Energy & Buildings 194 (2019) 289–300

Contents lists available at ScienceDirect

Energy & Buildings

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

Comparative study of data driven methods in building electricity use prediction

Aaron Zeng[a,*], Sheng Liu[b], Yao Yu[c,*]

[a] United Technologies, Pudong New District, Shanghai, PR China
[b] School of Architecture, The Chinese University of Hong Kong, New Territories, Hong Kong, PR China
[c] Department of Construction Management and Engineering, North Dakota State University, Fargo, United States

ARTICLE INFO

Article history:
Received 1 January 2019
Revised 25 March 2019
Accepted 16 April 2019
Available online 18 April 2019

Keywords:
Energy prediction
Data driven
Electricity consumption
Big data
Comparative

ABSTRACT

The energy use prediction of building systems is crucial to design a high-efficiency building and maintain low energy consumption operation, which is also important in optimizing building system control and retrofitting. This paper demonstrates a comparative study of four data-driven methods used in online building energy predictions involving large-scale data extracted from several types of buildings. The characteristics of building electricity use and data reliability were addressed through the data pre-treatment process including visualization, cleaning, parsing and filtering. Mathematical algorithms and their applications in previous studies were summarized and compared, and evaluation methods were developed. The performance and suitable application scenarios of the proposed algorithms were conducted via the comparison of monitoring data and predicted results. The study indicates that the most complex method which requires the highest computation ability, i.e., the Artificial Neural Network (ANN), does not lead to the highest accuracy, while as the fastest computation method, Gaussian Process Regression (GPR) usually has the results with the lowest accuracy. Support Vector Machine (SVM) and Multivariate Linear Regression (MLR) methods usually perform better in the case scenarios studied. All the prediction accuracies can meet the requirements of RMSE <30% and NMBE <10% proposed by ASHRAE, and the computation time varies from less than 1 s to 22 s per prediction. All these methods/algorithms worked well for buildings with stable energy use patterns. For buildings with complex and unstable occupancy schedules and energy use patterns, MLR and SVM methods have the ability to achieve a high accuracy with fast computation speed.

1. Introduction

Building energy use prediction plays a crucial role in building energy management and performance improvement, which is paramount in building commissioning especially for energy saving estimation [1], fault detection and diagnosis, efficiency optimization, and smart building practice [2]. There are five main categories of energy use that can be applied in building energy use prediction: whole building or sub-metering electricity, heating energy, cooling energy, fossil fuel and others [3]. The available prediction methods can also be used to simulate building control systems, such as the model predictive control [4].

The energy prediction methods used in buildings can be clustered to three categories: (i) physical-based approaches; (ii) data-driven approaches; (iii) hybrid approaches that combine the first two methods [2,5]. The physical-based approach requires kernel physical components, thermal performance, and their corresponding numeric values, while data driven methods use purely historical data to predict energy performance of a building [6]. The hybrid model can also be treated as a grey-box model, which can use partial data and is quite practical in online building energy predictions [7].

The physical model is the basis of several popular simulation tools such as DOE-2, EnergyPlus and DeST [8]. Since physical-based models require underlying assumptions and specified input parameter values, which made them time-consuming to establish a model but easier to integrate all the components in the building system. By contrast, data-driven methods are usually more efficient and can be compiled to current artificial intelligence (AI) systems used in buildings [9]. On the other hand, the data-driven building energy use prediction does not require the detailed energy analysis or data about the simulated building and alternatively learns from historical/available data for predictions [10].

Currently, the most popular data-driven methods in building energy predictions are the artificial neural network (ANN), support

[*] Corresponding authors.
E-mail addresses: zeng.building@foxmail.com (A. Zeng), yao.yu@ndsu.edu (Y. Yu).

https://doi.org/10.1016/j.enbuild.2019.04.029
0378-7788/© 2019 Elsevier B.V. All rights reserved.

vector machine (SVM), decision trees and statistical approaches [11], and some other methods, such as random forest and Gaussian mixture model, have also been used [12]. These data-driven prediction methods have been used for different building types and various data resources, and different combinations and strategies were widely discussed in the previous research [13].

Using 'Ensemble Bagging Trees' (EBT), with the input data of meteorological parameters and building-level occupancy and meters [6], hourly electricity demands can be predicted accurately, which is critical to estimate and predict utility bills. Another research used the national data from the Commercial Buildings Energy Consumption Survey (CBECS), and then a gradient boosting regression was proved to be more effective than the linear regression and SVM, since this gradient boosting regression model only demands five parameters [14].

In the UK, a supermarket and its gas and electricity usage data were taken as an example. The multiple regression was proved to be flexible with a high accuracy, and the temperature was found more influential than humidity [15]. Residential load forecasting was also discussed using the log-normal process and conventional Gaussian process prediction, while the log-normal process performed better in terms of accuracy [16]. Multiple linear regression and neural network models have revealed that the difference between household energy demand predictions is relatively small regardless of the effect of different methodologies, and the uncertainty level still has a great effect on energy use prediction results [17].

Some researchers also compared the usage of different prediction algorithms and their superiorities in certain scenarios. A city scale energy use prediction uses the ordinary linear regression (OLS), random forest, and support vector machine to fit the energy benchmarking data, including electricity and natural gas, while the OLS algorithm was identified as the best approach [18].

Among the AI approaches, ANN and SVM are widely used methods to improve the accuracy and other predication performances. Hybrid methods integrating SVM with other methods are preferred such as the Least Square Support Vector Machine (LS-SVM) [19]. These two methods (ANN and SVM) have also been used in heating and cooling energy predictions with low percentage errors [20]. Multiple regression and extreme learning machine can also be used in the heating system prediction, such as the thermal response time in an optimal control [21]. A work in predicting the electricity consumption of a building in Turkey was done to compare different methods, including SVM, LS-SVM, ANN and regression methods, and the LS-SVM was proved to be an accurate and fast computation approach among them [22].

Hybrid methods of other combinations are also recommended, such as combining neural network and optimization methods together [23]. For example, decision tree and ANN can be integrated as a hybrid model for both prediction and classification in the prediction process, and then a unified objective function will be resolved based on either continuous or discrete parameters [24].

The data extraction and selection procedure is also important and critical. A residential building in France showed an improved prediction and performance by using the data of only representative days instead of all of the days, which can mitigate the burden of data monitoring and collecting [25]. A principal component analysis was proved to be effective in the energy prediction of appliances and lighting systems, together with the occupant activity recognition [26]. For instance, a random forest approach has identified the educational feature to be the most influential factor for regional energy use density [27].

The previous methods usually used one consistent method based on historical data or simulated data in a software package by assuming a consistent schedule, which lacks enough capacity to deal with fluctuated operational complexity and thus cannot reflect the industrial need in efficient and accurate predictions. To eliminate such limitations, our research uses big data sets collected from six real buildings with different building types/functions and occupancy schedules. These data were stored in a dynamic Microsoft SQL server database, which have been cleaned and mapped in a common data schema [28]. Principal component analysis including meteorological parameters and occupancy schedules were carried out to reduce the computation complexity, and then energy use prediction results were compared to reach the least bias and computation burden. Commonly recognized methods, including support vector machine, artificial neural network, Gaussian process regression and multivariate linear regression, were used to verify their effectiveness and identify their optimal applications. Those methods were proved to be efficient and accurate in previous research based on simulated data by using the physical based methods, such as the data generated by using EnergyPlus. The kernel functions or inputs were selected based on our preliminary studies, and computational time was also used to evaluate their computational efficiency. Necessary data process methods and visualization were also carried out along with the optimal methods for the selected buildings.

2. Technical methods

The basic work flow of this research includes data collection, pre-processing, prediction algorithms, and result comparison and verification. The data are collected as the actual building operation data from its SQL server database, and then principal components were extracted according to its impact on the building electricity use. Data cleaning, parsing and filtering were carried out, and the electricity uses of six different buildings (Table 1) were predicted with the prediction algorithms including support vector machine, artificial neural network, Gaussian process regression and multivariate linear regression, which are all popular supervised machine learning methods based on the literature review. The prediction results of four different methods were then integrated to the large scale building energy monitoring system, to compare their results with the actual energy use obtained by looking at the pre-defined criteria including deviation values, percentage and statistical analysis. Their effectiveness and computational performance were then evaluated and compared to each other, including their suitable application scenarios.

Table 1 shows the building information of the six target buildings, all of which are using Variable Air Volume (VAV) systems with two of them having Monomer Air-Conditioner (MAC) installed in some special rooms/units.

The energy use of a building can be represented as a simplified function of multiple parameters, i.e., meteorological parameters and human activity schedules, used as independent variables to predict the electricity usage, and their optimal applications were conducted through several evaluations.

A simplified representation of our model is shown below:

$$y(x) = \delta(x_1) + \eta(x_2) + \varepsilon(x_1x_2) \tag{1}$$

The parameters refer to the building, climate and occupant behavior property, where x_1 means the meteorological parameters after data transformation, and x_2 refers to the human behaviors and operation schedule after data transformation, which are all verified through data cleaning, parsing and filtering process.

2.1. Support vector machine

Support vector machine (SVM) is a supervised machine learning method, which is popular in dimensional analysis, machine learning, image recognition, data classification and prediction. It is most

Table 1
Building information of case study buildings.

Building name & type	Building location	Building size (m^2)	HVAC system	Construction year	Annual electricity use (million kWh)
Office-1	West Shanghai	39,000	VAV	2011	3.32
Office-2	West Shanghai	50,400	VAV	2011	2.73
Hotel-1	West Shanghai	59,400	VAV	2013	2.14
Hotel-2	West Shanghai	49,800	VAV	2000	15.88
Shopping Mall-1	West Shanghai	22,000	VAV+MAC	2007	15.61
Shopping Mall-2	West Shanghai	Unknown	VAV+MAC	Unknown	12.49

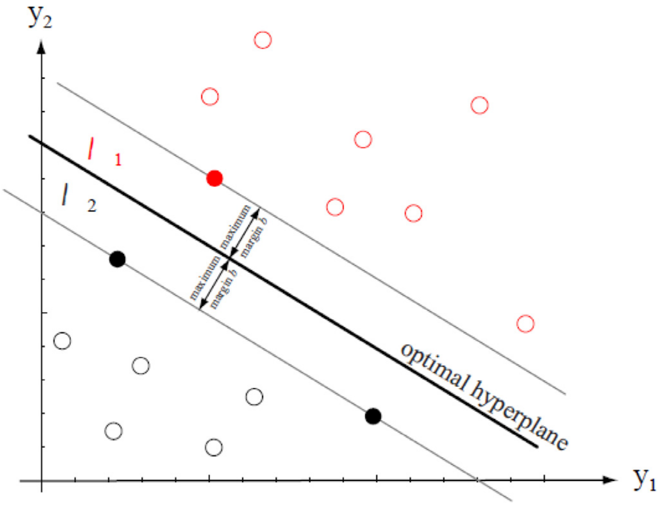


Fig. 1. Diagram of Support Vector Machine [29].

appropriate for a small sample with high dimensions, especially for the data with non-linear property.

If the data set contains data points in an n-dimensional to an infinite-dimensional space, the SVM can classify data into two different classes by identifying the best hyperplane to separate them. The best hyperplane for an SVM is the one with the largest margin of the different classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points. These data points that are closest to the separating hyperplane are identified as support vectors, as shown in the Fig. 1.

An optimal hyperplane can then be calculated, which is always in the n-1 dimension, and thus SVM can be used in the classification of image recognition and building system fault detection. It can also be altered to predict building energy use as a regression method.

In its regression application, the data were sampled to a high dimensional feature space by a kernel function first, and then the projected input data will be calculated by using a linear regression. The regression equation is shown below [29]:

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \mathrm{G}(x_i, x) + \beta, \quad \beta \in R \tag{2}$$

$$s.t. \quad \begin{cases} \sum\limits_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases}$$

The equations have three key components, where α_i and α_i^* are Lagrange multipliers, and C is the threshold which is called penalty parameter. The crucial equation in projecting original data to the high dimensional feature is G(x_i, x), and there are three types of popular kernel functions and other user-defined functions:

$$\text{Linear Kernel :} \quad \mathrm{G}(x_i, x) = x_i' x \tag{3}$$

$$\text{Gaussian Kernel :} \quad \mathrm{G}(x_i, x) = \exp(-\|x_i - x\|^2) \tag{4}$$

$$\text{Polynomial Kernel :} \quad \mathrm{G}(x_i, x) = (1 + x_i' x)^p \\ \text{where p is in the set } \{2, 3, \ldots\}. \tag{5}$$

Different kernel functions are evaluated in the prediction training process, indicating that the linear kernel is the most computation efficient kernel and the Gaussian kernel shows a good performance, which is suitable for analyzing non-linear data.

2.2. Artificial neural network

Artificial neural networks (ANN) are data mining algorithms created based on the biological neural networks, which are designed to be a supervised machine learning method. The ANN model is claimed to have a relatively high accuracy compared to other methods with low computation burdens [27]. Besides, the relationship between prediction and dependent (output) variables is not required before the model implementation because the supervised learning process will identify it during the model creation process. This study uses a multilayer feedforward network to establish the neural network using a backpropagation algorithm. A typical neural network has three layers: input, hidden, and output layers [28]. The input layer includes n neurons determined by the number of input data variables; the output layer has a single neuron for the dependent variable; and the hidden layer has 2n + 1 neurons for a preliminary structure.

The ANN model applied in the prediction of building electricity energy consumption can use the same data set as other methods, but with simplified climate parameters and occupancy schedules as the dependent variables and the energy use as the output variable. The simplified climate parameters were used because they can greatly decrease the computation time that is needed, but the prediction accuracy will not be greatly affected at the same time based on our preliminary study.

2.3. Gaussian process regression

Gaussian process regression (GPR) models are nonparametric kernel-based probabilistic models, which use a finite number of joint or multivariate Gaussian distribution to accumulate the actual data distribution.

A linear regression model based on Gaussian process regression is of the form:

$$y = x^T \beta + \varepsilon \tag{6}$$

where $\varepsilon \sim N(0, \sigma^2)$, and the coefficients matrix of β is estimated from the data as well as the error variance σ^2. A GPR model explains the response by introducing latent variables, $f(x_i)$, $i = 1$, 2, …, n from a Gaussian process (GP), and explicit basis functions, h. Then a mean function m(x) and covariance function, k(x,x′) will define the Gaussian process, so an instance of response y can be modeled as

$$P(y_i | f(x_i), x_i) \sim N(y_i | h(x_i)^T \beta + f(x_i), \sigma^2) \tag{7}$$

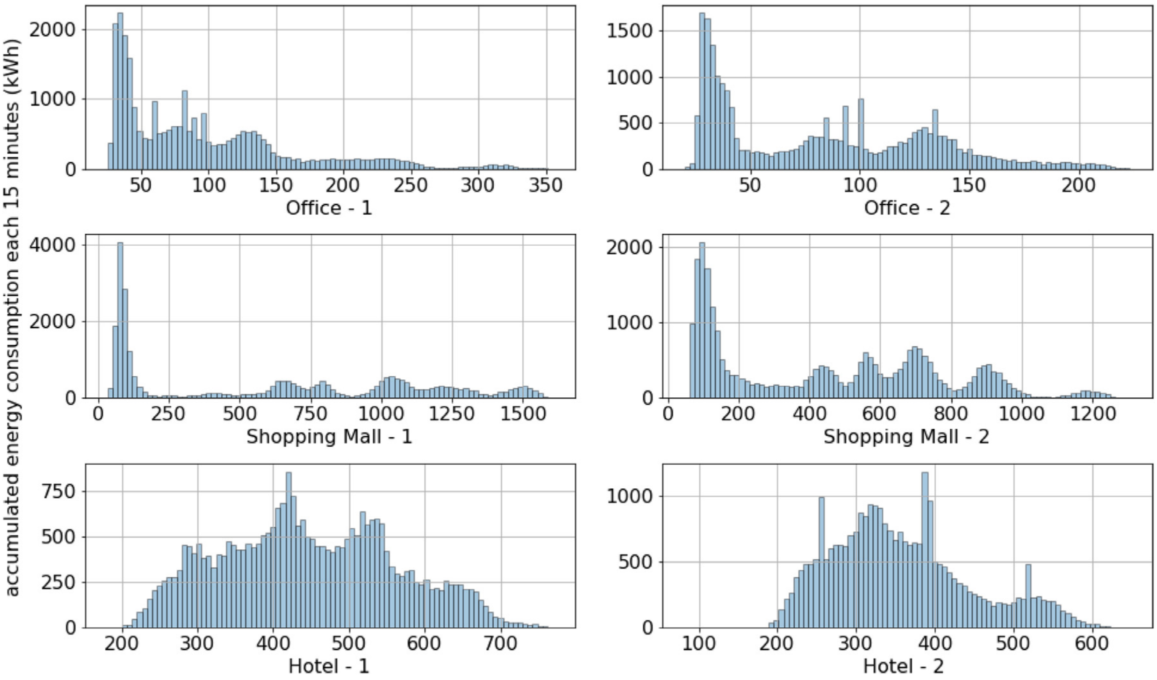


Fig. 2. Illustration of the joint Gaussian distribution of all case buildings' energy use.

Based on the physical knowledge of building energy usage, the energy use of a building system or a whole building should follow an appropriate Gaussian distribution under a similar meteorological condition. Then the energy use of a whole building during a long period can be identified as joint Gaussian distributions with different probabilities of energy use. This will be a suitable scenario to use the Gaussian Process Regression, and the energy use histograms are shown below in Fig. 2. It can be seen that four out of six buildings have joint Gaussian distributions excluding Office-1 and Hotel-2, and the Shopping Mall-1 and 2 are significant joint Gaussian distributed, which were revealed by visualization.

2.4. Multivariate linear regression

Multivariate linear regression is quite popular in the statistical analysis of most physical systems, due to its high computation efficiency and ability to clearly and directly illustrate the relationship between inputs and outputs. Our model is conducted as:

$$y_\mathrm{i} = \beta_0 + \beta_1 x_{\mathrm{i}1} + \beta_2 x_{\mathrm{i}2} + \ldots + \beta_\mathrm{p} x_{\mathrm{ip}}, \quad i = 1, \ldots, n. \tag{8}$$

where

- y_i is the *i*th response, which is the electricity energy use at *i*th time step;
- β_k is the *k*th coefficient, and β_0 is the constant term (also called intercept);
- x_ij is the *i*th observation on the *j*th dependent variables after data transformation, $j = 1,\ldots$p, which is the combination of meteorological and occupant behavior parameters.

2.5. Prediction result evaluation algorithms

The efficiency and accuracy levels of these algorithms were evaluated and compared on data sets collected through the Building Energy Management System, whose data were stored in a SQL server database. The necessary procedure, such as data cleaning, filtering and parsing, was integrated into the algorithms to increase the reliability of the data. The basic prediction algorithm was established by offline data to find out an optimal kernel function and inputs combination, and then tested in online prediction including training and testing in order to compare different algorithms on the compatibility of large scale actual building energy data.

Root Mean Squared Error (RMSE) and Normalized Mean Bias Error (NMBE) were used as statistical evaluation criteria to measure the deviation of the predicted values from actual operations [30]. To achieve a more accurate evaluation of the prediction results, coefficient of variance RMSE (CV-RMSE) was used instead of classical bias determinant RMSE to avoid ambiguity [31,32].

The variations of CV-RMSE and NMBE values are mainly determined by different operation schedules and weather conditions, and a result fitted with a lower CV-RMSE and NMBE is usually favorable. Based on ASHRAE criteria, when hourly calibration data are used, these requirements shall be 30% and 10%, respectively [38]. Since our data have a more dense time interval (15 min), if these criteria can meet the ASHRAE requirement, then the prediction deviations are within the allowable range, and a small amount of larger deviations beyond these ranges can be also acceptable.

Another coefficient of determination in the model evaluation is *R*-square [33]. Since more than one variable were utilized in the research, adjusted *R*-square was eventually utilized to improve the performance of *R*-square, which is designed to handle large size data with multiple inputs to eliminate the negative effect of extra explanatory variables. The adjusted *R*-square is represented as [34]:

$$\begin{aligned} Adjusted \quad R-square &= R^2 - (1-R^2)\frac{p}{n-p-1} \\ &= 1-(1-R^2)\frac{n-1}{n-p-1} \end{aligned} \tag{9}$$

In this equation, *n* is the size of data/sample size, and p is the number of exploratory variables, R^2 means *R*-square. Adjusted *R*-square can eliminate the bias by increasing exploratory variables, which is a measure of suitability of alternative nested sets, and thus it is particularly useful in the feature selection stage of building model [35]. A higher adjusted *R*-square is usually preferred.

Average deviation is another criteria which can reveal the percentage deviations of the whole data set, and it is represented as:

$$Average \quad Deviation = \frac{|y^* - y|}{y} \times 100\% \tag{10}$$

3. Data prediction process

3.1. Building site and data monitoring system

The studied buildings are all located in Shanghai, China, with a long hot and humid summer which can last for four and a half months, and also a contradictory long and cold winter which can last another four months. The temperature fluctuation during shoulder seasons is usually significant with the temperature variation higher than 10 °C, even though it has minor effect on the air-conditioning energy use based on our preliminary study using correlation factors. Because during that period, people used to adjust their clothing patterns to handle the temperature vibration instead of using air-conditioning systems to improve occupant thermal comfort. Energy sub-metering systems were widely used to record the electricity use of an entire building and the key components of the building systems including lighting and plugs, elevators and pumps, HVAC systems and miscellaneous items. All of these data were uploaded to a central server and stored in a SQL server database, which were used in this study for direct download, export and prediction, as well as further smart energy management applications [36]. These data were previously used in offline computation for a simple analysis and have great potential in further research and data mining during energy prediction and smart building operation.

In Shanghai, more than 1500 large commercial buildings (>20,000 m^2) [37] have their electrical energy use sub-metered, and most of their energy use are recorded and stored in a Microsoft SQL server database. Some buildings have their data accessible to the development of prediction algorithms in this research. It was assumed that the baseline energy consumption is a steady process with random white noise, and the load of the individual zone is relatively consistent within a day. The time interval of data monitoring is 15 min, and missing data are completed through the data pre-process using a linear interpolation.

In this research, six buildings were studied, including two shopping malls, two hotels, and two office buildings (Table 1). They all have different long-term and short-term operation schedules. Due to the common energy consumption pattern in Shanghai, these buildings all use electricity for lighting, office equipment, security, elevators and HVAC systems. In other words, the electricity usage equals to the whole building energy consumption.

3.2. Data preprocess and selected parameters

The development of the prediction methods went through two major stages: the preliminary offline model development and then the online large scale test. The offline model development involves small scale historical data analyzed using a personal laptop, while the online test is based on real data extracted from the SQL server database and the optimally selected methodologies. All the comparisons conducted in the study are based on the online test results.

Based on the offline data association analysis, multiple parameters collected through the building information monitoring system were tested to identify their correlation with electricity usage. Critical parameters in this hierarchy analysis include dry bulb temperature, wet-bulb temperature, enthalpy, human activity and operation schedule.

In the online prediction algorithm tests, a prediction program on a remote computer was set up to run automatically at a fixed time during different days, which would generate a numeric matrix as the result. The prediction will then be compared to the actual values simultaneously. The program can self-adjust its key coefficients and kernel variables according to the behavioral pattern of the most recent data.

Standardized association factors were used as feature extraction methods to investigate the correlation between input and output variables. The correlation analysis illustrated that the dry-bulb temperature, wet-bulb temperature and enthalpy are the most influential meteorological parameters, while other factors, including humidity level, etc., have negative or zero contribution to the energy usage. The standardization of parameters was proved to be beneficial to improve the reliability of original data, and dimension reduction approaches were applied in the study to reduce the complexity of computation.

The occupancy schedule is a kernel parameter in energy prediction, the most important factor is occupied and unoccupied hours which will greatly impact the energy use because occupants will use heating, electricity plug, office equipment, and air-conditioning devices. The schedule used in this study was extracted from the building operation and maintenance log considering the national holidays and off work hours.

The data being used in the prediction still have outliers or missing points that may lead to unexpected deviations. For example, Office-2 has a large number of zero energy use, which cover 9.62% of the total data points, and the number of zero values for Hotel-1 accounts for 0.13%, based on the maintenance log. There should be no zero values because these buildings are in operation constantly, and thus these zeros were intentionally eliminated in the prediction. Large deviations also exist. For example, 0.16% of accumulated energy use of Office-2 are much higher than the ordinary energy consumption, while 0.004% of values for Shopping Mall-1 are too large, which are thus suspected as monitoring bias, since these values are far beyond the reasonable range of the normal maximum energy use for these buildings. After eliminating all the zero values and large bias, all the rest data were used to generate prediction algorithms.

3.3. Training set and test set split

The development of the prediction methods relies on high quality training and test data split approaches, and in this study, a self-learning method was applied. Specifically, the algorithms take one day's data as the test data and the 29 days' data prior to this day as the training data, since the energy use of the testing day (related to the test data) is affected more significantly by the data for the days closer to it, and a moving weighted average was used to increase the weight of nearby days and also include the days prior to 29 days being used. After one prediction was done, the method used will automatically move to the next day and continue the same process, until the energy use of the last day in the whole time period was predicted successfully. In the end, the algorithm will combine all the prediction results and put them together according to their timestamps to compare with actual values.

After finalizing the training and test sets, to further improve the computational speed and accuracy, the data set was split based on time step. For example, all the data points at 6:00 pm were extracted from the training set and then regrouped to predict the energy use at 6:00 p.m. in the test set, and this process was repeated for all the time steps.

A simplified visualization is shown as below, where the above figure is an example of the training set, and the test set is shown in the figure below (Fig. 3).

In general, our training and test data have similar behaviors, indicating that neither significant over-fitting nor under-fitting issue occurs. For example, the above figure showed the energy use being predicted for a typical day. The statistical criteria for the training and test data, including RMSE, *R*-square and NMBE, all meet the ASHRAE's requirement with minor differences.

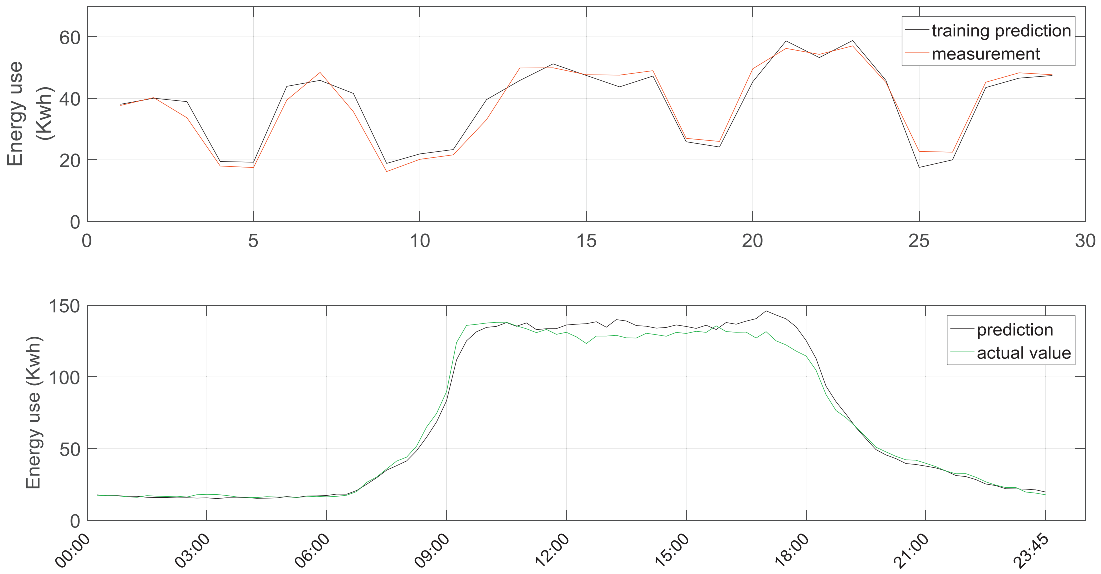


Fig. 3. Visualization of training data and test data and their prediction result.

4. Evaluation of prediction results

4.1. Data visualization of original values

Three data visualization approaches were applied to all data sets among all the time intervals, and some representative scenarios of certain days are shown in Fig. 4(a). A weekly box plot was used to reflect the weekly energy use of ten months during the year of 2017. Significant data deviations from typical data distributions were shown in red, and the histogram was used for the whole data set visualization and deviation analysis. The overall review of the data set indicates the significant energy use on a daily and weekly basis for the target building Office-1, which is an office building, during night and weekend or holiday, and the energy use is usually lower than that during the workday, which matches the original assumption based on energy use patterns, i.e., energy use is low during night and holiday while during workday the building consumes more energy. The other two visualization methods showed different energy use patterns of other buildings. For example, for the two shopping malls, there is no obvious difference between holiday and weekday in terms of the typical operation schedule, and thus the energy use curves are more driven by moving averaged ambient air temperature, as shown in Fig. 4(b). Fig. 4(c) illustrates the comparison of original energy use pattern of these buildings, and it can be seen that Shopping Mall −1, Shopping Mall-2, Hotel-1 and Hotel-2 are mainly driven by ambient air temperatures, while some maintenance works or system upgrading may affect their energy use curves significantly. Office-1 and Office-2 have another patterns, and their energy use is quite low during weekend and holiday due to the absence of occupants, indicating that their schedules have a more significant impact on energy use than other factors.

4.2. Statistical analysis of prediction results

Statistical analysis of prediction results in terms of statistical criteria and computation time are included in this study. In general, an algorithm with the least computation time, the highest *R*-square and the least NMBE and CV-RMSE is preferred. The data prediction results for Office-1 by using the support vector machine is shown below in Fig. 5(a). Fig. 5(b) shows the energy usage comparison between the predictions using different methods and the actual energy consumption of Office-1, the Office-1 was selected because it covered all typical prediction bias which represent all the scenarios for all buildings.

As shown in Fig. 5(a), the prediction results have the same tendency and variation with the actual energy use, including the change point, and only a small number of deviations were observed through the visualization, which were derived from the unsteady vibration in terms of occupant activities and schedules at noon. Another improvement is that the box plot (Fig. 5(a)) shows less faulty operations in the weekly data visualization, and the distribution becomes more steady and is close to the mean value of the operational data. In fact, the values of the faulty boxes are reduced to 0% in the prediction results. Fig. 5(b) illustrates the comparison of the actual and predicted values during a short time period, which represents a small amount of severe deviations. Still, the prediction tends to have more steady schedules than the actual data, which is represented through an optimized operation curve. The SVM curve shown in Fig. 5(b) represents the result for a short period of time with more significant deviations compared to the SVM results shown in Table 2, which represent the results for a longer time period (10 month).

It is clearly revealed by Fig. 5(b) that SVM gives a significant deviation during the first day, which might be a potential/inherent issue in using SVM in such applications. To find out the possible reason for that, more simulations and studies would be needed in the future study. ANN has relatively large deviations during the last two days that are weekends, while GPR and MLR have no such errors. Uncommon fault greatly decreases the potential for using ANN and SVM in a future large scale energy use prediction, and a study on the longer term of energy use prediction suggested that ANN tends to have more severe deviations than other methods.

Popular statistical criteria for prediction are also considered, as shown in Table 2, for Office-1 with high energy density and more steady operation schedule. The values for the Adjusted *R*-square, NMBE, and CV-RMSE of four different methods are quite similar, so the prediction evaluation results cannot generate any preference among them through the individual evaluation, and thus these criteria must be considered together. Still, the computation time has a severe difference. ANN usually took quite a longer time to generate reasonable results, while GPR was much faster than all the other methods. At the current stage, SVM, GRP and MLR can all satisfy the requirement of quick response predictions, while ANN is not

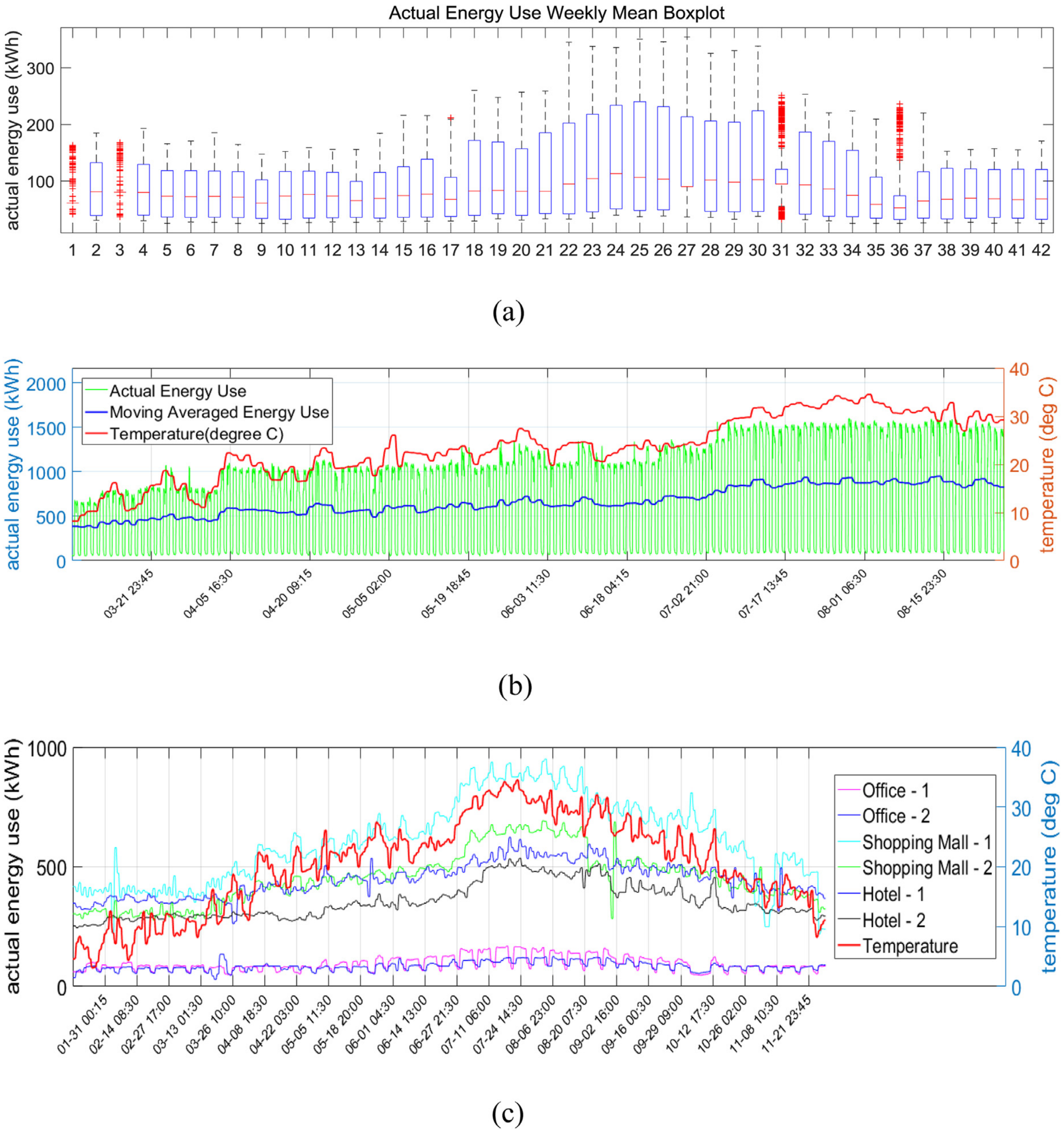


Fig. 4. Data visualization of typical building operations based on original data (a). representative scenarios of certain days of Office-1, (b) overall data review of Office-1, (c) time series plot of all six buildings in terms of energy usage and moving averaged temperature.

Table 2
Adjusted *R*-square, NMBE, CV-RMSE and computation time of the prediction @ Office-1.

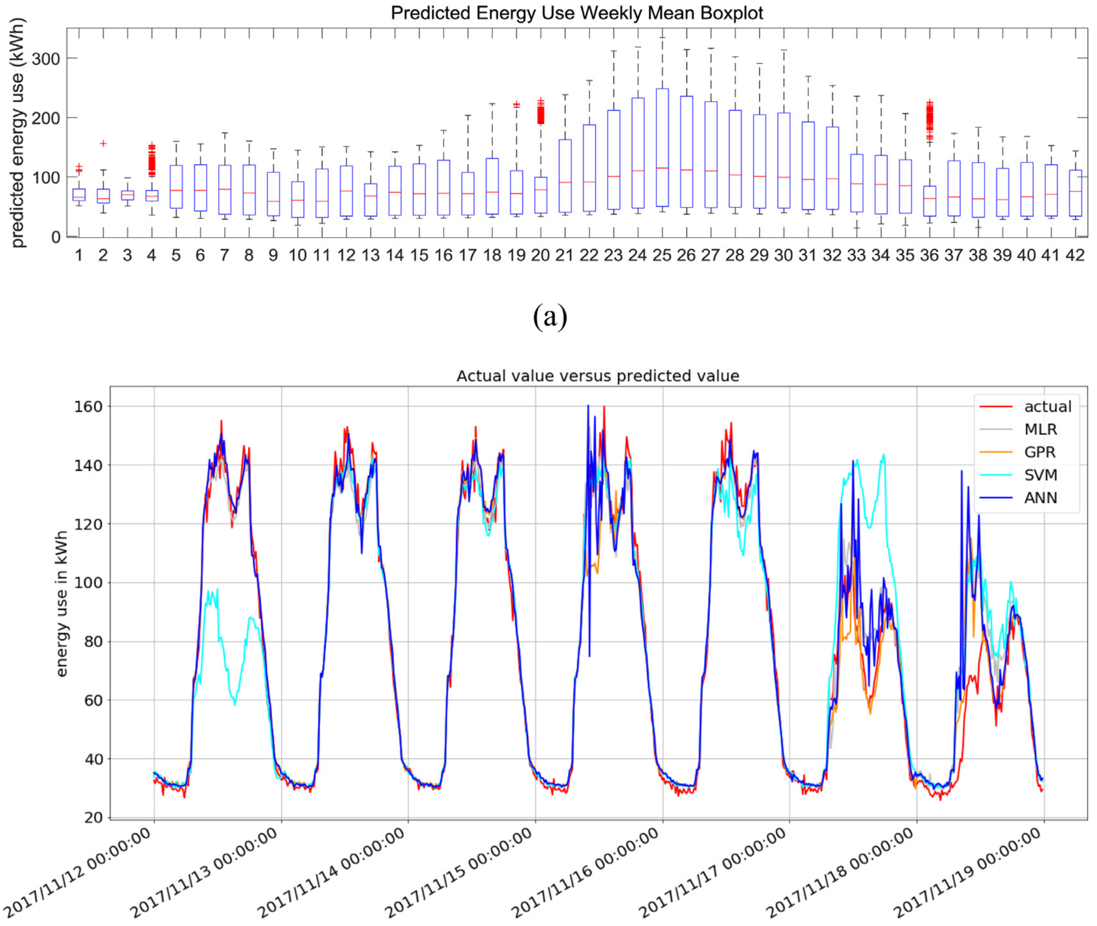
Prediction methods/Evaluation criteria	Adjusted *R*-square	CV-RMSE	NMBE	Computation time/CPU time (s/loop)
Artificial neural network	0.822	28.691	0.550	26.065
Gaussian process regression	0.878	23.778	0.362	0.02
Multivariate linear regression	0.870	24.550	2.610	2.670
Support vector machine	0.864	25.101	3.004	1.354

superior compared to the others. The computer used for all these simulations has a common configuration for a laptop, i.e., Intel(R) Core(TM) i5-5200 U CPU @ 2.20GHs 2.19 GHz as the processor.

4.3. Comparison of different methodologies for different buildings

To identify an optimal method in energy prediction for different types of buildings, the results of different methods are compared by looking at their deviations of energy use values, deviation percentages, and statistical criteria.

The commonly used criteria in the statistical regression and prediction are used here to investigate the goodness of data fitting, including the adjusted *R*-square, NMBE, and CV-RMSE. Average deviations are also calculated and represented. Their effects were considered together and optimal comparative evaluations of these methods were given.



(a)

(b) Comparison of original and predicted value in a short period by four methods for Office-1.

Fig. 5. Visualization of energy use prediction (a) weekly prediction results box plot of Office-1, (b) energy use prediction versus actual value of Office-1.

The average deviations of the six studied buildings vary a lot as shown in Fig. 6, and these high fluctuations reflect that there is not a once-for-all method suitable for all types of buildings. SVM behaves best for Hotel-1 and Hotel-2; MLR also behaves best on these two and slightly better than SVM. ANN behaves best for Hotel-1 and Hotel-2 but is still the worst among the four methods, and GPR has a different behavior pattern among these two buildings while it is worse than MLR in Hotel-2 but same as MLR in Hotel-1. For Office-2 and Shopping Mall-2, their high average deviations indicate that none of these methods are very suitable for these two buildings.

Based on the ASHRAE Guideline 14-2002 [38], all of our predictions can meet the requirements for NMBE and CV-RMSE, which indicates that the prediction methods are acceptable for further applications in real building energy predictions.

From the perspective of statistical analysis, the adjusted *R*-square, NMBE, and CV-RMSE should be evaluated together, and the methods with high *R*-square and low NMBE and CV-RMSE are identified as the optimal methods. Figs. 7–9 indicate that MLR and GPR methods usually stand out at some rare scenarios. ANN usually works better to minimize NMBE, while MLR tend to minimize the average deviation based on percentage, SVM works well for buildings with a steady operation schedule, and GPR works best for Office-1 and Shopping Mall-2, which typically have the highest uncertainty that is likely caused by unrecorded and unpredictable occupant behaviors.

It can be seen that for different comparison criteria, the same method might receive slightly different evaluation results, which reflects the complexity of our building energy use patterns. By averaging different criteria, however, we can evaluate the effectiveness of different methods.

As a conclusion, the optimal algorithm for different buildings varies a lot. Different methods have their own optimal application scenarios. MLR is considered to be optimal for the buildings of Office-2, Shopping Mall-1, Hotel-1 and Hotel-2. GPR is considered to be optimal for the Office-1 and Shopping Mall-2, while GPR and SVM are identified as the fastest method. For different deviation requirements, MLR is best when least average deviation, ANN is best when NMBE minimization is required but usually have low adjusted *R*-square coefficients, GPR can usually reach least CV-RMSE.

The comparison of prediction methods for Office-1 and Office-2 shows a great difference between their prediction results, even though they have a similar function/building type. Office-2 have

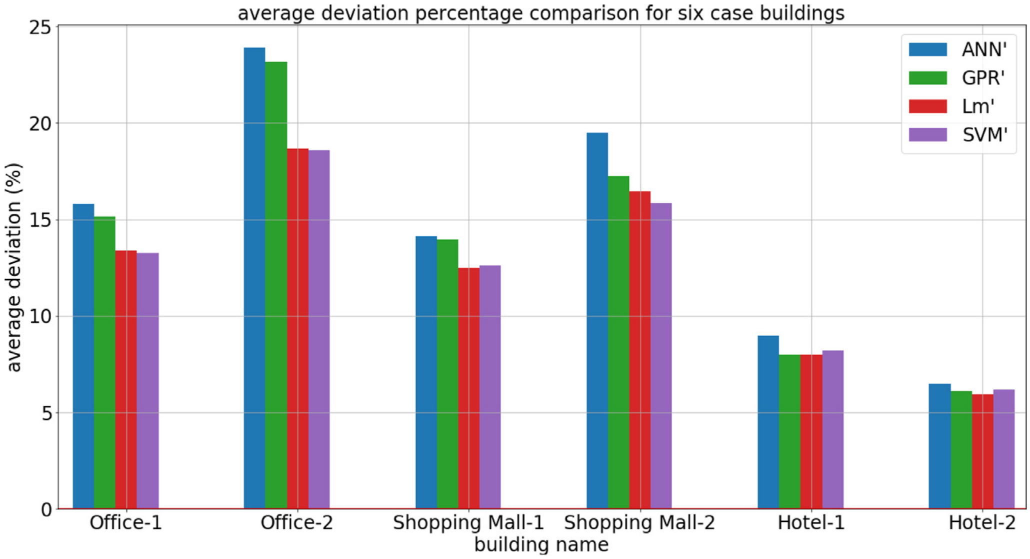


Fig. 6. Average deviations of the annual data prediction distribution of the six buildings.

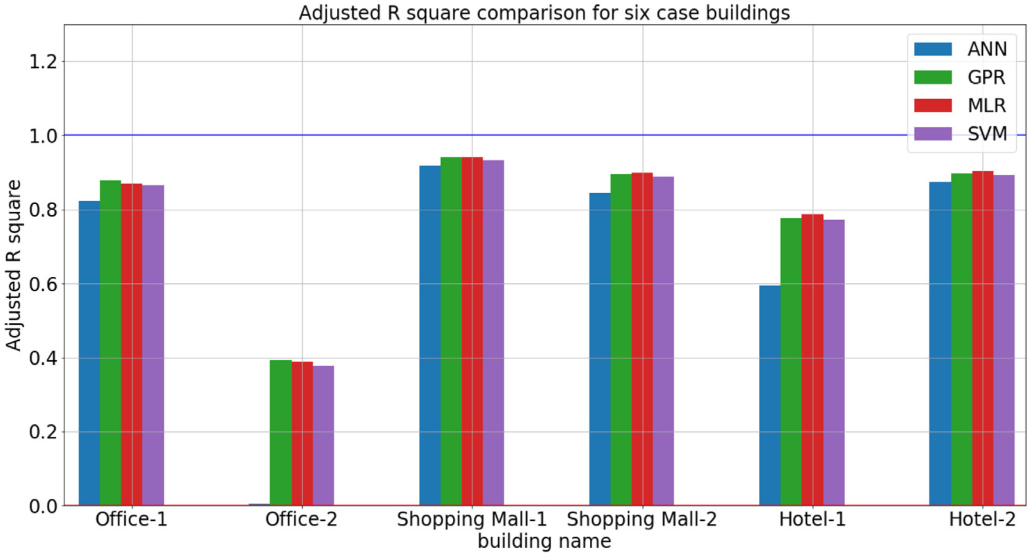


Fig. 7. Adjusted *R*-square of the annual data prediction distribution of the six buildings.

more data points effected by unknown reasons which lead to high deviations by all four methods especially ANN because ANN would tend to overfitting the energy use.

Another aspect of evaluating the prediction is data distribution, which illustrates the data proficiency and result reliability. Original data visualizations illustrate the energy usage as a non-ideal Gaussian distribution, and the temperature being recorded as the principal independent variable is non-Gaussian either, even though for certain buildings they can be treated as joint-Gaussian distributions.

The prediction result evaluations have more dimensions of data distributions that can be analyzed, and all percentage deviations related to the four prediction algorithms show quasi-Gaussian distributions, including a dual peak joint Gaussian distribution for Shopping Mall-2, which indicates that the difference between the actual and predicted values is partially dependent on a single factor. The central limit theorem has indicated that if the observations are huge enough, the whole data set will follow the Gaussian distribution, even if each observation is not Gaussian distributed [25]. It can be proved that the data sets used in all these prediction methods are big enough to represent the actual building operation conditions, as shown in the sample visualization plot in Fig. 10. In the Gaussian distributions as shown in this figure, y-axis illustrates the number of data points for the six buildings, and x-axis represents the different prediction algorithms. The Kolmogorov-Smirnov test shows our results are not ideal enough, which can be proved if more bins are used in the histogram. Hence, bigger data sets with high density are needed in the future analysis, and the histogram visualization can have more bins in order to increase their accuracy.

Despite the great performance of the overall prediction, severe deviations at certain time steps still exist. Based on the detailed data verification, the extreme bias in the prediction process is due to the original data faults, which requires the improvement of data quality and sensor calibration. Sensor malfunction or data acquisition bias were identified as the most relevant and common causes. For example, the office building has constant zero electricity consumption for two weeks, while the utility bills show the actual energy use are not zero during these two weeks. This could result in great deviations and thus negatively affect the training of

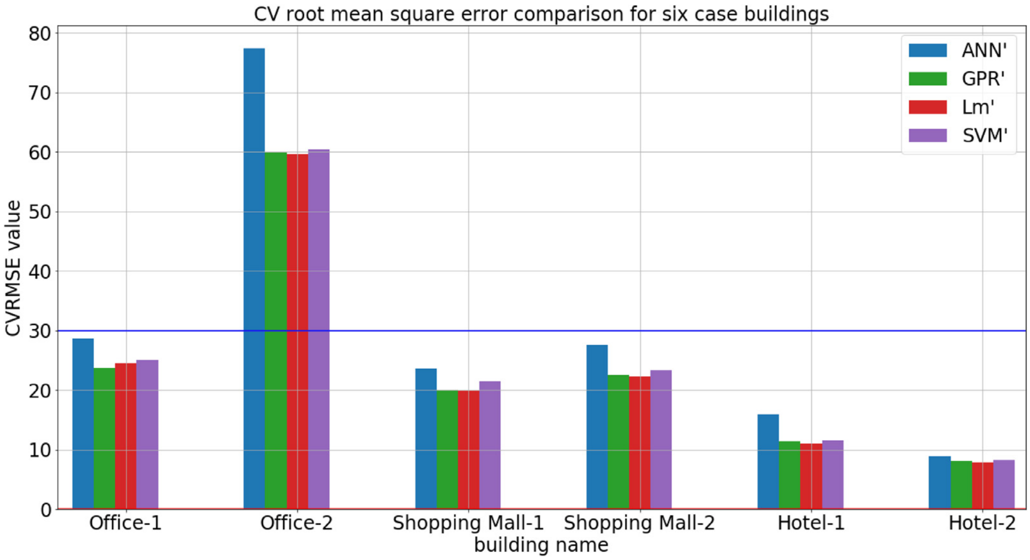


Fig. 8. CV-RMSE of the annual data prediction distribution of the six buildings (blue line indicates the threshold defined by the ASHRAE standard). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

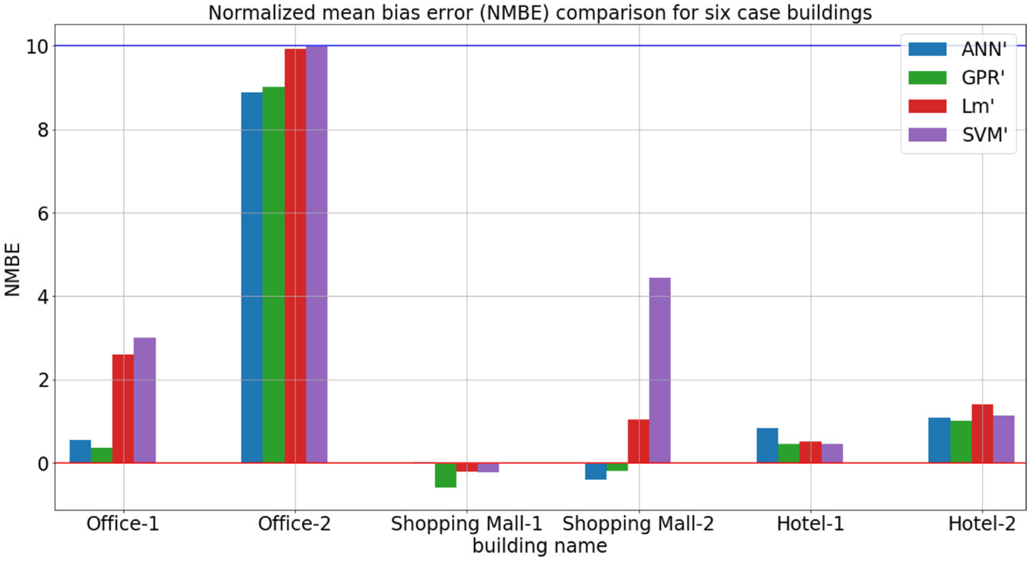


Fig. 9. NMBE of the annual data prediction distribution of the six buildings (blue line indicates the threshold defined by the ASHRAE standard). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the models. The data preprocess can eliminate most of these significant deviations, but minor errors cannot be all detected and deleted. Another cause of deviation is due to the meteorological data. Most of the sensors are not well located at their correct positions, which thus cannot reflect the actual weather conditions for the target buildings. Therefore, the weather monitoring sensors are going to be modified for certain buildings in the future analysis.

5. Discussion

The limitation of the accuracy that restricts the use of online prediction is human activity data, especially the sudden change of occupant numbers and equipment operation by maintenance staffs. By using the number of occupants with their schedule and clustering them to a simplified data set, our model can draw a conclusion to take occupancy into consideration without sacrificing too much computation efficiency. One alternative solution is to use the equipment operation status instead of direct human activities as an independent variable in a real-time prediction, since a large portion of energy use is determined by building mechanical equipment, which has the highest variation, and thus detailed data of the equipment are typically needed. The equipment operation can be recognized as a reflection of an occupant behavior and can be monitored by using building energy management systems (BMS). The utilization of these data can improve the energy prediction greatly. Thus the acquisition of data from both BMS and sub-metering systems is necessary. Most buildings in Shanghai have different suppliers for BMS and sub-metering systems, and thus it is needed to integrate the data among different resources or suppliers, rather than only rely on the sub-metering data in the current database.

Furthermore, the whole building energy use prediction is directly associated to utility bills, and it can help to smooth the operation curve by using energy storage devices, to decrease or shift the peak demands, and to reduce unnecessary utility bills and maximum demand (MD). It can also be used in a building design decision making process to determine whether an energy storage device or high frequency response control is needed, based on peer

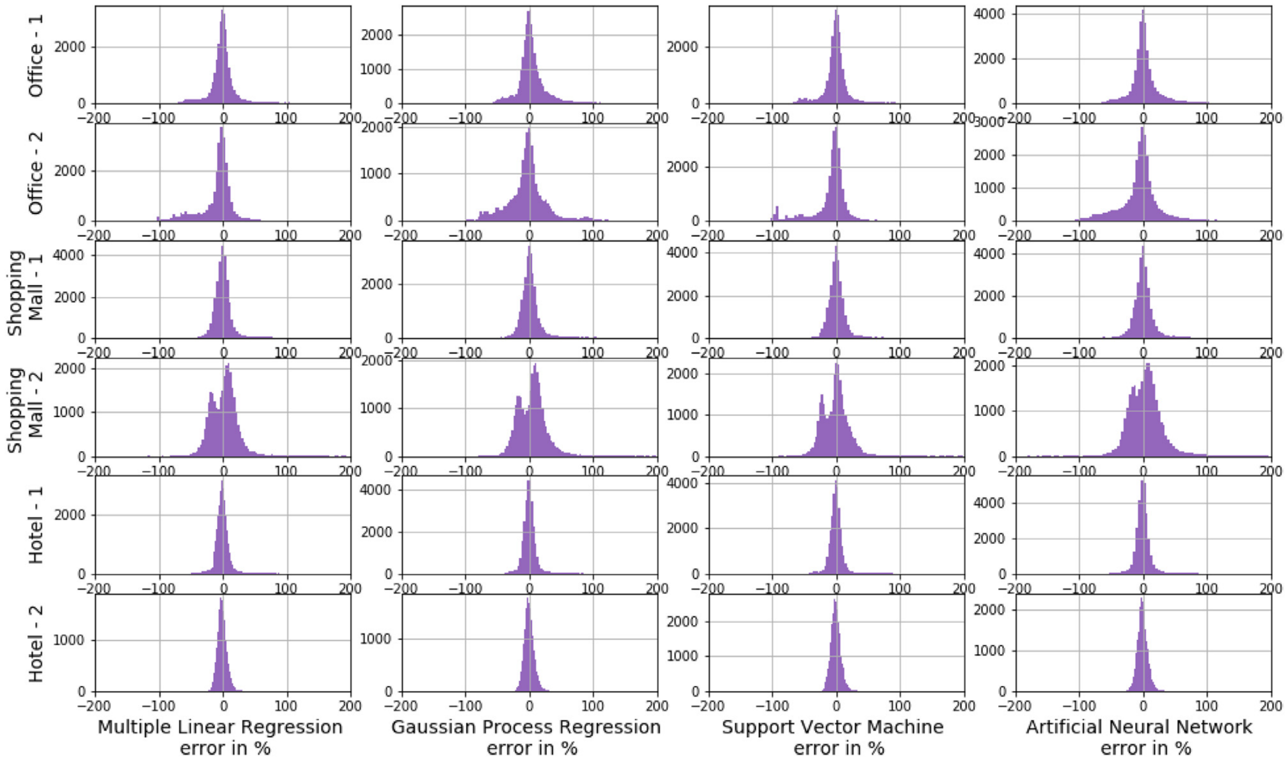


Fig. 10. Visualization of the annual data prediction distribution (x-axis represents percentage error of four algorithms, the y-axis represents number of data points).

group comparison of buildings with similar functions. Meantime, the Chinese government is promoting the use of whole building energy use sub-metering of different types of energy sources, including natural gas, water, hot water and steam, which will provide more data for energy prediction and management in the future.

6. Conclusion and future work

This paper introduced four electricity use prediction and forecasting methods, based on multivariate linear regression, Gaussian process regression, support vector machine, and artificial neural network. In addition, the accuracy of prediction and computation complexity were compared and discussed.

Based on the results shown above, several core conclusions were drawn and shown below.

1) The proposed electricity consumption algorithms can all be used in electrical energy prediction regardless of different weather conditions or occupancy schedules, and the best results were obtained in the energy use prediction of an office building due to its steady schedule.
2) The Gaussian process regression method can generate the prediction results with the least CPU time for most of the scenarios and MLR have the lowest relative deviation with low CPU time, since results of MLR can be best explained by using building physical knowledge, it is of great potential for use in a real-time monitoring system for building energy prediction with optimization algorithms.
3) The support vector machine method has a steady behavior with low accuracy deviations, and the computation time is relatively low. For small data set with nonlinear relationship between different parameters, SVM is recommended especially in real-time control system.
4) Artificial neural network has the least behavior performance in terms of predication accuracy, with high CPU time on all of these occasions. For ANN with many hidden layers and nodes, the deficiency of computation complexity would negatively affect its application in a real-time building monitoring system. Due to its low accuracy, in industrial applications involving large scale building energy use data, ANN is not preferred unless there are very high data uncertainties.
5) Considering the equilibrium of prediction accuracy and the least computation time, multivariate linear regression is considered as the best method in the case study with simplified inputs, while SVM is the best method with complex inputs and nonlinear relationship. To ensure a faster and more reliable prediction, large data sets and real-time methods can be utilized, and dimension reduction methods will be a good option. Detailed occupant activity can be a good option to greatly increase the prediction accuracy.
6) Based on the evaluation results of different buildings, it can be inferred that using principal components can generate great results without sacrificing much accuracy and can increase the computational efficiency significantly. So the input variables should be calibrated and verified as principal components before different methodologies can be concluded for different types of buildings.

The current prediction methods are based on one year's data focusing on only building electricity use. In the future work, it can be extended to a larger system by integrating thermal dynamics of building equipment and complex systems. These methods can also be used in adjusting a control system based on the deviation of prediction from the actual operation monitored by using a real-time monitoring system. As the tendency of smart cities, these methods could be used to predict the energy use of a district and guide the use of energy storage devices to reduce the peak demands and fix missing data or detect sensor value deviations and operational faults.

Additionally, the way to record the occupancy schedule can be improved further in the future study from the binary variables that were used in this study to more parameters in representing and reflecting the detailed occupant behaviors, such as occupant numbers, window on/off, or equipment on/off by occupants.

From the perspective of data science application, the current prediction methods focus on classical statistical learning methods, which have the limited parameter tuning ability [39] in the big data application. In the future, ensemble methods, such as Random Forest and deep learning methods, will be utilized to establish more accurate and reliable prediction algorithms based on elaborate data with detailed occupant behaviors, and hyper-parameter tuning, such as batch size or number of layers and nodes, will be used to eliminate under-fitting and over-fitting problems [39,40].

Conflict of interests

In this statement, we acknowledge no conflict of interest exits in the submission of this manuscript, thus the manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Acknowledgement

The authors gratefully acknowledge the online data from all these buildings as well as all the owners and engineers involved in the maintenance and operation of these building energy management system.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.enbuild.2019.04.029.

References

[1] T. Walter, M.D. Sohn, A regression-based approach to estimating retrofit savings using the building performance database, Appl. Energy 179 (2016) 996–1005. https://doi.org/10.1016/j.apenergy.2016.07.087.

[2] X. Lü, T. Lu, C.J. Kibert, M. Viljanen, Modeling and forecasting energy consumption for heterogeneous buildings using a physical-statistical approach, Appl. Energy 144 (2015) 261–275. https://doi.org/10.1016/j.apenergy.2014.12.019.

[3] T. Ahmad, H. Chen, Y. Guo, J. Wang, A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: a review, Energy Build. 165 (2018) 301–320. https://doi.org/10.1016/j.enbuild.2018.01.017.

[4] J. Drgoňa, D. Picard, M. Kvasnica, L. Helsen, Approximate model predictive building control via machine learning, Appl. Energy 218 (2018) 199–216. https://doi.org/10.1016/j.apenergy.2018.02.156.

[5] N. Fumo, A review on the basics of building energy estimation, Renew. Sustain. Energy Rev. 31 (2014) 53–60. https://doi.org/10.1016/j.rser.2013.11.040.

[6] Z. Wang, Y. Wang, R.S. Srinivasan, A novel ensemble learning approach to support building energy use prediction, Energy Build. 159 (2018) 109–122. https://doi.org/10.1016/j.enbuild.2017.10.085.

[7] H-x. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (2012) 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049.

[8] Z. Li, Y. Han, P. Xu, Methods for benchmarking building energy consumption against its past or intended performance : an overview, Appl. Energy 124 (2014) 325–334. https://doi.org/10.1016/j.apenergy.2014.03.020.

[9] J. Kneifel, D. Webb, Predicting energy performance of a net-zero energy building: a statistical approach, Appl. Energy 178 (2016) 468–483. https://doi.org/10.1016/j.apenergy.2016.06.013.

[10] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Applying support vector machine to predict hourly cooling load in the building, Appl. Energy 86 (2009) 2249–2256. https://doi.org/10.1016/j.apenergy.2008.11.035.

[11] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.

[12] A. Srivastav, A. Tewari, B. Dong, Baseline building energy modeling and localized uncertainty quantification using gaussian mixture models, Energy Build. 65 (2013) 438–447. https://doi.org/10.1016/j.enbuild.2013.05.037.

[13] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, et al., A review of data-driven approaches for prediction and classification of building energy consumption, Renew. Sustain. Energy Rev. 82 (2018) 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108.

[14] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M.A. Brown, et al., Machine learning approaches for estimating commercial building energy consumption, Appl. Energy 208 (2017) 889–904. https://doi.org/10.1016/j.apenergy.2017.09.060.

[15] M.R. Braun, H. Altan, S.B.M. Beck, Using regression analysis to predict the future energy consumption of a supermarket in the UK, Appl. Energy 130 (2014) 305–313. https://doi.org/10.1016/j.apenergy.2014.05.062.

[16] M. Shepero, D. van derMeer, J. Munkhammar, J. Widén, Residential probabilistic load forecasting: a method using gaussian process designed for electric load data, Appl. Energy 218 (2018) 159–172. https://doi.org/10.1016/j.apenergy.2018.02.165.

[17] A. Kialashaki, J.R. Reisel, Modeling of the energy demand of the residential sector in the united states using regression models and artificial neural networks, Appl. Energy 108 (2013) 271–280. https://doi.org/10.1016/j.apenergy.2013.03.034.

[18] C.E. Kontokosta, C. Tull, A data-driven predictive model of city-scale energy use in buildings, Appl. Energy 197 (2017) 303–317. https://doi.org/10.1016/j.apenergy.2017.04.005.

[19] A.S. Ahmad, M.Y. Hassan, M.P. Abdullah, H.A. Rahman, F. Hussin, H. Abdullah, et al., A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renew. Sustain. Energy Rev. 33 (2014) 102–109. https://doi.org/10.1016/j.rser.2014.01.069.

[20] J.S. Chou, D.K. Bui, Modeling heating and cooling loads by artificial intelligence for energy-efficient building design, Energy Build. 82 (2014) 437–446. https://doi.org/10.1016/j.enbuild.2014.07.036.

[21] Y. Guo, J. Wang, H. Chen, G. Li, J. Liu, C. Xu, Machine learning-based thermal response time ahead energy demand prediction for building heating systems, Appl. Energy 221 (2018) 16–27. https://doi.org/10.1016/j.apenergy.2018.03.125.

[22] F. Kaytez, M. Taplamacioglu, E. Cam, F. Hardalac, Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines, Int. J. Electr. Power . Energy Syst. 67 (2015) 431–438. https://doi.org/10.1016/j.ijepes.2014.12.036.

[23] C. Deb, F. Zhang, J. Yang, S.E. Lee, K.W. Shah, A review on time series forecasting techniques for building energy consumption, Renew. Sustain. Energy Rev. 74 (2017) 902–924. https://doi.org/10.1016/j.rser.2017.02.085.

[24] S. Banihashemi, G. Ding, J. Wang, Developing a hybrid model of prediction and classification algorithms for building energy consumption, Energy Procedia 110 (2017) 371–376. https://doi.org/10.1016/j.egypro.2017.03.155.

[25] S. Paudel, M. Elmitri, S. Couturier, P.H. Nguyen, R. Kamphuis, B. Lacarrière, et al., A relevant data selection method for energy consumption prediction of low energy building based on support vector machine, Energy Build. 138 (2017) 240–256. https://doi.org/10.1016/j.enbuild.2016.11.009.

[26] S. Ahmadi-Karvigh, A. Ghahramani, B. Becerik-Gerber, L. Soibelman, Real-time activity recognition for energy efficiency in buildings, Appl. Energy 211 (2018) 146–160. https://doi.org/10.1016/j.apenergy.2017.11.055.

[27] J. Ma, J.C.P. Cheng, Identifying the influential features on the regional energy use intensity of residential buildings based on random forests, Appl. Energy 183 (2016) 193–201. https://doi.org/10.1016/j.apenergy.2016.08.096.

[28] P.A. Mathew, L.N. Dunn, M.D. Sohn, A. Mercado, C. Custudio, T. Walter, Big-data for building energy performance: lessons from assembling a very large national database of building energy use, Appl. Energy 140 (2015) 85–93. https://doi.org/10.1016/j.apenergy.2014.11.042.

[29] R.O. Duda, P.E. Hart, D.G. Stork, in: Pattern Classication, second ed., John Wiley & Sons, 2001, pp. 282–333. chapter 5.

[30] E. Gayawan, A comparison of akaike, schwarz and r square criteria for model selection using some fertility models, Australia J. Basic Appl. 3 (2009) 3524–3530. http://insipub.net/ajbas/2009/3524-3530.pdf.

[31] Y. Fu, Z. Li, H. Zhang, P. Xu, Using support vector machine to predict next day electricity load of public buildings with sub-metering devices, Proc. Eng. 121 (2015) 1016–1022. https://doi.org/10.1016/j.proeng.2015.09.097.

[32] S. Touzani, J. Granderson, S. Fernandes, Gradient boosting machine for modeling the energy consumption of commercial buildings, Energy Build. 158 (2018) 1533–1543. https://doi.org/10.1016/j.enbuild.2017.11.039.

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

[34] M.D. Ugarte, A.T. Militino, A.T. Arnholt, Probability and Statistics With R, CRC Press, Boca Raton, FL, 2016.

[35] http://rstudio-pubs-static.s3.amazonaws.com/13750_6e070573890a4d1d9c16ffb53f5ddfd2.html.

[36] K. Zhou, C. Fu, S. Yang, in: Big Data Driven Smart Energy Management : fFrom Big Data to Big Insights, 56, 2016, pp. 215–225. https://doi.org/10.1016/j.rser.2015.11.050.

[37] http://www.shjzjn.org/#/home.

[38] K.L. Gillespie, J.D. Cowan, C.W. Frazell, J.S. Haberl, K.H. Heinemeier, J.P. Kummer, et al., ASHRAE GUIDELINE 14-2002, Measurement of Energy and Demand Savings, 8400, 2002.

[39] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition, Springer Series in Statistics Trevor, 2017.

[40] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006 Chapter 2 https://doi.org/10.1142/S0129065704001899.