Energy Reports 9 (2023) 4111–4121

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

Time-series clustering and forecasting household electricity demand using smart meter data

Hyojeoung Kim [a], Sujin Park [a], Sahm Kim [b,*]

[a] Department of Applied Statistics, Chung-Ang University, Seoul, Republic of Korea
[b] Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea

ARTICLE INFO

Article history:
Received 31 October 2022
Received in revised form 19 February 2023
Accepted 6 March 2023
Available online 14 March 2023

Keywords:
Time-series clustering
Residential electricity demand
Time-series forecasting
Smart meter data
Weather variables

ABSTRACT

This study forecasts electricity consumption in a smart grid environment. We present a bottom-up prediction method using a combination of forecasting values based on time-series clustering using advanced metering infrastructure (AMI) data, one of the core smart grid technologies. Remote data metering every 15 min to 1 h is possible with real-time communication on power generation information, consumption, and AMI development. Hence, its prediction is more challenging due to the large variation of each household's electricity. These issues were solved by time-series clustering methods using Euclidean distances and Dynamic Time Warping distance. The auto-regressive integrated moving average (ARIMA), ARIMA exogenous (ARIMAX), double seasonal Holt–Winters (DSHW), trigonometric, Box–Cox transform, autoregressive moving average errors, trend and seasonal components (TBATS), neural network nonlinear autoregressive (NNAR), and nonlinear autoregressive exogenous (NARX) models were used for demand forecasting based on clustering. The result showed that the time-series clustering method performed better than that using the total amount of electricity demand regarding the mean absolute percentage error and root mean squared error.

Hence, various exogenous variables were considered to improve model accuracy. The model considering exogenous variables—cooling degree day, humidity, insolation, indicator variables, and generation power consumption performed better than that without exogenous variables.

1. Introduction

Electricity generated from fossil fuels emits carbon dioxide (CO_2), causing air pollution and global warming. Interest in renewable energy has increased, including participation in RE100 to use 100% new renewable energy to cope with global warming and climate change. However, the power generation cost of renewable energy is higher than that of other energy sources; hence, expanding the supply without considering demand is impossible. Therefore, forecasting electricity consumption is essential for energy planning, management, and conservation (Amasyali and El-Gohary, 2018).

Time series, machine learning, and other methods have been used to predict electricity consumption. Høverstad et al. (2015) performed load prediction by extracting the characteristics of seasonal elements—daily and weekly dates—and confirmed that the performance of the double seasonal Holt–Winters (DSHW) algorithm was excellent at 6.3% based on the mean absolute percentage error (MAPE). Al-Musaylh et al. (2018) used multivariate adaptive regression spline, support vector regression(SVR), and autoregressive integrated moving average (ARIMA) models for electricity demand forecasting. Furthermore, Rahid (Rashid, 2018) employed such techniques as ARMIA and external smoothing to predict abundant electricity consumption accurately using smart grids. Kim et al. (2019) used ARIMA and ARIMA generalized autoregressive conditional heteroskedasticity(-GARCH) models, multiple seasonal exponential smoothing, and artificial neural network (ANN) models. They demonstrated that the ANN model with external variables (weather and holiday variables) worked best from 1 h to 1 day prior to forecasting. Furthermore, Pallonetto et al. (2022) applied long short-term memory (LSTM) and support vector machine (SVM)models for 1-h and 1-day-ahead load forecasting. Moreover, Hafeez et al. (2020) proposed a modified mutual information (MMI) factored conditional restricted Boltzmann machine (FCRBM) genetic wind-driven optimization (GWDO) hybrid model that incorporated preprocessing based on MMI, FCRBM for forecasting, and the GWDO algorithm for optimization to supplement nonlinear electrical load data.

However, predicting electricity consumption is challenging due to various factors, such as the physical properties of building, installed equipment (e.g., heating ventilation and air-conditioning

* Corresponding author.
E-mail addresses: hj8217@cau.ac.kr (H. Kim), bsujin314@gmail.com (S. Park), sahm@cau.ac.kr (S. Kim).

https://doi.org/10.1016/j.egyr.2023.03.042

system), outdoor weather conditions, and energy-use behavior of the building affecting consumption (Kwok and Lee, 2011). Several studies have investigated the relationship between electricity consumption and weather variables. Franco and Sanstad (2008) found that the effects of space cooling via air conditioners and using other appliances at high temperatures predominate. Furthermore, electricity consumption during weekends and holidays is lower than on weekdays. Hekkenberg et al. (2009) investigated the electricity demand pattern in the relative temperate climate of the Netherlands for possible changes regarding the increased use of cooling applications. They revealed a significant increment in the temperature dependence of electricity demand in May, June, September, October and during the summer holidays between 1970 and 2007. Maia-Silva (Maia-Silva et al., 2020) proposed that the AT-based models using temperature and humidity could predict electricity demand better during the historical period and in a warmer climate. The AT-based models projected higher demand across all regions compared with the temperature-only-based models. Furthermore, studies on the number of exogenous variables have also been recently conducted. Moreover, Román-Portabales et al. (2021) analyzed smart grid electrical load papers using ANN-based models. They verified that the number of exogenous variables varies depending on the prediction period.

As mentioned, many studies have used exogenous variables to predict electricity consumption for electricity usage for several reasons. Jain et al. (Jain et al., 2014) used electricity consumption for the previous two time steps, current temperature, current solar flux, an indicator variable denoting the weekend/holiday or weekday, sine of the current hour, and cosine of the current hour to predict the electricity consumption of a multi-family residential building. Georgescu (Georgescu et al., 2014) considered various weather variables, such as outdoor air temperature, relative humidity, solar radiation, wind speed, and wind direction as input variables. Furthermore, Yang et al. (2005) predicted energy consumption by considering weather variables (outdoor temperature, relative humidity, rainfall, wind speed, bright sunshine duration, and solar radiation) and the occupancy area and rate to predict non-residential energy consumption. Moreover, Lai et al. (2008) used the date, outdoor temperature/humidity, indoor temperature/humidity (bedroom and living room), and water temperature as input variables.

Consumer electricity usage patterns have been analyzed through clustering using advanced metering infrastructure (AMI) data besides introducing various exogenous variables and machine learning techniques. The AMI, a key smart grid technology, enables real-time remote metering between 15 min and 1 h using two-way communication technology. With the introduction and spread of AMI, more detailed predictions are possible with hourly usage information. However, this technology has developed significant variability for each household due to detailed household information. Accordingly, predictions have become difficult. his problem can be solved by analyzing and classifying patterns of household power usage through clustering to reduce volatility.

Various clustering analysis cases exist based on the spread of AMI. Rhodes et al. (2014) used *k*-means clustering on clustered households with similar electricity usage patterns per hour for each season, varying each season. However, variables such as telework status, television viewing time per week, and education level were significantly correlated with the average profile shape, owing to comparing clustering results and questionnaire answers through probit regression analysis. Bedi and Toshniwal (2019) generated seasons (summer, rainy, winter) through the *k*-means clustering analysis, comparing the prediction results with the ANN, recurrent neural network, and SVR models. Afterward, the results of ANNs, cyclic neural networks, and SVR models were compared on a season-by-season basis, demonstrating the best performance of LSTM.

Several other papers have analyzed the pattern of AMI users using *k*-means clustering algorithms (Qiu et al., 2016; Quilumba et al., 2014; Guerrero-Prado et al., 2020). Some studies have considered a hierarchical clustering technique besides the *k*-means. Son et al. (2020) proposed a demand prediction method based on the time-series cluster analysis using smart meter data. Normalized periodogram-based and autocorrelation-based distances were used as a hierarchical cluster analysis method. Electricity demand forecasting methods have been applied with autoregressive moving average (ARMA) errors, TBATS, DSHW, fractional ARIMA, ARIMAX, and NNAR. Moreover, Lee and Kim (2020) clustered households using hierarchical clustering methods, such as dynamic time warping (DTW) and periodogram for household AMI data. Power usage was then predicted in summer and winter using the NN-AR and TBATS models. The research found that predicting power usage by a cluster of households with similar usage was better than predicting all power usage at once. The DTW method displayed a stark visual between clusters compared with the periodogram method.

This study compares the effectiveness of forecasting the residential electricity consumption of a multi-family household using statistical and artificial intelligence-based models, including exogenous variables, such as weather variables (outdoor temperature, humidity, and solar radiation), and an indicator variable (weekend/holiday or weekday). This study proposes a method to cluster households using the time-series cluster analysis. We fitted a prediction model for each cluster and predicted the electricity demand of clusters to consider the electricity usage pattern of various households in predicting domestic housing AMI data. A cluster analysis method suitable for time-series data should be used (not a general cluster analysis method) for electricity usage data to consider the time-series characteristics. Therefore, this work performed calculations using the commonly used Euclidean distance, and DTW, which exhibited good performance in previous studies. Households were clustered using the *k*-means method. Fig. 1 presents a flowchart which reflect the proposed paradigm. The predictive performance of total residential power usage was compared using the following time-series models: ARIMA, ARIMAX, DSHW, TBATS, NN-AR, and NARX by cluster.

The contributions of this paper are presented as follows. A comprehensive comparison of multiple statistical and artificial intelligence-based models and result analysis applying time-series clustering while considering various exogenous variables have never previously been performed.

(1) The DSHW and TBATS, which are the univariate seasonal time-series models, display excellent results without clustering; however, the NARX model—the multivariate model—has the best accuracy with clustering.
(2) Weather variables (temperature, humidity, and solar radiation) and indicator variables (weekends, weekdays, or weekdays) act as effective variables for forecasting electricity demand, regardless of clustering.

The remainder of this paper is organized as follows. Section 2 explains the forecasting models and time-series cluster analysis methodologies in this study. Section 3 discusses the AMI electricity consumption data and preprocessing methods. Furthermore, Section 4 compares the results of the time series cluster analysis with predictive performance for each model. Section 5 details the necessity of cluster analysis and prediction results in predicting housing power demand. Finally, conclusions are drawn from an excellent model and future research directions are presented.

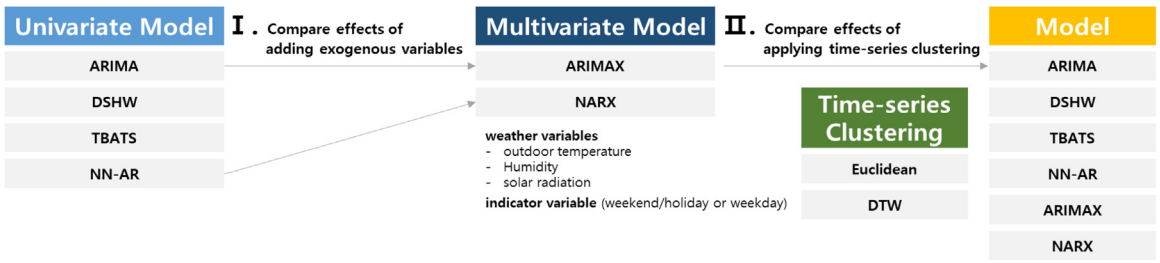

Fig. 1. Analysis process flowchart.

2. Methodology

2.1. Time-series forecasting models

2.1.1. ARIMA

The ARIMA model is a time-series data-based analysis technique depending on actions based on past knowledge or experience. It was first introduced by Box and Jenkins in 1976 (Box et al., 2015). It is a generalization of the ARMA model, which uses past observations and errors to describe current time-series values. Furthermore, it is an analysis technique used to predict the following indicators quarterly, semester-wise, or annually, review them weekly or monthly, and monitor trends for outliers. The model can also be applied to the analysis target demonstrating unstable and non-stationary characteristics (Peter and Silvia, 2012). The general form of the ARIMA(p,d,q) model is as follows:

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)\epsilon_t,$$

where $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$..

$$\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q, \tag{1}$$

where $\phi_p(B)$ corresponds to the equation for the autoregressive model, p denotes the order of the current model, $\theta_q(B)$ represents the equation for the moving average model, q is the order of the current model, d denotes the degree to which the first difference was included, ϵ_t indicates an error term or white noise with a mean of zero and a constant σ^2 value, and B corresponds to a backward shift operator.

2.1.2. ARIMAX

The ARIMAX model adds exogenous variables to the ARIMA model. It has been used as a prediction model in various fields, like the ARIMA model. When the degree of ARIMA is p,d,q and the number of exogenous variables is k, the exogenous variables are denoted as x_{it}, and the ARIMAX (p,d,q) model is as follows

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)\epsilon_t + \sum_{i=1}^{k} r_i x_{it}, \tag{2}$$

where $\phi_p(B)$ corresponds to the equation for the autoregressive model, p denotes the order of the current model, $\theta_q(B)$ is the equation for the moving average model, q represents the order of the current model, d denotes the equation containing the first difference, ϵ_t corresponds to an error term or white noise, and r_i is a coefficient of the exogenous variable, x_{it}.

2.1.3. DSHW model

Taylor (2003) (Taylor, 2003) proposed the DSHW model with two seasonal cycles. By adding one more seasonality, this model has two seasonal cycles. This study implemented Holt–Winter's dual-seasonal version to consider the day-by-day pattern of housing AMI data. This dual-seasonal addition method is more suitable for one-step head forecasting than the multiplication method. Finally, the DSHW model is defined as follows (Son et al., 2020).

$$L_t = \alpha\left(y_t - S_{t-s_1} - D_{t-s_2}\right) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$
$$S_t = \gamma\left(y_t - L_t - D_{t-s_2}\right) + (1-\gamma)S_{t-s_1}$$
$$D_t = \delta\left(y_t - L_t - S_{t-s_1}\right) + (1-\gamma)D_{t-s_2}$$
$$F_{t+h} = L_t + T_t \times h + S_{t+h-s_1} + D_{t+h-s_2} + \phi^h[y_t - (L_{t-1} - T_{t-1} - S_{t-s_1} - D_{t-s_2})]$$
$$L_{s_1} = \frac{1}{s_1}\sum_{t=1}^{s_1} y_t,\ L_{s_2} = \frac{1}{s_2}\sum_{t=1}^{s_2} y_t$$
$$T_{s_1} = \left(\frac{1}{s_1^2}\sum_{t=s_1+1}^{2s_1} y_t - \sum_{t=1}^{s_1} y_t\right),\ T_{s_2} = \left(\frac{1}{s_2^2}\sum_{t=s_2+1}^{2s_2} y_t - \sum_{t=1}^{s_2} y_t\right)$$
$$S_1 = y_1 - L_{s_1}, \ldots, S_{s_1} = y_{s_1} - L_{s_1}$$
$$D_1 = y_1 - L_{s_2}, \ldots, S_{s_2} = y_{s_2} - L_{s_2} \tag{3}$$

where y_t represents the real data and S_t and D_t denote the seasonal component over time t. Furthermore, L_t and T_t indicate the level and trend of the series at time t, respectively. In addition, F_{t+h} describes the forecasting value of h ahead of time t. Moreover, α, β, γ, andϕ correspond to smoothing parameters, which can be user-specified or internally estimated (Høverstad et al., 2015).

2.1.4. TBATS model

The TBATS model, introduced by De Livera et al. (2011), is a triple seasonality model that complements several limitations of previous models. First, we solved the nonlinearity problem of the dependent variable and assumed that the error term could follow the ARMA model without white noise. The periodicity of the non-constant cycle that could not be resolved within the existing model can be considered through this supplementation. It also reduced the time to estimate the model by expressing periodicity as a trigonometric function. The following formula represents this model:

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{\omega}-1}{\omega}, \omega \neq 0, \\ logy_t, \omega = 0, \end{cases}$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t,$$
$$b_t = (1-\phi)\bar{b} - \phi b_{t-1} + \beta d_t,$$
$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} + \gamma_i d_t,$$
$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t,$$
$$s_{j,t}^{*(i)} = -s_{j,t-1} sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t,$$

$$d_t = \sum_{i=1}^{p} \psi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t,$$

$$y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t, \tag{4}$$

where $y_t^{(\omega)}$ is the Box–Cox transformed observation for parameter ω at time t. l_t denotes the local level data, b represents the long-term trend, b_t is the short-term trend within period t, and ϕ denotes the damping parameter for the trend. Furthermore, p and q correspond to the orders of the ARMA error, ψ_i and θ_i represent the coefficients of ARMA(p,q), $\bar{b}$ is the long-term trend value, and αandβ correspond to the parameters for the level and trend, respectively. Moreover, $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$ are smoothing parameters, $\lambda_j^{(i)} = 2\pi/m_i$, k_i is the number of trigonometric functions consisting of the ith periodicity($s_t^{(i)}$), $s_{j,t}^{(i)}$ represents the stochastic level of the ith seasonal component by $s_t^{(i)}$, and $s_{j,t}^{*(i)}$ denotes the stochastic level of the ith seasonal component (Kim et al., 2019).

2.1.5. NN-AR

The ANN is a mathematical model of biological neurons. It was proposed by Warren McCulloch and Walter Pitts in 1943 (McCulloch and Pitts, 1943). In 1958, Frank Rosenblatt proposed the structure used in ANNs today (Rosenblatt, 1958). The ANN consists of interconnected artificial neurons that learn the relationship between the input and output by adjusting the weights between the neurons through a backpropagation algorithm. Fig. 2 shows the common structure of ANNs. The ANN, a non-linear modeling technique, was introduced to solve linear inaccurate modeling issues, gaining popularity in the energy prediction field (Mustapa et al., 2020). An input layer of the model is entered with the data, and the result is output through an output layer via one or more hidden layers. This model is called the feed-forward neural network (FFNN) because information travels only in one direction (i.e., forward, in this model). The NN-AR models were designed to forecast a time series from past values. Lagged values of the time series can be used as input to a neural network using the time-series data, like those used in a linear autoregression model. Additionally, the model used in this study is a single hidden layer. The number of nodes in the hidden layer was 14, and a logistic sigmoid function was used as the activation function. The NN-AR is applied to time-series data with discrete, nonlinear, and autoregressive tendencies. It can be written as follows (Ruiz et al., 2016):

$$y(t) = f(y(t-1), y(t-2), \ldots, y(t-p)), \tag{5}$$

where p corresponds to the past values of the series, y represents the predicted value at time t, and $f(\cdot)$ is a nonlinear function. This model is easy to understand and applicable to many problems because of its simple configuration.

2.1.6. NARX

This study used NARX to facilitate using exogenous variables to predict housing AMI data. Like the NN-AR model, the NARX model consists of an FFNN structure in which an input value and a value multiplied by weight are combined after data are entered into the input layer. Afterward, a resulting value is output through the activation function. The structure of the NARX is represented by the following equation (Lin et al., 1996):

$$\begin{aligned} y(t) &= f[u(t-D_u), \ldots, u(t-1), u(t), \\ &\quad y(t-D_y), \ldots, y(t-1)] \\ &= \psi[u(t), y(t-1), \ldots, y(t-D)], \end{aligned} \tag{6}$$

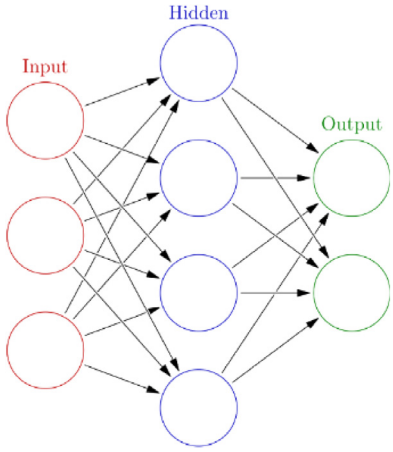

Fig. 2. Artificial neural network structure.

where u corresponds to the value of the exogenous variables and y is the predicted value of the network at time t. Furthermore, D_u and D_y denote the order of the exogenous variables and predicted values. Moreover, f is a nonlinear function, which can be approximated using a multilayer perceptron. The resulting system is called a NARX. This study expressed the function corresponding to the mapping performed by the multilayer perceptron as ψ. The number of hidden layers was designated as a single layer. The number of nodes of the hidden layer was set to 16, and the activation function was a logistic sigmoid function.

2.2. Time series clustering algorithm

2.2.1. Euclidean distance

The Euclidian distance formula is the most widely used distance function in the clustering context; it determines the distance between two points in n-dimensional space (Bouhmala, 2016). The distance between the two points, q and c, with coordinates, $(q_1, q_2, \ldots, q_n)$ and $(c_1, c_2, \ldots, c_n)$, respectively, is expressed by the following Euclidean distance formula:

$$\begin{aligned} Dist\ (Euclidean) &= \sqrt{(q_1 - c_1)^2 + (q_2 - c_2)^2 + \cdots + (q_n - c_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2}\ , \end{aligned} \tag{7}$$

Specifically, Euclidean distance is the simplest method of clustering, where the distance between two points is calculated. The points are considered to be in the same cluster if the distance is less than or equal to a certain distance.

2.2.2. DTW distance

Kruskal (1938) first addressed the DTW distance, which was proposed to determine patterns of time series by Berndt and Clifford (1994) (Berndt and Clifford, 1994). The DTW replaces the one-to-one point comparison used in Euclidean distance with a many-to-one (and vice-versa) comparison, as illustrated in Fig. 3. The main feature of this approach is that it allows recognizing similar shapes, even if they represent signal transformations (Kumar and Baboo, 2017). Given that the two time-series sequences $Q = q_1, q_2, \ldots, q_i, \ldots, q_n$, and $C = c_1, c_2, \ldots, c_j, \ldots, c_m$, are given here, an $n \times m$ matrix is created through two time series, and the (i, j)th element of this matrix indicates the Euclidean distance $(q_i - c_i)^2$ between the two points q_i and c_i, which is used to search for the optimal wapping path. The warping path $(W = w_1, w_2, \ldots, w_k, \ldots, w_K,)$ is a set of wapping distances

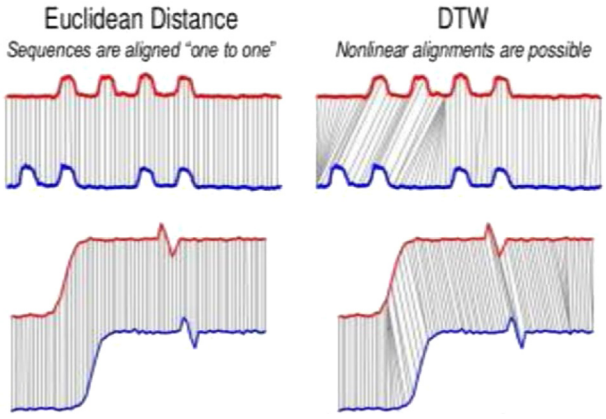

Fig. 3. Difference between Euclidean and DTW distance.

representing the mapping between Q and C, and must be continuous. The kth element in W is defined as w_k, which is called the wapping distance. Finally, it can be expressed as a *Dist* (*DTW*) equation because it is the same as finding a path in which the sum of these warping distances (w_k) is minimized (i.e. the cost of the warping path is minimized):

$$Dist\ (DTW) = \min\{\sqrt{\sum_{k=1}^{K} w_k/K}\}, \tag{8}$$

Part of the effectiveness of the DTW is due to the algorithm "searching" for better mapping. There are various step patterns suited for different situations, and Eq. (9) can be considered a standard step pattern and is used in conjunction with the DTW algorithm in this paper (Ma and Angryk, 2017):

$$D\ (\mathrm{i},\mathrm{j}) = (q_i - c_j)^2 + \min\begin{Bmatrix} D(i-1,j-1) \\ D(i,j-1) \\ D(i-1,j) \end{Bmatrix}, \tag{9}$$

2.2.3. K-means clustering

The k-means algorithm belongs to unsupervised machine learning. It was introduced by Fames MacQueen in 1967 (MacQueen, 1967). The algorithm combines data into k clusters. k represents the number of clusters and "means" implies the average of each cluster by grouping data with similar characteristics. Specifically, the average of each cluster is used and grouped into k clusters. The average implies the center of each cluster and the average distance of the data. The algorithm first determines the required number of clusters, k, and sets the initial centroid. Afterward, it traverses all data and allocates them to the cluster to which the nearest centroid belongs. The centroid is then moved to the center of the cluster. The process is repeated until no data are left to be allocated to the cluster. The Euclidean distance is selected as the similarity index for a given data set X containing n multidimensional data points and the category k to be divided. The clustering targets minimize the sum of the squares of the various types, that is, it minimizes the following:

$$d = \sum_{k=1}^{k}\sum_{i=1}^{n} \|(x_i - u_k)\|^2, \tag{10}$$

where k denotes the number of centers in the cluster, u_k represents the kth center, and x_i denotes the ith data in the data set.

This study referenced the silhouette coefficient, calculated by considering the mean intra-cluster distance a and the mean nearest-cluster distance b for each data point, to determine the number of clusters (Shahapure and Nicholas, 2020). The value of the silhouette coefficient $s((i)$ for the ith $x((i)$ is defined by Eq. (11), as follows:

$$s\ (i) = \frac{b^{(i)} - a^{(i)}}{max\left(a^{(i)}, b^{(i)}\right)}, \tag{11}$$

where $a((i)$ represents data cohesion in a cluster and is the average distance from the rest of the data in the same cluster as $x(i)$. A smaller distance indicates higher cohesion. Additionally, $b((i)$ represents intercluster separation, which is the average distance between $x((i)$ and all data in the closest cluster. Furthermore, $b(i)$, $a(i)$, and $s((i)$ should be large, small, and close to 1, respectively, to optimize the number of clusters.

When the k value is determined, the k-means algorithm randomly specifies k centroids from the data set, and each data point is allocated as a group of the nearest centroids. In the assigned group, the process is repeated until the centroids converge by reassigning them. Furthermore, the group is the side closest to the convergent final centroids. (Kim et al., 2022)

3. Data

3.1. Electricity consumption data

We collected empirical data from a building called Suwon Ggumegreen, which comprised 32 17-story multi-family residential buildings in Suwon in the Republic of Korea. The Ggumegreen data set consists of electrical consumption data for 138 households with 1-h intervals. They agreed to collect data as a part of the survey of AMI installation households. A set consists of 13 weeks of summer data from 00:00 on June 3, 2019, to 23:00 on September 1, 2019, comprising hour-by-hour data.

The outliers were confirmed by examining the total power values by day and hour using box plots, as illustrated in Fig. 4(a and b). The daily total power consumption in Fig. 4(a) reveals that consecutive usage numbers up to 60 kW, breaking off, and long after, they reoccurred at more than 100 kW. A total of 136 households were analyzed, considering the average total daily electricity usage by household was less than 10 kW, and the total daily power consumption of 100 kW or more occurred in only two households, excluding cases where the sum of the daily electricity consumption of households was 100 kW or more or zero (or missing). The "three-sigma rule" was used in hourly total power consumption. As a result of applying the three-sigma rule, the range was from −6 to 7 kW. As depicted in Fig. 4(b), no negative numbers are in the data, and when the number exceeds 7 kW, it is a very small ratio of 0.097% of the total, so the maximum value was set to 7 kW, and when it exceeds that, it was converted to 7 kW.

Fig. 5(a and b) presents the entire power consumption boxplots after preprocessing by day and hour, respectively. The number displays a natural continuous flow without breaking off. Fig. 6(a) and (b) depicts before and after outlier processing. The difficulty of observing the overall data distribution was solved due to the extreme outlier value.

The average electricity usage by the hour was examined by day of the week to determine the characteristics of the data (Fig. 7). Saturdays and Sundays were verified to display different patterns compared to weekdays. Furthermore, electricity usage was verified to be higher during the daytime on weekends than on weekdays. Therefore, the average electricity usage was examined by the hour by dividing June 6 to August 15 into weekends (including holidays) and weekdays, considering the difference between the day on and day off (Fig. 8). We confirmed that the electricity usage pattern on the day on and the day off was noticeably different. Accordingly, the days were converted into categorical variables (the day on = 1 and the day off = 0) and were added to the exogenous variables.

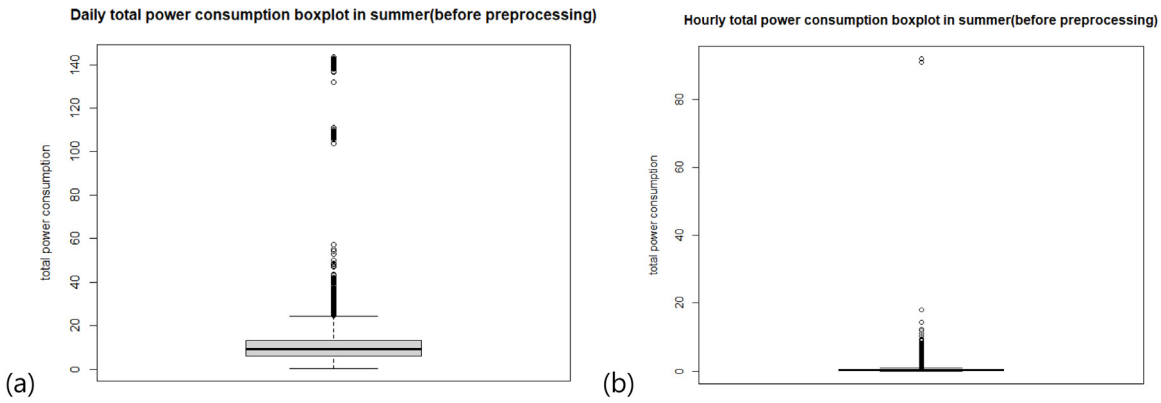

Fig. 4. Total power consumption boxplots before preprocessing: (a) Daily and (b) Hourly.

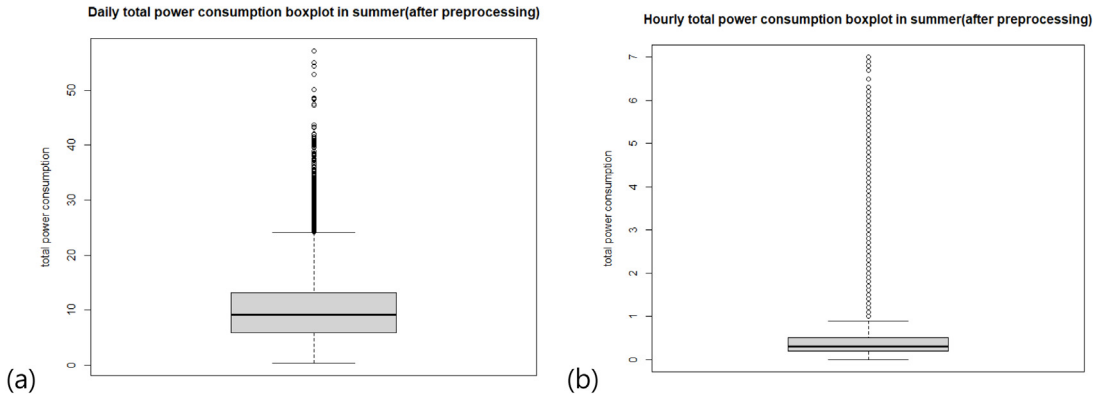

Fig. 5. Total power consumption boxplots after preprocessing: (a) daily and (b) hourly.

3.2. Weather data

Weather information was obtained from a weather station 3 km away from a multi-family house whose power consumption was observed. We extracted temperature, humidity, and solar radiation data, which have frequently been used in previous papers. We considered heating degree day (HDD) and cooling degree day (CDD) instead of temperature because temperature significantly depends on air-conditioning and heating devices. The data were not affected because they were taken during the summer. Therefore, the temperature was converted to the CDD as follows (Jung and Kim, 2014).

$$\text{HDD} = \begin{cases} 18 - T_t, & \text{if } T_t \leq 18 \\ 0, & \text{else} \end{cases}$$

$$\text{CDD} = \begin{cases} T_t - 24, & \text{if } T_t \geq 24 \\ 0, & \text{else} \end{cases} \tag{12}$$

The value of solar radiation for the time excluding 6 to 20 o'clock was converted to zero for the solar radiation, considering the sunrise and sunset times in summer in Korea. The missing values of the remaining meteorological elements were treated using linear interpolation considering the continuous flow of meteorological data.

4. Application of the models

4.1. Clustering results

Time-series cluster analysis was performed using Euclidean and DTW distance calculations. The complete data required a long time to calculate the cluster distance. Therefore, the distance was calculated using electricity every hour from 00:00 on Monday to 23:00 on Sunday by extracting a week that best represents the electricity usage cycle for housing. The silhouette score was calculated after households were classified into clusters from 2 to 10 by Euclidean distance (Fig. 9). The score decreased for the first five clusters;, therefore, it increased slightly with six clusters and then decreased again. The closer the silhouette score is to 1, the more optimized the number of clusters. The silhouette score was the best when the number of clusters was two, but the number of clusters was judged too small to distinguish the pattern, and the number of clusters was set to six. In addition, the DTW was equally divided into six clusters to compare with the Euclidean distance.

We searched for the number of households in each group using a histogram to examine the characteristics of the six groups (Fig. 10). The first and last groups consisted of three and 51 households, respectively. Fig. 11 presents the weekly power usage by clustering. There is a difference in power usage for each time zone and on weekends between the six detected cluster patterns. Cluster 2 started in the morning and peaked in the late afternoon, then gradually decreased in the evening, demonstrating continuous power use from morning to evening. Clusters 3, 4, and 5 displayed a flow of power consumption in the morning, cut off, and resumed in the evening. Cluster 6 is rarely used in the morning and consumes considerable power in the evening, so it is estimated that Cluster 6 has a high commuting rate for family members. Furthermore, Clusters 2 and 5 were less used on the weekends, whereas Clusters 4 and 6 were more used on weekends. The power usage size indicates that Clusters 2, 5, and 6 have high power usage over time for the peak point, whereas Clusters 3, 4, and 5 revealed relatively low power consumption.

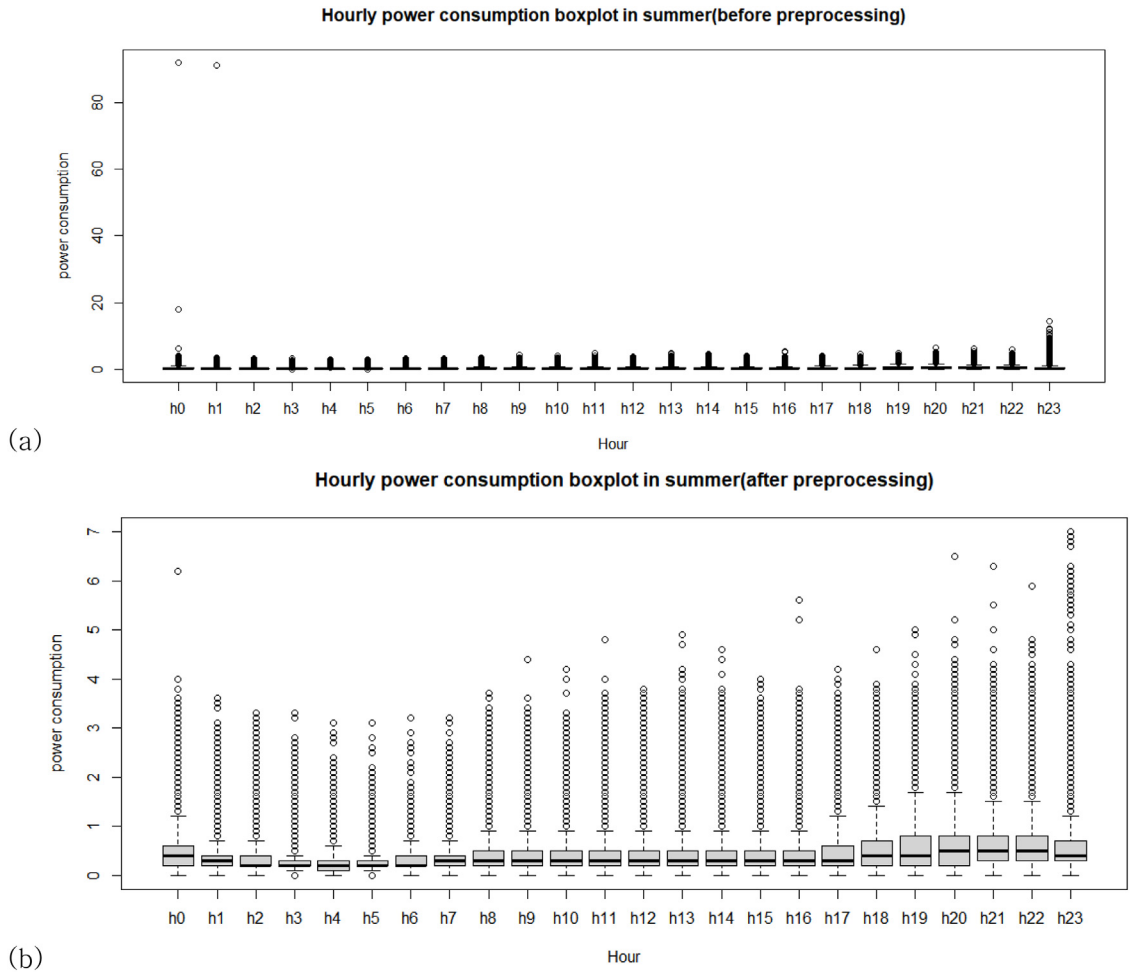

Fig. 6. Hourly load demand boxplots: (a) before and (b) after preprocessing.

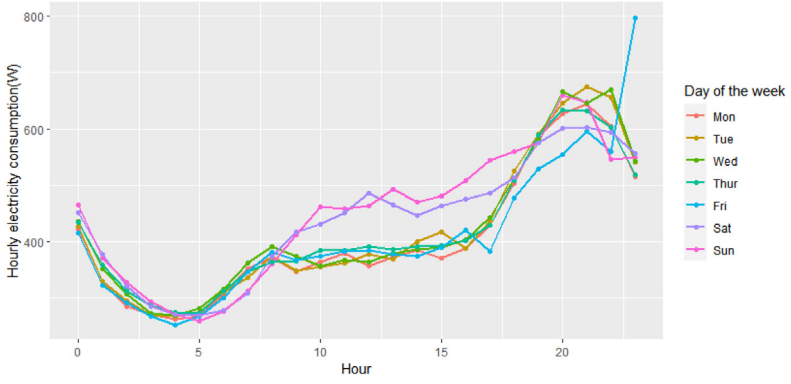

Fig. 7. Average daily load demand plot (by day of the week).

The number of households was examined by clustering six DTWs (Fig. 12). If the number of households in a cluster is too small, changing the power usage pattern of any one household in such a cluster can result in overfitting, so caution is required. For the Euclidean distance, Cluster 1 contains three households, whereas DTW is more evenly distributed because there are more than 10 households distributed in all clusters. Fig. 13 presents the weekly power usage by clustering. As with the Euclidian distance method, differences exist between the six detected cluster patterns depending on the time zone, weekend, and power usage. Cluster 1 continues to consume power regardless of the time zone. Clusters 2, 4, 5, and 6 used power in the morning and late afternoon to evening. Among them, Cluster 4 seems to have higher morning power usage than other clusters. Furthermore, Cluster 2 used less power on weekends than on weekdays, but Clusters 5 and 6 used more power than usual on weekends during the afternoon. According to the power usage, in Cluster 3, households have various periods when power usage increases, but they do not usually use substantial power, but the power usage tends to increase suddenly. Cluster 3 demonstrated a significant

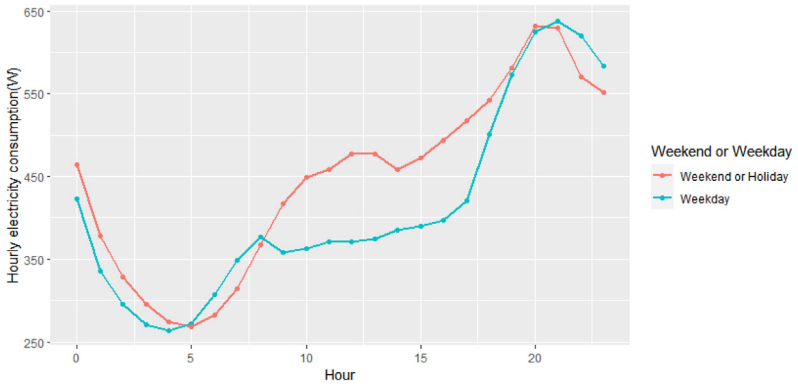

Fig. 8. Average daily load demand (by day-off or on).

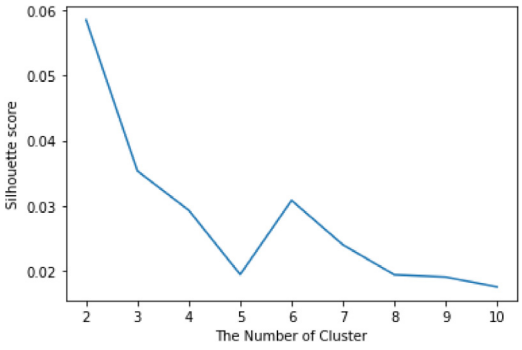

Fig. 9. Silhouette score using Euclidean distance.

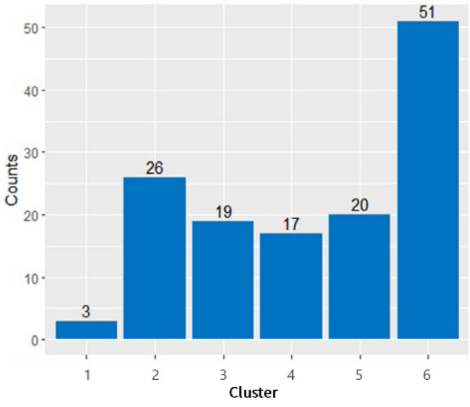

Fig. 10. Histogram of the cluster using Euclidean distance.

variation in daily electricity usage. In addition, Cluster 5 has a higher average power consumption than Cluster 6.

4.2. Forecasting results

The prediction result was calculated for each cluster after fitting the prediction model using the corresponding value after calculating the total power usage of households in each of the six clusters. Finally, two weeks were predicted for all households using electricity for housing by adding all the prediction results by cluster. Furthermore, training data were used to train the prediction model for 11 weeks from 00:00 on June 3, 2019 to 23:00 on August 18, 2019. Test data were used for two weeks from 00:00 on August 19, 2019, to 23:00 on September 1, 2019, to evaluate prediction performance. Furthermore, the ARIMA, ARIMAX, DSHW, TBATS, NN-AR, and NARX models were used

Table 1
Parameter estimations of the ARIMA(2,1,3) model.

Parameter	Estimates
ϕ_1	1.5463
ϕ_2	−0.7284
θ_1	−2.1105
θ_2	1.6597
θ_3	−0.5421

to fit the prediction model of the households belonging to each cluster. The predicted performance was evaluated by comparing the predicted value using the total electricity consumption data without clustering with the value calculated by adding the predicted values for each cluster.

We chose the model with the lowest MAPE in each cluster. Afterward, the forecasting values were compared using the root mean squared error (RMSE) and MAPE. Performance evaluation indices are widely used to evaluate model performance, especially for short-term load forecasting, and MAPE is defined as follows:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \widehat{y_t}}{y^t}\right|. \tag{13}$$

Furthermore, the RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\widehat{y_t} - y_t)^2}, \tag{14}$$

where y_t represents the actual value and $\widehat{y_t}$ corresponds to the forecasted demand at time t.

Households were divided into six clusters using each cluster analysis method. Furthermore, the optimal prediction model was fitted for each cluster to predict two weeks (336 h) of data. The prediction was made using the time-series cross-validation method. Furthermore, all data on the day previous to the prediction day were kept as learning data to maximize the learning data (Fig. 14). The predictive results of the appropriate prediction model were presented using the total usage without clustering to compare the predictive performance of the method of fitting individual models by dividing households into clusters; Tables 1 to 4 present the parameters estimated from the entire non-clustered training set. Furthermore, the ARIMA and ARIMAX models were implemented for forecasting in each cluster. The parameters were automatically specified for each cluster using auto.arima function in R (Hyndman and Khandakar, 2008). Table 5 provides the forecast results for the summer data.

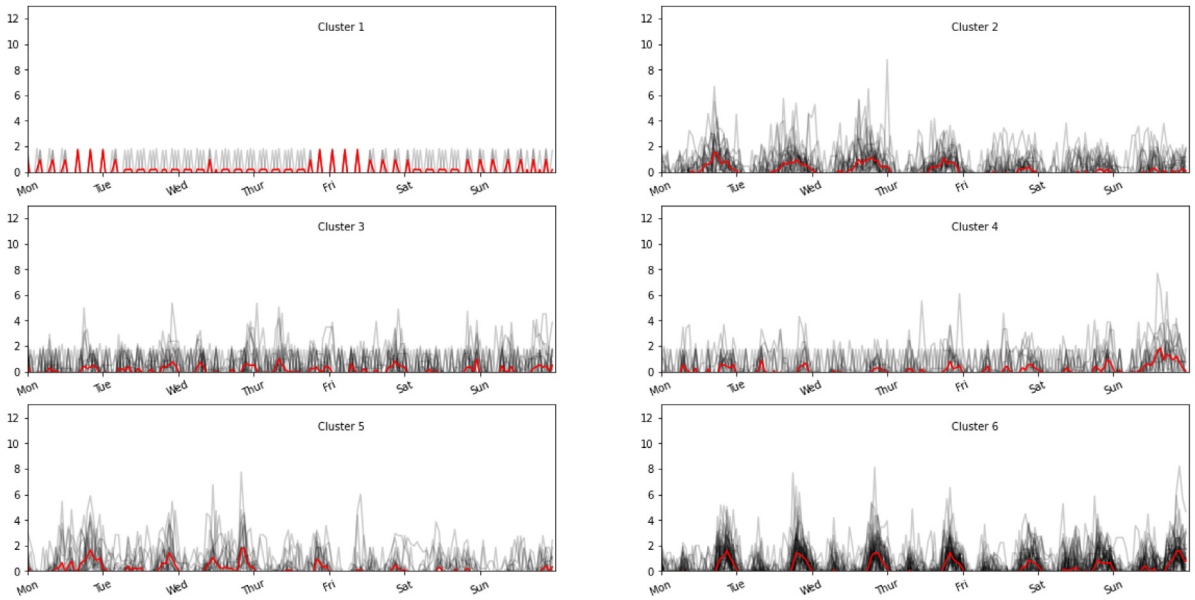

Fig. 11. Household electricity usage of six clusters by Euclidean distance during a week.

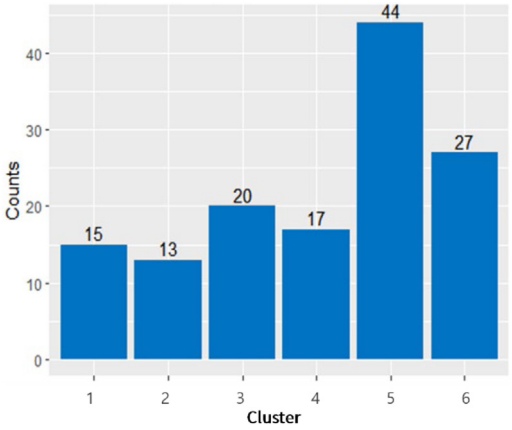

Fig. 12. Histogram of the cluster using DTW distance.

Table 2
Parameter estimations of the ARIMAX(2,1,3) model.

Parameter	Estimates
ϕ_1	1.5819
ϕ_2	−0.7987
θ_1	−2.2065
θ_2	1.8208
θ_3	−0.6089
Humidity	−0.0857
Solar radiation	−5.9592
CDD	3.3915
Indication variable (Day-off or not)	2.2029

Table 3
Parameter estimations of the DSHW model.

Parameter	Estimates
α (Level)	0.1250
β (Trend)	0.0038
γ (Seasonal 1)	0.0231
δ (Seasonal 2)	0.1123

Table 4
Parameter estimations of the TBATS model.

Parameter	Estimates
α (Level)	0.0949
$\gamma_1^{(24)}$	−0.000057
$\gamma_1^{(168)}$	−0.000074
$\gamma_2^{(24)}$	0.000013
$\gamma_2^{(168)}$	0.000033
ϕ_1	1.6542
ϕ_2	−0.9035
θ_1	−1.5454
θ_2	0.8432

Table 5
Forecast performance evaluations in terms of RMSE and MAPE.

Model	Cluster X		Euclidean		DTW	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
ARIMA	14.793	26.199	12.769	22.391	13.055	23.320
ARIMAX	13.098	22.365	11.047	17.594	11.538	20.159
DSHW	6.828	9.055	6.663	8.981	6.626	8.849
TBATS	6.692	8.889	6.644	8.804	6.717	8.887
NN-AR	21.730	19.528	6.973	10.226	6.989	9.363
NARX	13.723	12.741	6.348	8.799	5.939	7.778

First, the total electricity consumption of all households was predicted by fitting the ARIMA, ARIMAX, DSHW, TBATS, NN-AR, and NARX models without clustering households, resulting in the MAPE error rates of 26.199%, 22.365%, 9.055%, 8.889%, 19.528%, and 12.741%, respectively. The introduction of regressors (i.e., covariates) in extreme smooth (ETS) models is not feasible because the DSHW and TBATS were ETS based models (forecastability is the ETS equivalent of the invertibility of the ARIMA models) (Hyndman et al., 2008). Therefore, exogenous variables were considered only in the ARIMA and NN-AR models. The ARIMAX model, which added exogenous variables to the ARIMA model, performed better than ARIMA in terms of the MAPE and RMSE values, with an RMSE from 14.793 to 13.098. Similarly, the NARX model, which added exogenous variables to the NN-AR model, displayed a significant performance improvement for

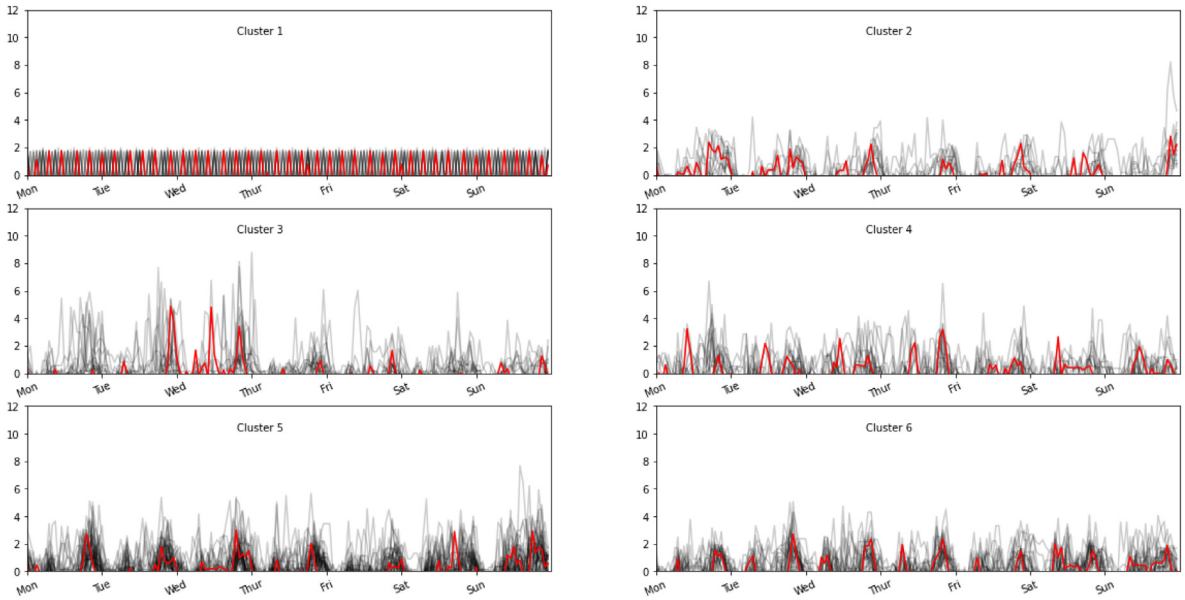

Fig. 13. Household electricity usage of six clusters using DTW distance during a week.

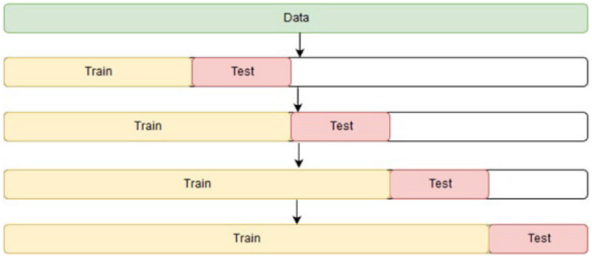

Fig. 14. Time-series cross-validation process.

MAPE and RMSE, with an RMSE from 21.730 to 13.723. The exogenous variables (outdoor temperature, humidity, solar radiation, and weekend indicator variable) affected the electricity demand forecast, as better accuracy was calculated by adding exogenous variables.

The univariate seasonal time-series models (DSHW and TBATS) demonstrated more accurate predictive power without clustering. The prediction results clustered using the Euclidean or DTW distance by fitting the model to each cluster were better than non-cluster options in all predictive models. The results improved significantly after clustering using the ARIMA, ARIMAX, NN-AR, and NARX models. The performance was improved in the case of ETS-based DSHW and TBATS although, the effect was smaller than in the other models. Regardless of the Euclidian and DTW distance calculation methods, both methods with clustering demonstrated better performance compared to those without clustering, but each model has a different optimal clustering distance calculation method. The best models for Euclidean distance calculation were the ARIMA, ARIMAX, and TBATS model, whose MAPE improved by 26.2% to 22.4%, 22.4% to 17.6%, 8.9% to 8.8% and 9.1% to 8.8% respectively. Moreover, the optimal models for the DTW distance calculation method were the DSHW, NN-AR, and NARX models, and their MAPE values improved by 9.1% to 8.8%, 19.5% to 9.4%, and 12.7% to 7.8% respectively.

These results revealed that the forecasting performance after classifying the households with similar power usage patterns using cluster analysis was higher than when predicting the total electricity usage without classification. The clustering effect based on the distance calculation was slightly different for each model. However, in the case of NARX, which performed best among several models, the RMSE was 6.348 for the Euclidean distance calculation, but for DTW, it was 5.939, revealing a significant difference depending on the distance calculation method. In NARX, the DTW distance calculation method is advantageous.

5. Conclusion

This study performed a time-series cluster analysis using 1-h unit electricity consumption data gathered from 136 households for household AMI data under a smart grid environment. Furthermore, it predicted electricity consumption by cluster after clustering each household. The Euclidean and DTW distance calculation methods were used as time-series cluster analysis methods. All households were divided into six clusters. The prediction performance of each model was compared by the time-series cluster analysis method using the ARIMA, ARIMAX, DSHW, TBATS, NN-AR, and NARX models.

Furthermore, the results were predicted and compared to prove the excellence of the method of predicting the total electricity consumption for all households via a bottom-up method using cluster analysis and predicting and summing the power usage for each cluster. Regardless of the model, clustering households with similar electricity usage patterns and predicting electricity usage for each cluster was better than predicting the total electricity usage for all households without clustering. The MAPE and RMSE, indicators for evaluating predictive performance, performed best, especially when the NARX model was used as a cluster with the DTW distance calculation. Fig. 15 illustrates the results of the NARX model with the DTW cluster, which had the best performance with the actual value during the prediction period of two weeks. Except for approximately four days, the predicted value from the low point to the peak moved quite similarly to the actual value.

Furthermore, this study primarily used derivative variables, such as weather and dates. This study confirmed that performance improved when exogenous variables (e.g., the CDD, humidity, and solar radiation) and indication variables (day off or on) were added compared to the basic model using only electricity consumption. However, household electricity demands can be affected by various factors, such as electrical consumer behavioral

Fig. 15. Time series cross-validation process. Electricity demand forecasting plot using NARX with DTW cluster.

patterns, and building information (e.g., the number of household members, telecommuting status, residence type, and house size). Therefore, future studies should classify households in further detail using electricity consumption data, apartment characteristics (e.g., space size and location), household characteristics (e.g., the number of household members and telecommuting ratio), and various climate factors in the region. If an analysis of the influence of electricity consumption on various information is performed for each finely clustered household, stable peak loads in summer and winter can be predicted even in rapidly changing situations (e.g., rapid climate change).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sahm kim reports financial support was provided by National Research Foundation of Korea(NRF).

Data availability

The authors do not have permission to share data.

References

Al-Musaylh, M.S., Deo, R.C., Adamowski, J.F., Li, Y., 2018. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Adv. Eng. Inform. 35, 1–16.

Amasyali, K., El-Gohary, N.M., 2018. A review of data-driven building energy consumption prediction studies. Renew. Sustain. Energy Rev. 81, 1192–1205.

Bedi, J., Toshniwal, D., 2019. Deep learning framework to forecast electricity demand. Appl. Energy 238, 1312–1326.

Berndt, D.J., Clifford, J., Using dynamic time warping to find patterns in time series. In: KDD workshop, Vol. 10, 16, pp. 359–370.

Bouhmala, N., 2016. How good is the euclidean distance metric for the clustering problem. In: 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). IEEE, pp. 312–315.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.

De Livera, A.M., Hyndman, R.J., Snyder, R.D., 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. J. Amer. Statist. Assoc. 106 (496), 1513–1527.

Franco, G., Sanstad, A.H., 2008. Climate change and electricity demand in California. Clim. Change 87 (1), 139–151.

Georgescu, M., Eccles, E., Manjunath, V., Swindle, E., Mezic, I., 2014. Machine learning methods for site-level building energy forecasting and data rectification. Build. Simul. Optim. 133.

Guerrero-Prado, J.S., Alfonso-Morales, W., Caicedo-Bravo, E., Zayas-Pérez, B., Espinosa-Reza, A., 2020. The power of big data and data analytics for AMI data: A case study. Sensors 20 (11), 3289.

Hafeez, G., Alimgeer, K.S., Khan, I., 2020. Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid. Appl. Energy 269, 114915.

Hekkenberg, M., Benders, R.M.J., Moll, H.C., Uiterkamp, A.S., 2009. Indications for a changing electricity demand pattern: The temperature dependence of electricity demand in the Netherlands. Energy Policy 37 (4), 1542–1551.

Høverstad, B.A., Tidemann, A., Langseth, H., Öztürk, P., 2015. Short-term load forecasting with seasonal decomposition using evolution for parameter tuning. IEEE Trans. Smart Grid 6 (4), 1904–1913.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. J. Stat. Softw. 27, 1–22.

Hyndman, R., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with Exponential Smoothing: The State Space Approach. Springer Science & Business Media.

Jain, R.K., Smith, K.M., Culligan, P.J., Taylor, J.E., 2014. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl. Energy 123, 168–178.

Jung, S.W., Kim, S., 2014. Electricity demand forecasting for daily peak load with seasonality and temperature effects. Korean J. Appl. Statist. 27 (5), 843–853.

Kim, H., Kim, J.M., Kim, S., 2022. Frost forecasting considering geographical characteristics. Adv. Meteorol..

Kim, Y., Son, H.G., Kim, S., 2019. Short term electricity load forecasting for institutional buildings. Energy Rep. 5, 1270–1280.

Kumar, R., Baboo, C.S., 2017. Motif discovery comparison using multivariate rhythm sequence technique and dynamic time warping (DTW) in time series data.

Kwok, S.S., Lee, E.W., 2011. A study of the importance of occupancy to building cooling load in prediction by intelligent approach. Energy Convers. Manage. 52 (7), 2555–2564.

Lai, F., Magoules, F., Lherminier, F., 2008. Vapnik's learning theory applied to energy consumption forecasts in residential buildings. Int. J. Comput. Math. 85 (10), 1563–1588.

Lee, J.Y., Kim, S., 2020. Time series clustering for AMI data in household smart grid. Korean J. Appl. Statist. 33 (6), 791–804.

Lin, T., Horne, B.G., Tino, P., Giles, C.L., 1996. Learning long-term dependencies in NARX recurrent neural networks. IEEE Trans. Neural Netw. 7 (6), 1329–1338.

Ma, R., Angryk, R., 2017. Distance and density clustering for time series data. In: 2017 IEEE International Conference on Data Mining Workshops. ICDMW, IEEE, pp. 25–32.

MacQueen, J., 1967. Proceedings of the fifth berkeley symposium on mathematical statistics and probability. In: Some Methods for Classification and Analysis of Multivariate Observations. pp. 281–297.

Maia-Silva, D., Kumar, R., Nateghi, R., 2020. The critical role of humidity in modeling summer electricity demand across the United States. Nature Commun. 11 (1), 1–8.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5 (4), 115–133.

Mustapa, R.F., Dahlan, N.Y., Yassin, A.I.M., Nordin, A.H.M., 2020. Quantification of energy savings from an awareness program using NARX-ANN in an educational building. Energy Build. 215, 109899.

Pallonetto, F., Jin, C., Mangina, E., 2022. Forecast electricity demand in commercial building with machine learning models to enable demand response programs. Energy and AI 7, 100121.

Peter, Ď., Silvia, P., ARIMA vs. ARIMAX–which approach is better to analyze and forecast macroeconomic time series. In: Proceedings of 30th International Conference Mathematical Methods in Economics, Vol. 2, pp. 136–140.

Qiu, W., Zhai, F., Bao, Z., Li, B., Yang, Q., Cao, Y., 2016. Clustering approach and characteristic indices for load profiles of customers using data from AMI. In: 2016 China International Conference on Electricity Distribution. CICED, IEEE, pp. 1–5.

Quilumba, F.L., Lee, W.J., Huang, H., Wang, D.Y., Szabados, R., 2014. An overview of AMI data preprocessing to enhance the performance of load forecasting. In: 2014 IEEE Industry Application Society Annual Meeting. IEEE, pp. 1–7.

Rashid, M.H., 2018. AMI smart meter big data analytics for time series of electricity consumption. In: 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). IEEE, pp. 1771–1776.

Rhodes, J.D., Cole, W.J., Upshaw, C.R., Edgar, T.F., Webber, M.E., 2014. Clustering analysis of residential electricity demand profiles. Appl. Energy 135, 461–471.

Román-Portabales, A., López-Nores, M., Pazos-Arias, J.J., 2021. Systematic review of electricity demand forecast using ANN-based machine learning algorithms. Sensors 21 (13), 4544.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65 (6), 386.

Ruiz, L.G.B., Cuéllar, M.P., Calvo-Flores, M.D., Jiménez, M.D.C.P., 2016. An application of non-linear autoregressive neural networks to predict energy consumption in public buildings. Energies 9 (9), 684.

Shahapure, K.R., Nicholas, C., 2020. Cluster quality analysis using silhouette score. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics. DSAA, IEEE, pp. 747–748.

Son, H.G., Kim, Y., Kim, S., 2020. Time series clustering of electricity demand for industrial areas on smart grid. Energies 13 (9), 2377.

Taylor, J.W., 2003. Short-term electricity demand forecasting using double seasonal exponential smoothing. J. Oper. Res. Soc. 54 (8), 799–805.

Yang, J., Rivard, H., Zmeureanu, R., 2005. On-line building energy prediction using adaptive artificial neural networks. Energy Build. 37 (12), 1250–1259.