Microsystem Technologies (2022) 28:121–135
https://doi.org/10.1007/s00542-019-04479-z

TECHNICAL PAPER

Applying the naïve Bayes classifier to HVAC energy prediction using hourly data

Chang-Ming Lin[1] · Sheng-Fuu Lin[1] · Hsin-Yu Liu[2] · Ko-Ying Tseng[2]

Received: 25 February 2019 / Accepted: 10 May 2019 / Published online: 22 June 2019

Abstract

Heating ventilation and air conditioning (HVAC) accounts for approximately 50% of the total energy consumption of buildings. Therefore, many studies have been focused on the simulation and optimal control of HVAC power consumption or the prediction of energy consumption through the construction of energy consumption models and the improvement of HVAC power consumption through energy management methods. The prediction of energy consumption by optimal energy-saving control or energy baseline is dependent on an accurate energy consumption model, however, the accuracy of the energy consumption model is influenced by the model variables. In addition, different operating periods and load conditions also lead to different changes in energy consumption, which will affect the accuracy of optimal energy consumption control or prediction of energy consumption. The present study proposes a method to enhance the accuracy and sensitivity of HVAC power consumption prediction, which involves the use of a clustering technique to locate clusters with similar information within hourly data, the construction of energy consumption models by converting the clustered hourly data into monthly data, and the application of the proposed Naïve Bayes classifier to classify hourly data under different operating conditions into the energy consumption model with the smallest prediction error. A multiple variable regression model and an artificial neural network (ANN) model were compared with the models developed in the present study, and the normalized mean bias error (NMBE) and the coefficient of variation of the root mean squared error (Cv-RMSE) were used as criteria for the predicted energy consumption values.

1 Introduction

According to data from the International Energy Agency (IEA), the annual average growth rate of global energy demand from 2011 to 2035 is predicted to be 2.5%. The buildings sector is the largest energy-consuming sector, accounting for approximately 33% of the global energy consumption. Many studies have shown that heating ventilation and air conditioning (HVAC) is responsible for approximately 50% of the total energy consumption of buildings (Lombard et al. 2008; U.S. Energy Information Administration 2012; Office of Energy Efficiency & Renewable Energy (EERE) 2012), therefore, the development of energy-saving control technologies for HVAC systems has become extremely important. The steady rise of global temperatures in recent years and the utilization of HVAC systems in office buildings to maintain thermal comfort in the indoor environment have led to a continuous increase in HVAC power consumption. Many studies on energy control have been focused on the construction of HVAC power consumption models for the estimation of minimum energy consumption and subsequent control of HVAC operating parameters to achieve energy savings. (Lee et al. 2011) developed a chiller energy consumption model using part load ratios (PLR) for the estimation of minimum energy consumption and employed a differential evolution algorithm to obtain the optimal model, which was used to solve the optimal chiller loading (OCL) problem for energy conservation. (Chang 2004) used PLR to build a nonlinear regression model of chiller efficiency to achieve optimal control of chiller load. In another study, (Chang et al. 2005) constructed kW-PLR curves for air-

✉ Chang-Ming Lin
linbarry@itri.org.tw

[1] Institute of Electrical and Control Engineering, National Chiao Tung University, No. 1001 University Road, Hsinchu 30010, Taiwan, ROC

[2] Energy and Environment Research Laboratories, Industrial Technology Research Institute, Rm.820, Bldg.51, 8F, 195 Sec.4, Chung Hsing Rd, Chutung, Hsinchu 31040, Taiwan, ROC

conditioning systems and employed a genetic algorithm (GA) to solve the OCL problem. In a multi-chiller system, the best operation occurs when the total energy consumption of the chillers with respect to the demanded load is minimized. (Salari and Askarzadeh 2015) generated a model based on the relationship between centrifugal chiller energy consumption and PLR to calculate the minimum total energy consumption of multi-chiller systems, and used the General Algebraic Modeling System (GAMS) to solve the OCL problem. (Chen et al. 2014) used neural networks (NN) to build models of chiller power consumption for energy saving in chillers.

The energy consumption model can also be used as an energy baseline, which serves as a reference standard for the quantification of energy performance and the calculation of energy savings, thus providing a reference for the comparison of energy performance before and after the implementation of energy saving measures. The energy baseline, which is constructed by establishing relationships between variables that influence energy use, plays a critical role in the measurement and verification (M&V) process (Fig. 1). The majority of literature on the improvement of the control of building energy consumption utilized an energy baseline for the verification of the actual energy savings achieved with improvement measures. (Lee and Cheng 2012) compared the results of a simulation–optimization approach with an energy baseline to verify the overall energy savings of a chilled water system. In addition, the energy consumption model can be used for the prediction of the energy consumption of facilities to assist in the implementation of energy saving measures or energy management in buildings. Lei and Hu (2009) used meteorological data as variables for the construction of a regression model and compared the prediction accuracy of a simple linear regression model and a multiple variable regression model. From the results of the study, it was found that the monthly average outdoor dry-bulb temperature was the most important variable that affected model accuracy, and a simple linear regression model was sufficient for the simulation of energy use. Manjarres et al. (2017) proposed an optimal energy-efficient predictive control framework to achieve the minimization of HVAC power consumption, and compared energy savings through the use of an energy baseline. Kissock and Kelly (1993) used four weather parameters to construct energy consumption models for the prediction of energy use in commercial buildings. Carpenter et al. (2018) used change-point and Gaussian process models to create baseline energy models in industrial facilities, and compared the verification results of two energy consumption models in accordance with the ASHRAE Guideline 14 requirements on the normalized mean bias error (NMBE) and coefficient of variation of the root mean square error (Cv-RMSE). Kissock and Eger (2008) proposed the application of multivariable change-point models in the estimation of energy savings in industrial facilities, as the models can reduce the effects of actual temperature changes on the model-predicted values. Amiri et al. (2015) used stepwise analysis to identify the most effective variables and developed a multiple variable regression model to predict the energy consumption of commercial buildings. The selection of variables is extremely important in the creation of baseline energy models. Mustapa et al. (2017) used the simple regression method and multiple variable regression method with different variables to develop models for educational buildings to identify the variables with the greatest effect on energy consumption in educational buildings, and compared the mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) of the constructed models. Regardless of whether the energy consumption model is used in the estimation of the minimum energy consumption in energy-saving control or as an energy baseline for the comparison of energy performance, the prediction accuracy of the model is extremely critical. The accuracy of energy consumption models is influenced by the model variables; however, different operating periods and load conditions also result in changes in energy consumption. The use of a single energy consumption model for the dynamic simulation of energy consumption at different working levels will result in excessively large errors in the estimated energy consumption values. Ko et al. (2017) proposed a cluster

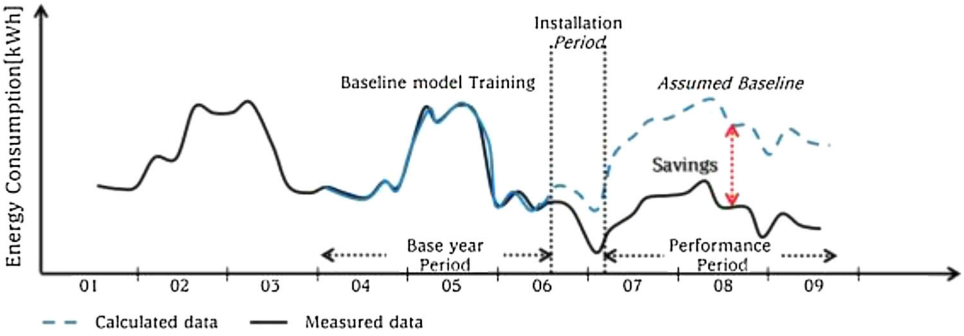


Fig. 1 Illustration of verification M&V for retrofitting

inverse model, which involved the clustering of daily data using the K-means clustering algorithm and the subsequent use of the clustered data to create linear models of temperature and energy consumption for the estimation of energy consumption in office buildings. Tnag et al. (2014) used a clustering algorithm and regression analysis to predict HVAC power consumption, and subsequently achieved the enhancement of model accuracy through data mining and machine learning methods.

Data mining is the process of discovering knowledge from data and the utilization of the discovered knowledge to describe the clusters, rules, and associations within the data. Bayindir et al. (2017) applied the Naïve Bayes classifier to the prediction of daily total photovoltaic energy generation by using daily average temperature, daily total sunshine duration, and daily total global solar radiation as input parameters. Granderson et al. (2016) expands recent analyses of public-domain whole-building M&V methods, focusing on more novel M&V 2.0 modeling approaches that are used in commercial technologies and present a testing procedure and metrics to assess the performance of whole-building M&V methods. Hong et al. (2013) proposed a modern approach that takes advantage of hourly information to create more accurate and defensible forecasts with several MLR models. The paper showed the predictive models attained from hourly data, over the classical methods of forecasting using monthly or annual peak data.

The development of energy consumption models proposed in the present study involved the clustering of hourly data using the K-means clustering algorithm, consolidation of the clustered hourly data into monthly average data by month, and the construction of energy consumption models using the simple linear regression method, so as to achieve regression models with high reliability. In addition, the Naïve Bayes classifier was used to classify data obtained under different operating conditions into the energy consumption model with the smallest prediction error. Therefore, the present study aimed to enhance the accuracy and reduce the prediction error of energy consumption prediction through the integration of the K-means clustering algorithm and the Naïve Bayes classifier.

2 Research background

The main objective of the present study was to set up a conditional probability model using a data mining method, the Naïve Bayes classifier, and to calculate energy consumption models corresponding to the data clusters using a set of attributes and Baye's theorem, so as to reduce the error in energy consumption prediction. The process flow chart of the methods used in the present study is shown in Fig. 2. Outdoor temperature data and HVAC power consumption data were clustered using the K-means clustering algorithm, and the energy consumption models of the respective data clusters were constructed using the simple linear regression method. Calendar data, outdoor temperature data, and HVAC on/off status were used as attributes for the classification of the hourly data into the corresponding HVAC power consumption model. The detailed methods used will be described in the following subsections.

3 Methodology

3.1 Data collection and variable selection

With HVAC power consumption data collected by digital meters at 3-min intervals, hourly energy consumption data were obtained by calculating the average of every 20 data points. Similarly, outdoor weather data, including outdoor temperature, outdoor humidity, wind speed, sunshine duration, and rainfall, are collected by meteorological stations at 3-min intervals, and hourly weather data were subsequently obtained by calculating the average of every 20 weather data points. HVAC power consumption increases with outdoor temperature, and the HVAC on/off status of many office buildings is distinctly different between non-working hours on non-working days and working hours on working days. In the present study, the hourly outdoor temperature data and hourly HVAC power consumption from May to Oct 2015 were compiled into a training dataset, while the hourly outdoor temperature data and hourly HVAC power consumption data during working hours from May to Oct 2016 were compiled into a testing dataset. The training and testing datasets consisted of hourly outdoor temperature data, hourly HVAC power consumption data, and calendar data, with the units of hourly outdoor temperature data and hourly HVAC power consumption data being °C and kWh, respectively.

3.2 K-means clustering algorithm

The K-means clustering algorithm, which is a partitioning-based clustering method, is one of the most widely applied methods in cluster analysis. It involves the organization of data objects into several mutually exclusive clusters, which are used to satisfy an optimal objective partitioning criterion. During the initialization of the partitioning algorithm, the number of clusters is first selected, and partitioning quality is assessed using an objective function to ensure mutual similarity among data objects within the same cluster and mutual dissimilarity among data objects of different clusters.

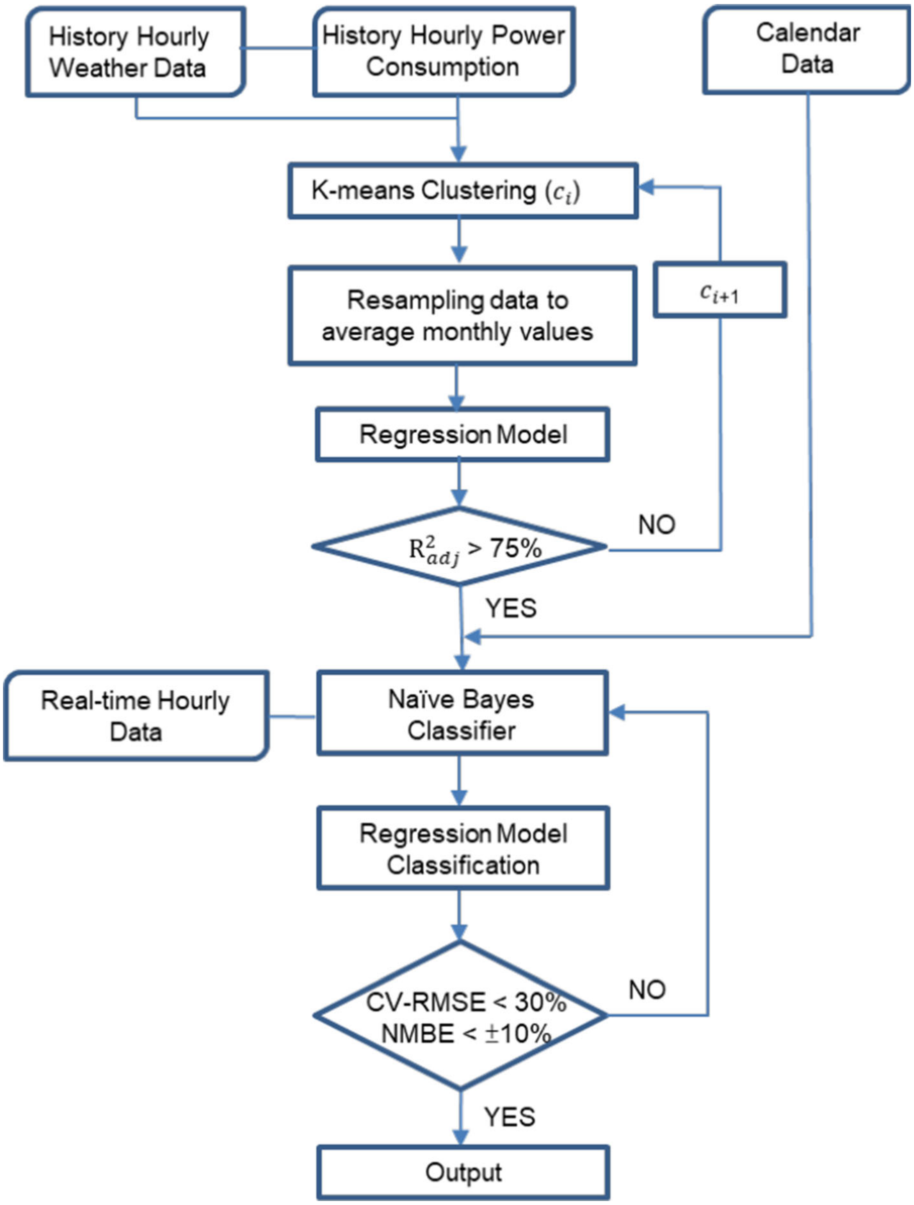

History Hourly Weather Data
History Hourly Power Consumption
Calendar Data
K-means Clustering (c_i)
Resampling data to average monthly values
c_{i+1}
Regression Model
R^2_{adj} > 75%
NO
YES
Real-time Hourly Data
Naïve Bayes Classifier
Regression Model Classification
CV-RMSE < 30%
NMBE < ±10%
NO
YES
Output

Fig. 2 Flowchart of the cluster and classifier method

The K-means clustering algorithm is a centroid-based partitioning-based clustering method whereby each cluster is represented by its centroid, which is also the mean of the cluster. Given a dataset S consisting of Q data points: $S = \{x_i | 1 \le i \le Q, x_i \in R\}$, K-means clustering aims to partition the data in S into k clusters to generate a cluster set $K = \{c_i | 1 \le i \le k\}$, where c_i represents the i-th cluster of data points. Additionally, the clusters are mutually exclusive, i.e. $c_i \cap c_j = \emptyset$, and $1 \le i, j \le k, i \ne j$. The union of all clusters is equivalent to the dataset S, as shown in Eq. (1):

$$S = \bigcup_{i=1}^{k} c_i \tag{1}$$

With the K-means clustering algorithm, the centroid z_i of each cluster c_i is initially generated, with the initial value of z_i determined by random sampling from the dataset S. Subsequently, the two steps described below are repeatedly executed until the data points within all clusters no longer change or until a specific termination condition is fulfilled.

3.2.1 Cluster allocation

Each data point x_i within S is allocated to a cluster c_j by locating the centroid that is closest to x_i. The measure of closeness is the Euclidean distance D, and the closest c_j to x_i is determined based on $D\left(x_i, c_j\right) \le D(x_i, c_l)$, where $1 \le j, l \le k, j \ne l$. The Euclidean distance between any two points $a = [a_1, a_2, \ldots, a_d]^T$ and $b = [b_1, b_2, \ldots, b_d]^T$ is calculated using Eq. (2):

$$D(a,b) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2} \tag{2}$$

3.2.2 Centroid relocation

$c_j = \{x_i^{cj} | 1 \leq i \leq n_{cj}\}$, where x_i^{cj} represents the i-th data point of cluster c_j and n_{cj} represents the number of data points of cluster c_j. The updated centroid c_j is calculated using Eq. (3):

$$c_j = \frac{\sum_{i=1}^{n_{cj}} x_i^{cj}}{n_{cj}} \tag{3}$$

3.3 Regression model

The linear regression model investigates the linear relationship between one or more independent variables (W) and a dependent variable (Y), and the values of the dependent variable are predicted based on the independent variables through the construction of an appropriate mathematical equation. The mathematical equation that relates the independent variables with the dependent variables is known as a regression model. The simple linear regression model is as follows:

$$Y_i = \alpha + \beta W_i + \varepsilon_i i = 1, 2, \ldots, n \tag{4}$$

where Y_i is the observed value, α is the intercept, β is the slope, and ε_i is the random error. α and β are known as the regression coefficients.

One of the most common methods used for the estimation of regression coefficients is the ordinary least squares method (OLS). The estimated regression model is as follows:

$$\hat{Y_i} = \hat{\alpha} + \hat{\beta} W_i i = 1, 2, \ldots, n \tag{5}$$

where $\hat{Y_l}$ is the estimated value of each observed value Y_i, and $\hat{\alpha}$ and $\bar{\beta}$ represent the estimated values of α and β, respectively.

Residual e_i is defined as the difference between the observed value and fitted value of the i-th data point, i.e. $e_i = Y_i - \hat{Y_i}$. The residual sum of squares (RSS) is used as the criterion to assess the fit of the model, with a smaller RSS indicating a better fit of the linear regression model to the observed value. The RSS is calculated using Eq. (6):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y_i}\right)^2 \sum_{i=1}^{n}[Y_i - (\alpha + \beta W_i)]^2 \tag{6}$$

The coefficient of determination (R^2) is a measure of the goodness of fit and explanatory power of the regression model, and is calculated using Eq. (7):

$$R^2 = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2} \tag{7}$$

The R^2 value ranges from 0 to 1, with a higher value indicating a higher explanatory power of the model for the variance in the dependent variable. As R^2 increases with increased model complexity, the explanatory power of high-complexity models for variance is usually overestimated. To address this concern, the adjusted R^2 (R^2_{adj}) was developed. The R^2_{adj} value is calculated as follows:

$$R^2_{adj} = 1 - \frac{\sum \frac{\left(Y - \hat{Y}\right)^2}{(n-1)}}{\sum \frac{\left(Y - \bar{Y}\right)^2}{(n-p)}} \tag{8}$$

where n is the number of points in the data sample, and p is the number of independent variables.

3.4 The proposed Naïve Bayes application

The Naïve Bayes classifier is a probabilistic model based on the analysis of relations between attributes and response variables in data. It is used for the classification of sample data through the application of Bayes' theorem to update probabilities as new information is acquired, and serves as a basis for classification and inference. The Bayes' theorem is described below [19], where P(A\B) is the probability of event A in case B occurs, P(B\A) is the probability of event B in case A occurs, P(A) is the probability of A and P(B) is the probability of B.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{9}$$

In particular, the Naïve Bayes classifier works as follows:

Each instance in the learning set D is defined by an n-dimensional attribute vector $X = (x_1, x_2, \cdots, x_n)$. Assume that m categories $C = (C_1, C_2, \ldots, C_m,)$ exist, and the naïve Bayes classifier predicts that X belongs to the category of maximum posterior probability if—and only if—the following conditions are met:

$$P(C_i|X) > P\left(C_j|X\right) 1 \leq j \leq m, j \neq i \tag{10}$$

According to (9), the maximized posterior probability $P(C_i|X)$ yields the following equation:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{11}$$

Because $P(X)$ is a constant, only the maximum value of $P(X|C_i)P(C_i)$ must be determined. If the prior probabilities of the categories are unknown, each category is generally assumed to possess the same prior probability $P(C_1) = P(C_2) = P(C_3) = \ldots P(C_m)$. Therefore, only the maximum value of $P(X|C_i)$ must be determined. Assuming that the attribute relationships in each category are independent of each other, $P(X|Ci)$ can be estimated through (9).

$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^{n} P(x_k|C_i) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i) \end{aligned} \tag{12}$$

Finally, the class label of X is predicted using the following equation:

$$P(X|C_i)P(C_i) > P\left(X|C_j\right)P\left(C_j\right) for 1 \leq j \leq m, j \neq i \tag{13}$$

3.5 Model verification criteria

In accordance with the requirements and recommendations of ASHRAE Guideline 14 and the FEMP's M&V Guideline 4.0, energy consumption models must be verified to ensure that the statistical results satisfy acceptable criteria. In the present study, the normalized mean bias error (NMBE) and the coefficient of variation of the root mean squared error (Cv-RMSE) were used to correct the monthly and hourly data of the model. Smaller NMBE and Cv-RMSE values indicate higher prediction accuracy.

The NMBE is used to estimate the deviation between predicted and actual measured values. A positive NMBE value indicates that the actual measured value is higher than the predicted value, while a negative NMBE value indicates that the measured value is lower than the predicted value. The NMBE is computed according to Eq. (14):

$$NMBE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)}{(n-p) \times \bar{Y}} \times 100 \tag{14}$$

The root mean squared error (RMSE) is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}} \tag{15}$$

For a multiple linear regression model, RMSE is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_i)^2}{n-p}} \tag{16}$$

The coefficient of variation of the root mean squared error (Cv-RMSE) indicates the uncertainty inherent in the model, which is computed as follow:

$$Cv - RMSE = 100 \times \frac{RMSE}{\bar{Y}} \tag{17}$$

The required values are dependent of data sampling frequency as listed in Table 1. ASHRAE Guideline 14 only provides requirements for monthly and hourly models.

4 Case study and results

4.1 Test site description

The building selected in the present study was a four-story office building with a floor area of approximately 3495 m^2. The operating periods of HVAC in the building were distinctly divided into non-working days and working days, while the load conditions were distinctly different between summer and non-summer seasons. HVAC power consumption accounted for approximately 50% of the total power consumption of the building, and the peak period of power consumption occurred during working hours (8 am to 6 pm) from May to October each year. The HVAC system of the building, which supplied cold air to the entire building, consisted of one 90 RT variable speed chiller, one 15 HP chilled water pump, one 15 HP cooling water pump and one 5 HP cooling tower.

The monthly average HVAC power consumption of the office building in 2015 is shown in Fig. 3. Within the year, monthly HVAC power consumption showed a peak in the summer seasons of June and July and gradual decreases on both sides of the peak (preceding and subsequent months). In Fig. 4, which shows the monthly average outdoor temperature in 2015, it can be seen that the highest outdoor temperatures were also concentrated in the months of June and July. Therefore, these results indicate a definite link between HVAC power consumption and outdoor temperature.

Table 1 Required values for the baseline model according to ASHRAE Guideline 14

	Monthly (%)	Hourly (%)
NMBE	5	10
Cv-RMSE	15	30

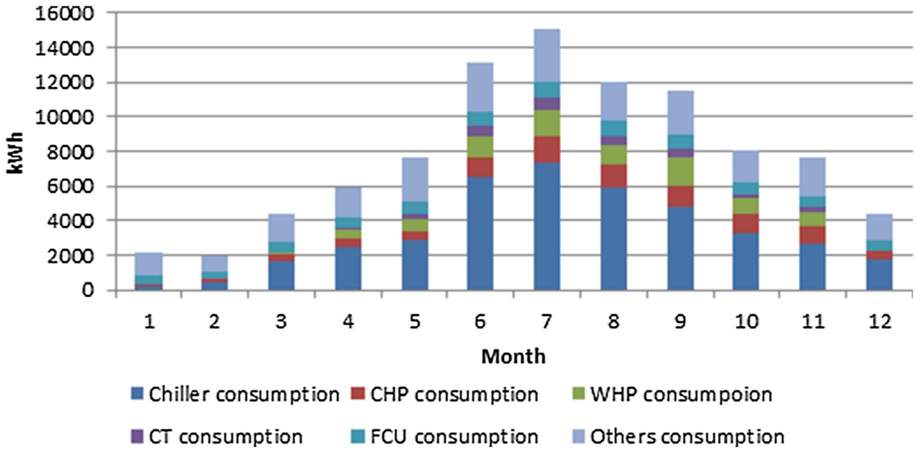


Fig. 3 Bar chart of monthly HVAC consumption of the B1 office building in Hsinchu, Taiwan

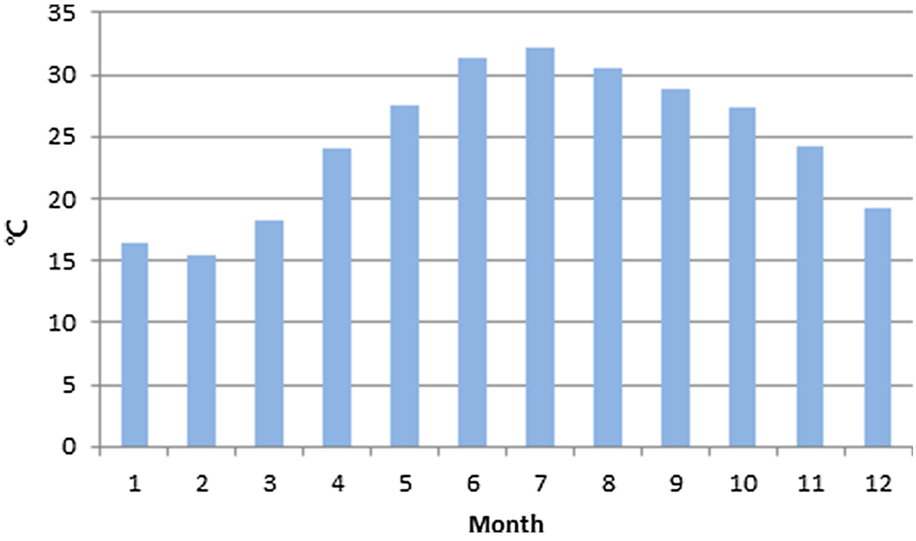


Fig. 4 Bar chart of monthly outdoor temperatures in Hsinchu (2015)

4.2 Training of the clustering algorithm

In the present study, the K-means clustering algorithm was used for the clustering of outdoor temperature and HVAC power consumption data, so as to identify the HVAC power consumption data corresponding to different outdoor temperatures and organize similar data into clusters. The clustered hourly data were subsequently converted to monthly average data by month, and energy consumption models were constructed using the simple linear regression method. Lastly, the Naïve Bayes classifier was used to classify data obtained under different operating conditions into the energy consumption model with minimum prediction error.

The HVAC power consumption models were developed using the training dataset, which consisted of monthly average HVAC power consumption data and monthly average outdoor temperature data from May to Oct 2015. The constructed models were then used to simulate the HVAC power consumption data during the working hours of May to Oct 2016, which were subsequently compared with the HVAC power consumption measurement data during the same period.

During the initialization of the K-means clustering algorithm, the number of clusters must first be decided. In the present study, the adjusted coefficient of determination (R^2_{adj}) for the regression model was used as the criterion for determining the number of clusters k for the K-means clustering algorithm. If $R^2_{adj} < 75\%$, the number of clusters was increased and reclustering was performed until $R^2_{adj} > 75\%$ (Fig. 2). The number of clusters was set as k = 3, k = 4, and k = 5 for comparison and the respective R^2_{adj} values were calculated, as shown in Table 2. Results indicated that $R^2_{adj} > 75\%$ was achieved for 2 clusters when k = 3, therefore the optimal k value of 3 was used in the case study.

The data of the training dataset were normalized to avoid the situation whereby direct comparison could not be

Table 2 Comparison of the coefficients of determination (R^2_{adj}) of the regression models

	c_1 (%)	c_2 (%)	c_3 (%)	c_4 (%)	c_5 (%)
$k = 3$	31.8	84	83.2	–	–
$k = 4$	34	5	59.4	18.7	–
$k = 5$	− 1.6	59.3	70.4	36.9	11.3

performed due to different data sizes. Normalization was performed using Eq. (18):

$$x_i' = \frac{x_i - \bar{x}}{S_x} \tag{18}$$

where x_i is the i-th data point,x_i' is the normalized value of the i-th data point, $\bar{x}$ is the mean value of the data point, and S_x is the standard deviation of the data point.

Using the K-means clustering algorithm, the normalized data were divided into three clusters $c_1, c_2, \mathrm{and} c_3$, with the x-axis being outdoor temperature and y-axis being HVAC power consumption, as shown in Fig. 5.

4.3 Validation of the clustering algorithm

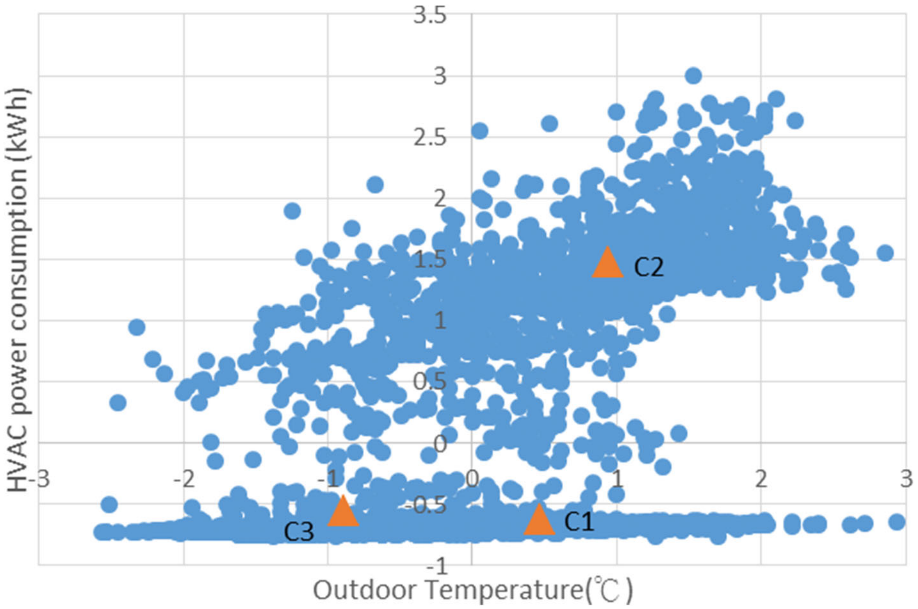
Energy consumption models were constructed for each cluster by performing simple linear regression. Each model represents the relationship between outdoor temperature and HVAC power consumption data within the cluster, and was established by converting the hourly HVAC power consumption data and hourly outdoor temperature data into monthly average data by month and constructing a simple linear regression model. Figures 6, 7, 8 indicate the energy consumption models for the respective clusters of outdoor temperature and HVAC power consumption data and the adjusted coefficients of determination R^2_{adj}. In particular, the R^2_{adj} of the second (c_2) and third (c_3) clusters were 84% and 83.2%, respectively, while the R^2_{adj} of the first cluster (c_1) was only 31%. This is due to the fact that the centroid of c_1 was representative of the median outdoor temperature and low HVAC power consumption, c_2 consisted of high outdoor temperature and high HVAC power consumption data points, while c_3 consisted of low outdoor temperature and low HVAC power consumption data points. In addition, the data of c_1 were mostly data obtained during non-working hours, therefore the average outdoor temperatures ranged between 20 to 25 °C and the HVAC system was not in operation.

4.4 Training of the Naïve Bayes classifier

The outdoor temperature data of the training dataset were respectively input into the three HVAC power consumption models for the calculation of three sets of simulated HVAC power consumption values. Subsequently, for each data point, the residual between the simulated and measured HVAC power consumption values was calculated and the regression model with the smallest residual was set as the HVAC power consumption model for classification using the Naïve Bayes classifier.

The training dataset consisted of hourly HVAC power consumption data and hourly outdoor temperature data with a time format, which amounted to a total of 3995 data points. In the present study, three parameters, namely calendar data (including month, week, and time), outdoor temperature and on/off status of the HVAC system, were

Fig. 5 Scatterplot of clusters related to HVAC consumption



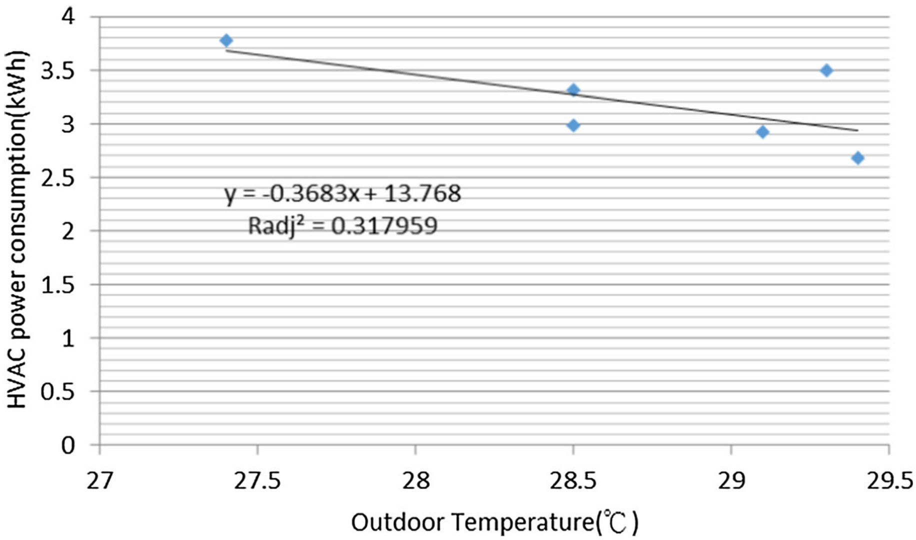


Fig. 6 Correlation between energy consumption and outdoor temperature of C1

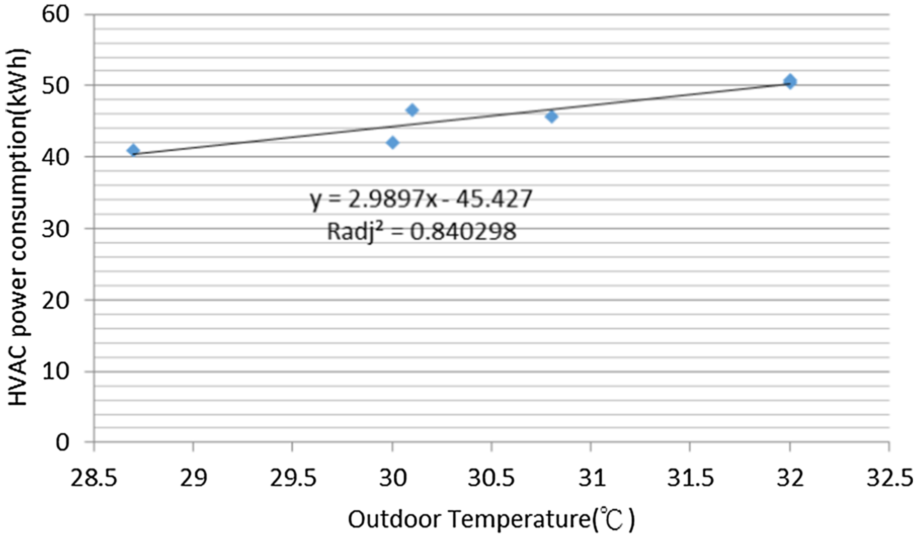


Fig. 7 Correlation between energy consumption and outdoor temperature of C2

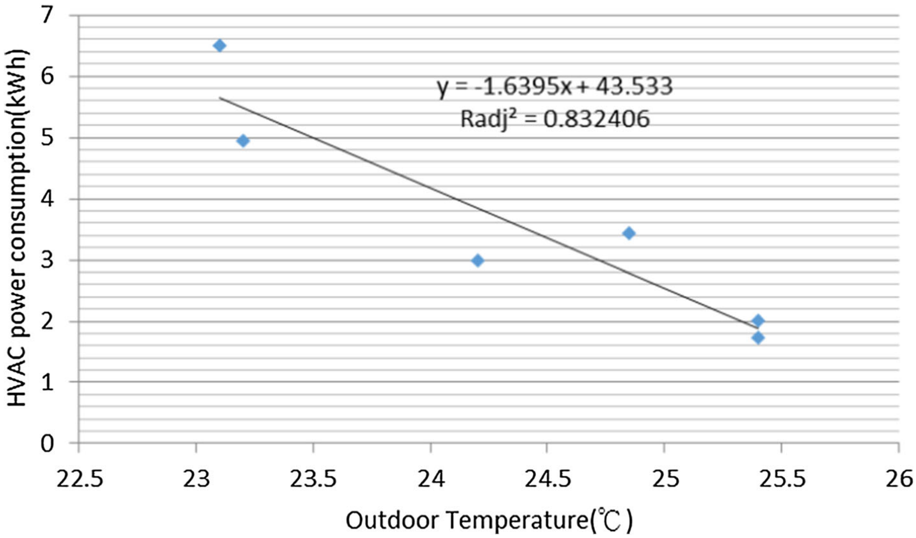


Fig. 8 Correlation between energy consumption and outdoor temperature of C3

Table 3 Value ranges corresponding to each label and parameter

Label	Parameter	Attributes	Description
1	Outside Temperature	A	0–5 °C
2			5–10 °C
3			10–15 °C
4			15–20 °C
5			20–25 °C
6			25–30 °C
7			30–35 °C
8			35–40 °C
1	Month	B	May
2			June
3			July
4			August
5			September
6			October
1	Week	C	Weekday
2			Weekend
1	Time of day	D	else
2			08:00 ~ 18:00
0	On/Off	E	Off
1			On

Table 4 Conditional probability distribution for the naïve Bayes application using the training dataset

	C1	*C2*	*C3*
$P(Ci)$	0.479599	0.331414	0.188986
$P(A1\vert Ci)$	0	0	0
$P(A2\vert Ci)$	0	0	0
$P(A3\vert Ci)$	0	0	0
$P(A4\vert Ci)$	0.017223	0.001511	0.003974
$P(A5\vert Ci)$	0.499478	0.108761	0.103311
$P(A6\vert Ci)$	0.3262	0.327795	0.892715
$P(A7\vert Ci)$	0.152401	0.535498	0
$P(A8\vert Ci)$	0.004697	0.026435	0
$P(B1\vert Ci)$	0.209812	0.132175	0.133775
$P(B2\vert Ci)$	0.150835	0.147281	0.150993
$P(B3\vert Ci)$	0.157098	0.202417	0.203974
$P(B4\vert Ci)$	0.136221	0.185801	0.264901
$P(B5\vert Ci)$	0.186326	0.179003	0.149669
$P(B6\vert Ci)$	0.159708	0.153323	0.096689
$P(C1\vert Ci)$	0.570459	1	0.670199
$P(C2\vert Ci)$	0.429541	0	0.329801
$P(D1\vert Ci)$	0.749478	0.059668	0.863576
$P(D2\vert Ci)$	0.250522	0.940332	0.136424
$P(E0\vert Ci)$	0.981733	0.003021	0.993377
$P(E1\vert Ci)$	0.018267	0.996979	0.006623

classified into 20 labels and 5 attributes (A, B, C, D, and E). Table 3 shows the value range and type of status of each label. The HVAC power consumption models of the three clusters were set as three categories $c_1, c_2, \mathrm{and} c_3$, and the data points of the training dataset were used to calculate the prior probability $P(Ci)$ of the three categories and the conditional probability $P(X|C_i)$ of each attribute. The conditional probability distribution of the categories and attributes is shown in Table 4. The respective probabilities that each hourly data point of the training dataset belonged to each category were calculated by substituting the probability values of the different attributes in Table 4 into Eq. (12). Subsequently, each data point of the training set was classified into the corresponding energy consumption model according to the maximum value calculated by Eq. (13).

4.5 Validation of the Naïve Bayes classifier

Results indicated that the sensitivity of the Naïve Bayes classifier in correctly predicting the HVAC power consumption models for the training dataset was 82.703%. Sensitivity indicates the rate of correct prediction and is calculated using the following equation:

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{19}$$

where: TP: true positives where both the actual and the software-predicted values are positive, i.e. correct prediction

FN: false negatives where the actual value is positive but the software-predicted value is negative, i.e. incorrect prediction. This represents a misjudgment of a correct result.

4.6 Validation of the proposed algorithm

The testing dataset was used to assess the accuracy of the HVAC power consumption models. Each data point of the testing dataset was classified into a HVAC power consumption model based on the probability calculated using the Naïve Bayes classifier, and the hourly outdoor temperature data of the testing dataset were input into the corresponding HVAC power consumption model to obtain the predicted hourly HVAC power consumption values. The normalized mean bias error (NMBE) and the coefficient of variation of the root mean squared error (Cv-RMSE) between the predicted hourly HVAC power consumption and the hourly HVAC power consumption data of the verification dataset were then compared.

4.7 Comparsion with other methods

Multiple variable regression and artificial neural network (ANN) are two methods commonly used to construct energy consumption models. As such, the energy consumption models established with the energy modeling application of the Naïve Bayes classifier for accuracy enhancement in the present study were compared with multiple variable regression and ANN models to assess the differences in accuracy of HVAC power consumption prediction. The multiple variable regression model was constructed with monthly average outdoor temperature and monthly average outdoor humidity as independent variables and monthly average HVAC power consumption as the dependent variable. For the ANN model, a feedforward network was adopted, which included an input layer, two hidden layers, and an output layer, and the only output of the model was predicted HVAC power consumption. The energy consumption models proposed in the present study were constructed by dividing hourly data into three clusters using the K-means clustering algorithm, converting the hourly data in each cluster into monthly average data, and performing simple linear regression. Subsequently, the Naïve Bayes classifier was applied to classify data obtained under different operating conditions to the energy consumption model with the minimum prediction error. Figures 9, 10, 11 show the comparison of the predicted and measured hourly HVAC power consumption data for the three methods. The data used for comparison was the testing dataset, which consisted of hourly HVAC power consumption data during the working hours of May to October 2016. Figure 9 shows the comparison for the multiple variable regression model, while Figs. 10 and 11 show the comparison for the ANN model and energy consumption models with the application of the Naïve Bayes classifier, respectively. The months at the left and right ends of Figs. 9 to 11 are May and October, respectively, which experienced greater fluctuations in outdoor temperature. Therefore, the HVAC power consumption of these two months exhibited dense, jagged patterns with large fluctuations. In contrast, the middle portions of the figures show data obtained from June to September, which are months with high outdoor temperatures, therefore the HVAC power consumption exhibited a more stable pattern with smaller fluctuations. With the multiple variable regression method, a mathematical relation of outdoor temperature and outdoor humidity with HVAC power consumption was established, and an R_{adj}^2 value of 92.6% was achieved. However, the predicted HVAC power consumption values in Fig. 9 were calculated solely based on outdoor weather data, therefore low HVAC power consumption during off-peak periods or non-working days could not be effectively estimated. In addition, high outdoor temperatures usually resulted in the overestimation of power consumption as the load condition of the HVAC system was not considered in the prediction process. Compared with the multiple variable regression model, the ANN model in Fig. 10 and energy consumption models with the application of the Naïve Bayes classifier in Fig. 11 were both able to achieve predicted power consumption

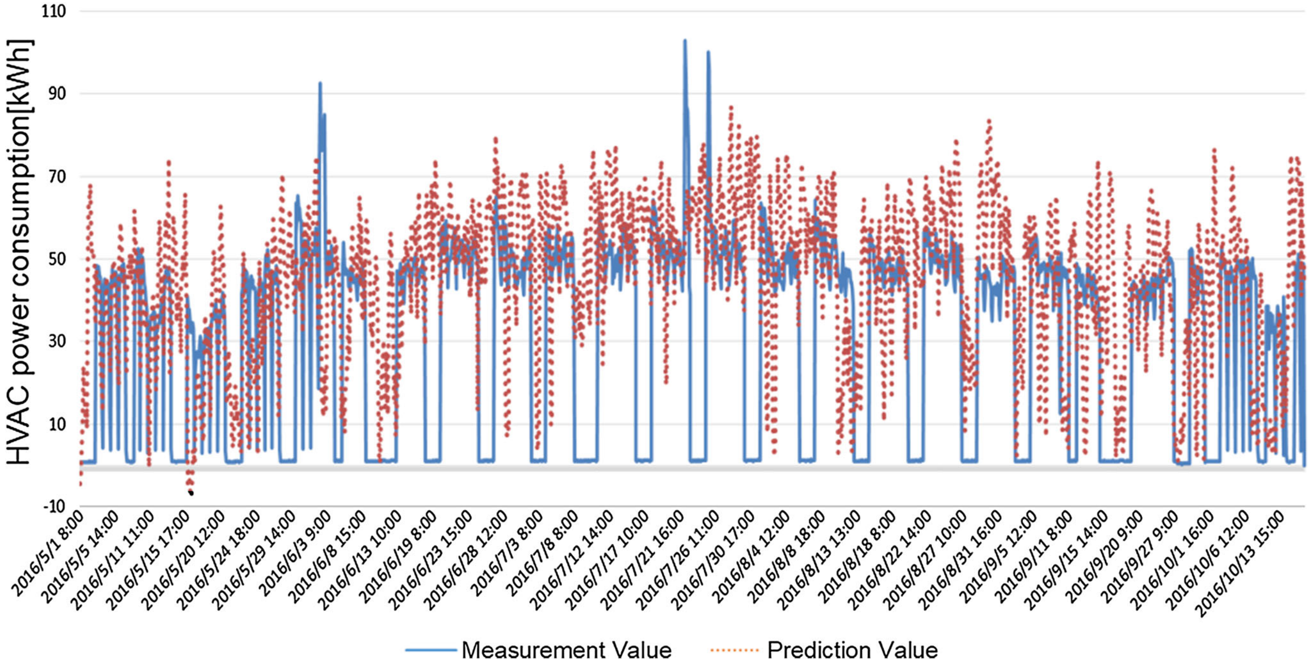


Fig. 9 Comparisons of the baseline model predictions for the multivariate regression model with the measured HVAC power consumption data

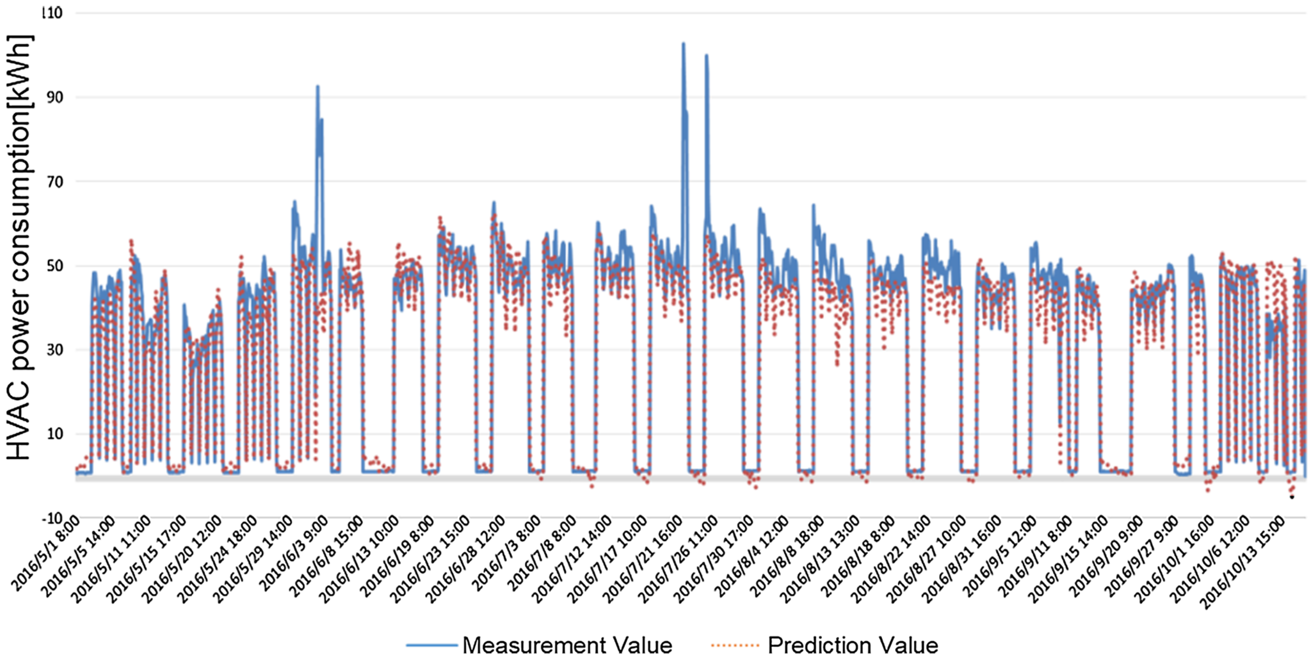


Fig. 10 Comparisons of the baseline model predictions for the ANN with the measured HVAC power consumption data

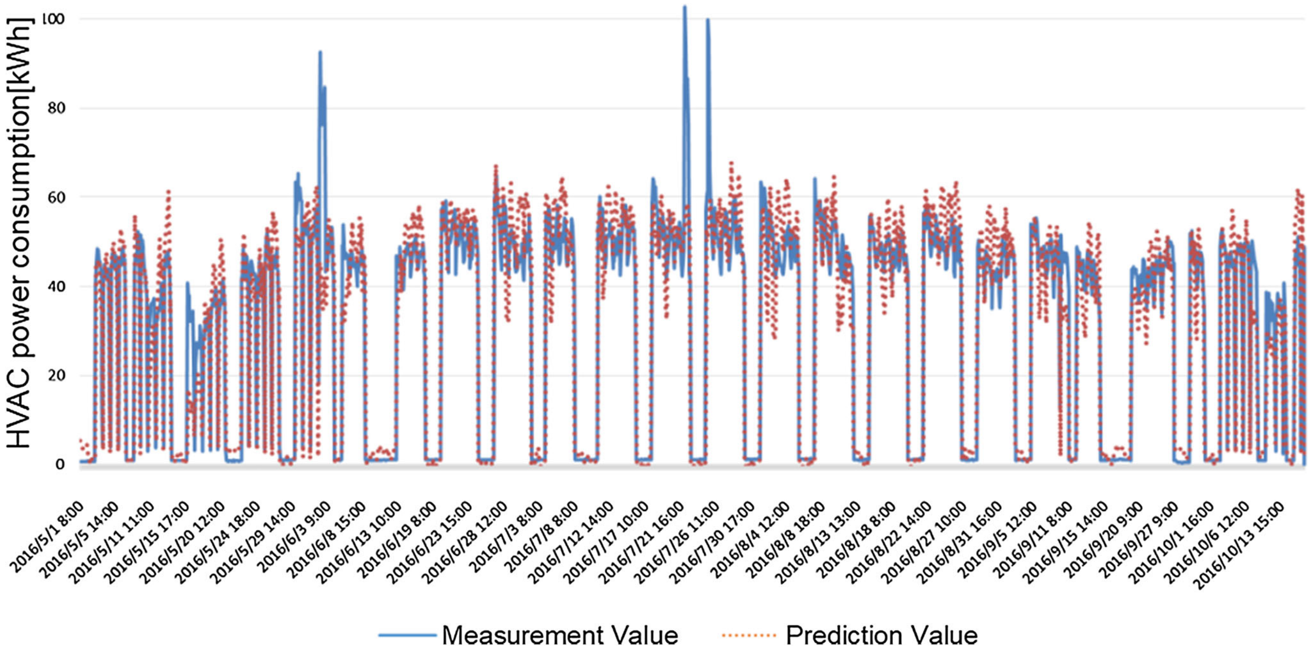


Fig. 11 Comparisons of the baseline model predictions for the proposed method with the measured HVAC power consumption data

values that were close to the actual measured values. The inputs to the ANN model were the attributes used in the Naïve Bayes classifier defined in the present study, i.e. outdoor temperature, calendar data, and on/off status of the chiller system, while the output was the predicted HVAC power consumption. For Fig. 11, the Naïve Bayes classifier was applied to classify each data point into the corresponding energy model according to the attributes of outdoor temperature data, calendar data and the on/off status of the HVAC system. With this method, HVAC power consumption during different operating periods and load conditions can be effectively predicted, thus resulting in smaller prediction errors. The comparison of the R^2_{adj}, NMBE and Cv-RMSE values of the multiple variable regression model, ANN model and energy consumption

Table 5 Statistical metrics of the multiple variable regression, ANN, and naïve Bayes classifier to classify power consumption models

	Multiple variable regression (%)	ANN (%)	Energy modeling application of Naïve Bayes classifier (%)
R^2_{adj}	92.6	–	31.8/84/83.2
NMBE	− 41.43	− 5.29	0.73
Cv-RMSE	93.2	19.54	22.36

models with the application of the Naïve Bayes classifier is shown in Table 5.

Table 5 shows the comparison of statistical metrics among different models. Based on the statistical performance of the three methods, we can conclude that the Naïve Bayes application inverse model has slightly better statistical performance in terms of the lowest NMBE compared to the other two methods. Although ANN can well predict the power consumption and has lower Cv-RMSE, it also has high model complexity, poor extrapolation possibility, and difficulty in calculating uncertainty. For the energy consumption models with the application of the Naïve Bayes classifier for prediction accuracy enhancement, the NMBE and Cv-RMSE of hourly data were 0.73% and 22.36%, respectively, which indicate that the energy consumption prediction results were superior to the requirements of ASHRAE Guideline 14, i.e. NMBE $<$ 10% and Cv-RMSE $<$ 30%. For the multiple variable regression model, although the R^2_{adj} value of 92.6% was the highest, the NMBE and Cv-RMSE values were − 41.43% and 93.2%, respectively, which indicate the model's ineffectiveness in predicting energy consumption and high uncertainty in the predicted data. For the ANN model, the NMBE and Cv-RMSE were 5.29% and 19.54%, respectively. In particular, the Cv-RMSE value was the smallest among the three methods, which shows that the energy consumption values predicted by the ANN model were lower than the actual values, however, the uncertainty of the predicted data was lower. The R^2_{adj} values of the energy consumption models corresponding to the three clusters obtained using the K-means clustering algorithm were 31.8, 84, and 83.2%, while the NMBE and Cv-RMSE were − 0.73% (smallest among the three methods) and 22.36%, respectively. These results show that the method proposed in the present study had the highest accuracy, i.e. the predicted data were the closest to the measured data, and the predicted data had low uncertainty, therefore the present study has provided a high-accuracy method for the prediction of building energy consumption.

5 Conclusions

The optimal energy-saving control or energy baseline of HVAC systems are dependent on an accurate energy consumption model for the prediction of power consumption. The use of hourly data to calibrate energy consumption models provides higher accuracy compared with the use of monthly or daily data, and facilitates discernment of the operating mode of the HVAC system. In the present study, an energy modeling application of the Naïve Bayes classifier to classify energy consumption models for the enhancement of prediction accuracy was proposed. Hourly outdoor temperature data and hourly HVAC power consumption data were divided into clusters using the K-means clustering algorithm, and the hourly data of the respective clusters were converted to monthly average data for the construction of energy consumption models using simple linear regression. The HVAC power consumption models of the clusters were set as the categories, and outdoor temperature, month, week, time, and on/off status of the HVAC system were set as the attributes for the calculation of conditional probabilities by the Naïve Bayes classifier, so as to obtain HVAC power consumption models with the smallest prediction error. Lastly, hourly data were used to increase the accuracy of the energy consumption models. The obtained models were subsequently compared with a multiple variable regression model and an ANN model. The NMBE of the multiple variable regression model and ANN model were − 41.43% and − 5.29%, respectively, while the NMBE and Cv-RMSE of the models developed in the present study were 0.73% and 22.36%, respectively, which indicated the fulfilment of the requirements of ASHRAE Guideline 14 as well as the substantial enhancement of the accuracy of power consumption prediction. With the application of the Naïve Bayes classifier, the operating trends of HVAC systems during different operating periods and load conditions can be shown through hourly data, and with the clustering of data, outlier power consumption values within the same operating period and load conditions can be identified for the diagnosis of power consumption abnormalities. Moreover, (Ballarini et al. 2014) presented a methodology for the identification of reference buildings, according to the IEE-TABULA project (2009–12) aimed at creating a harmonized structure for "European Building Typologies". The results show the enormous potentialities of energy savings even with basic energy retrofit actions. (Costanzo et al. 2018) reported he outcomes of a one-year measurement campaign of a passive house built in the Mediterranean climate of Cesena (Italy) in terms of thermal comfort parameters temperature and relative humidity and Indoor Environmental Quality (IEQ) parameter CO2 concentrations. The Passivhaus Standard can still be regarded as a good reference for designing low-energy and comfortable houses in a Mediterranean climate if some simplifications are made according to detailed building performance simulations. (McLeod et al. 2012) proposed a new method for the generation of current and future probabilistic micro regional climatic data in Passivhaus design, such data should provide a more robust basis for future cost and performance optimisation in low energy and passive building design. (Granderson et al. 2015) reported documents the application of a general statistical methodology to assess the accuracy of baseline energy models, focusing on its application to Measurement and Verification (M&V) of whole-building energy savings.

Low-cost energy saving and indoor comfort are the key directions for energy saving improvement in the future. In the future, the classification modeling method of this study can be combined with indoor temperature and outdoor temperature to establish an HVAC model that achieves indoor comfort. The HVAC system can be optimally adjusted according to different usage requirements inside the building, and the cooling air required for use space is supplied.

Acknowledgments We appreciate National Chiao Tung University for their generous assistance of computing time and Bureau of Energy, Ministry of Economic Affairs, R.O.C.(Taiwan) for the finical support.

References

Amiri SS, Mottahedi M, Asadi S (2015) Using multiple regression analysis to develop energy consumption indicators for commercial building in the U.S. Energy Build 109:209–216. https://doi.org/10.1016/j.enbuild.2015.09.073

Ballarini I, Corgnati SP, Corrado V (2014) Use of reference buildings to assess the energy saving potentials of the residential building stock: the experience of TABULA project. Energy Policy 273–284. https://doi.org/10.1016/j.enpol.2014.01.027

Bayindir R, Yesilbudak M, Colak M, Genc N (2017) A novel application of naïve bayes classifier in photovoltaic energy prediction. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/icmla.2017.0-108

Carpenter J, Woodbury KA, O'Neill Z (2018) Using change-point and Gaussian process models to create baseline energy models in industrial facilities: a comparison. Appl Energy 213:415–425. https://doi.org/10.1016/j.apenergy.2018.01.043

Chang Y-C (2004) A novel energy conservation method—optimal chiller loading. Electr Power Syst Res 69:221–226. https://doi.org/10.1016/j.epsr.2003.10.012

Chang Y-C, Lin J-K, Chuang M-H (2005) Optimal chiller loading by genetic algorithm for reducing energy consumption. Energy Build 37:147–155. https://doi.org/10.1016/j.enbuild.2004.06.002

Chen C-L, Chang Y-C, Chan T-S (2014) Applying smart models for energy saving in optimal chiller loading. Energy Build 68:364–371. https://doi.org/10.1016/j.enbuild.2013.04.030

Costanzo V, Fabbri K, Piraccini S (2018) Stressing the passive behavior of a Passivhaus: an evidence-based scenario analysis for a Mediterranean case study. Building and Environment 265–277. https://doi.org/10.1016/j.buildenv.2018.06.035

Granderson J, Touzani S, Custodio C, Sohn M, Fernandes S, Jump D (2015) Assessment of Automated Measurement and Verification (M&V) Methods. Lawrence Berkeley National Laboratory, LBNL-187225, 2015

Granderson J, Touzani S, Custodio CD, Sohn M, Jump D, Fernandes S (2016) Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings. Appl Energy 173:296–308. https://doi.org/10.1016/j.apenergy.2016.04.049

Hong T, Wilson J, Xie J (2013) Long term probabilistic load forecasting and normalization with hourly information. IEEE Trans Smart Grid 5(1):456–462. https://doi.org/10.1109/TSG.2013.2274373

Kissock JK, Eger C (2008) Measuring industrial energy savings. Appl Energy 85:347–361. https://doi.org/10.1016/j.apenergy.2007.06.020

Kissock, Kelly J (1993) A methodology to measure retrofit energy savings in commercial buildings. Doctoral dissertation. Dissertation. Texas A&M University

Ko J-H, Kong D-S, Hun J-H (2017) Baseline building energy modeling of cluster inverse model by using daily energy consumption in office building. Energy Build 140:317–323. https://doi.org/10.1016/j.enbuild.2017.01.086

Lee K-P, Cheng T-A (2012) A simulation–optimization approach for energy efficiency of chilled water system. Energy Build 54:290–296. https://doi.org/10.1016/j.enbuild.2012.06.028

Lee W-S, Chen Y-T, Kao Y (2011) Optimal chiller loading by differential evolution algorithm for reducing energy consumption. Energy Build 43:599–604. https://doi.org/10.1016/j.enbuild.2010.10.028

Lei F, Hu P (2009) A baseline model for office building energy consumption in hot summer and cold winter region. In: 2009 international conference on management and service science. https://doi.org/10.1109/icmss.2009.5301031

Lombard L-P, Ortiz J, Pout C (2008) A review on buildings energy consumption information. Energy Build 40:394–398. https://doi.org/10.1016/j.enbuild.2007.03.007

Manjarres D, Mera A, Perea E, Lejarazu A, Gil-Lopez S (2017) An energy-efficient predictive control for HVAC systems applied to tertiary buildings based on regression techniques. Energy Build 152:409–417. https://doi.org/10.1016/j.enbuild.2017.07.056

McLeod RS, Hopfe CJ, Rezgui Y (2012) A proposed method for generating high resolution current and future climate data for Passivhaus design. Energy Build 55:481–493. https://doi.org/10.1016/j.enbuild.2012.08.045

Mustapa R F, Dahlan N Y, Yassin I M, Mohd Nordin A H, Mahadan M E (2017) Baseline energy modeling in an educational building campus for measurement and verification. 2017 International Conference on Electrical, Electronics and System Engineering (ICEESE). https://doi.org/10.1109/iceese.2017.8298383

Office of Energy Efficiency & Renewable Energy (EERE) (2012) 2011 Building energy data book. https://openei.org/doeopendata/dataset/buildings-energy-data-book/resource/3edf59d2-32be-458b-bd4c-796b3e14bc65. Accessed Mar 2012

Salari E, Askarzadeh A (2015) A new solution for loading optimization of multi-chiller systems by general algebraic modeling system. Appl Therm Eng 84:429–436. https://doi.org/10.1016/j.enbuild.2015.03.057

Tnag F, Kusiak A, Wei X (2014) Modeling and short-term prediction of HVAC system with a clustering algorithm. Energy Build 82:310–321. https://doi.org/10.1016/j.enbuild.2014.07.037

U.S. Energy Information Administration (2012) Annual energy outlook early release Energy Information Administration (EIA). U.S. Department of Energy. https://www.eia.gov/outlooks/aeo/. Accessed Jun 2012