K-Nearest Neighbors (KNN) Algorithm for Energy Prediction Models

Method · July 2021

CITATIONS
0

READS
803

1 author:

Tarannom Parhizkar
University of California, Los Angeles
72 PUBLICATIONS **519** CITATIONS

SEE PROFILE

Journal of Cleaner Production 279 (2021) 123866

Contents lists available at ScienceDirect

Journal of Cleaner Production

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

Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction

Tarannom Parhizkar [a, *], Elham Rafieipour [b], Aram Parhizkar [c]

[a] Sharif Energy Research Institute, Sharif University of Technology, Tehran, Iran
[b] Department of Mechanical Engineering, Shahid Chamran University of Ahvaz, Khuzestan, Iran
[c] Faculty of Natural Science, Tabriz University, Iran

ARTICLE INFO

Article history:
Received 18 April 2020
Received in revised form
15 August 2020
Accepted 18 August 2020
Available online 23 August 2020

Handling editor. Bin Chen

Keywords:
Energy consumption prediction
Data analytics
Principal component analysis
Regression tree
K-nearest neighbors
Support vector regression
Random forest
Linear regression

ABSTRACT

The building sector is a major source of energy consumption and greenhouse gas emissions in urban regions. Several studies have explored energy consumption prediction, and the value of the knowledge extracted is directly related to the quality of the data used. The massive growth in the scale of data affects data quality and poses a challenge to traditional data mining methods, as these methods have difficulties coping with such large amounts of data. Expanded algorithms need to be utilized to improve prediction performance considering the ever-increasing large data sets.

In this paper, a preprocessing method to remove noisy features is coupled with predication methods to improve the performance of the energy consumption prediction models. The proposed preprocessing method is based on the well-known principal component analysis (PCA) and treats the historical meteorological and energy data of buildings. The cleaned and processed data are used in five prediction models including linear regression, support vector regression, regression tree, random forest and K-nearest neighbors.

The proposed methodology is applied to four case studies with different climate zones (cold, mild, warm-dry and hot-humid) to study the effect of dataset patterns on the feature reduction and prediction performance. The results show that the proposed method enables practitioners to efficiently acquire a smart dataset from any big dataset for energy consumption prediction problems. In addition, the best prediction model for each climate zones with considering mean square error, R^2, residual values and execution time is proposed.

1. Introduction

The building sector is one of the largest consumers of energy (39–40%) and emitters of greenhouse gases (38%–39%) in the world (Becerik-Gerber et al., 2013). Energy consumption prediction (Pham et al., 2020) and monitoring (Parhizkar et al., 2019, 2020) in buildings helps to increase the effectiveness and efficiency of decisions made to reduce energy demand and carbon emissions. Energy consumption models are mostly used in the first step of energy management and efficiency improvement models, such as optimizing operations and reducing costs (Li et al., 2020); sizing thermal energy storage to improve energy efficiency (Lin et al., 2019); modeling heating, ventilation and air conditioning (HVAC) systems to reduce energy consumption (Chen et al., 2020); and planning urban energy systems (Moghadam et al., 2017).

Two approaches have been used to predict energy consumption: principle-based modeling (white box) and data-driven modeling (black box) (Guermoui et al., 2020). In the principle-based approach, the inputs can be weather features, geographic locations, building designs, building material properties, occupancy characteristics and operating schedules, and the outputs are building load estimates (Gan et al., 2020). EnergyPlus, Ecotect, and eQuest are energy simulation software tools used to predict buildings' energy consumption. Principle-based models have a high accuracy in modeling buildings' energy consumption (Gan et al., 2020). However, due to lack of access to building design and material characteristics, principle-based approaches can be uncertain and time consuming.

In contrast, data-driven models, which has received attention in

* Corresponding author. Teymori Blvd., TarashtSharif Energy Research Institute, Sharif University of Technology, Tehran, P.O. Box: 1459777611, Iran.
E-mail addresses: Parhizkar.t@gmail.com (T. Parhizkar), Elham.rfp@gmail.com (E. Rafieipour), Parhizkarr.a@gmail.com (A. Parhizkar).

https://doi.org/10.1016/j.jclepro.2020.123866
0959-6526/

List of abbreviations

GB	Gradient boosting
GBM	Gradient boosting machine
HVAC	Heating, ventilation and air conditioning
KNN	K-nearest neighbors
MLR	Multiple linear regression
MNN	Multilayer neural network
MSE	Mean square error
PC	Principle component
PCA	Principal component analysis
SHGC	Solar heat gain coefficient
SVM	Support vector machine
SVR	Support vector regression

recent years, no longer requires the building design and material characteristics. These models base the predictions on the system historical data (Parhizkar et al., 2017, 2018). One of the basic data-driven prediction methods is the linear regression model. In (Iwafune et al., 2014), linear regression method is utilized to predict daily energy consumption of residential buildings. Outside temperature and date are considered as two affecting factors on energy consumption. The model is used to perform demand side management of residential buildings. In (Zhang et al., 2020), multivariate regression method is used to predict energy consumption for optimal design of building environment. In (Alobaidi et al., 2018), a regression model for predicting the average daily energy consumption of individual households is proposed. This framework utilizes information diversity to predict the day-ahead average energy consumption. To further enhance generalizability, a robust regression component was proposed. The proposed method was applied to a case study in France, and the results illustrate significant improvement in alleviating the unstable prediction problems that exist in other models. Afroz et al. (2018) developed an indoor temperature prediction model for commercial buildings. In this study, a nonlinear autoregressive network that considers exogenous input-based system identification was proposed to predict indoor temperatures. The optimal input parameters, size of network, and size of training data affected the performance of the model. Using sensitivity analysis, the researchers proposed a model that provided an accurate prediction for up to 28 days ahead.

In addition to regression methods, other machine learning algorithms can be used to predict building energy consumption (Zhou et al., 2016). Artificial neural networks (ANNs) are one of the most popular data-driven energy predication models that are designed based on the basic functions of human brain including processing units and biological neurons. A network consists of one or multiple processing units arrayed in layer, which are connected via connections. The method is presented comprehensively in (Bagnasco et al., 2015). Using this method, the hourly overall (Ilbeigi et al., 2020), cooling (Luo, 2020) and heating (Bui et al., 2020) energy consumption of buildings could be predicted. In (Rahman et al., 2018), a recurrent neural network model is proposed to predict the energy consumption of commercial and residential buildings. In this study, a deep neural network was used to perform imputation on datasets containing segments of missing values. The method was applied to datasets for a commercial building and a residential building. The results illustrate that the recurrent neural network model corresponded to a lower relative error compared to a conventional multi-layered perceptron neural network in the commercial building. However, in the residential building, the proposed model did not provide high accuracy in comparison to the multi-layered perceptron model.

Another popular data-driven method is the support vector regression (SVR). Multiple studies have used this method to predict hourly cooling (Li et al., 2009), heating (Chou and Bui, 2014) and overall (Shao et al., 2020) energy consumption of buildings. For instance, Ma et al. (2019) proposed an SVR model to predict building energy consumption in southern China. Multiple features, including weather data and economic factors, were taken as inputs, and the prediction model performance was evaluated using data provided by the Chinese National Bureau for four provinces of southern China. The results indicated that the SVR method has a high accuracy in predicting building energy consumption.

Random forest is one of the most widely used decision tree methods in the field of buildings energy consumption prediction. In (Fan et al., 2014), daily energy consumption of a non-residential building is predicted using random forest method. Maximum dry-bulb temperature, average dry-bulb temperature, minimum dry-bulb temperature, average dew point temperature, average relative humidity, average pressure, average amount of cloud, total rainfall, number of hours of reduced visibility, solar radiation, total evaporation and average wind speed are considered as affecting factors. Data of one year is used to train the model and resulted in 3.17% mean absolute percentage error. K-nearest neighbors is another statistical algorithm that has been used in this study for energy consumption prediction. This method could predict overall energy consumption of the building with 4.01% mean absolute percentage error.

There are several studies that have compared data-driven models in different case studies. For instance, Candanedo et al. (2017) presented a data-driven predictive model to predict electricity loads. This study compared four data-driven methods: multiple linear regression (MLR), support vector machine (SVM) with radial kernel, the random forest approach and gradient boosting machines (GBMs). The results showed that the GBM method, with a variance of 97% (R^2) in the training set and 57% in the testing set, was the most efficient when all predictors were used. Guo et al. (2018) compared four machine learning methods to predict the energy demand of building heating systems: support vector regression (SVR), MLR, the extreme learning machine approach and a backpropagation neural network. Data on building heating using a ground source heat pump system were used to test and compare the performances of the models, which take meteorological parameters, operating parameters, time and indoor temperature parameters as inputs. Their results indicated that the performance of an extreme learning machine model with 11 hidden layer nodes and feature set 4 is better than the other methods. In (Gassar et al., 2019), multiple machine learning methods for predicting gas and electricity consumption in London's residential buildings are compared. The study considered the multilayer neural network (MNN), MLR, random forest and gradient boosting (GB) methods, and the input features examined were socio-demographic, economic and building characteristics. The results show that household income, number of households and building characteristics are the most important features of gas and electricity consumption. The MNN models outperformed the MLR, random forest and GB models at predicting energy consumption by London's residential buildings. In another study conducted by Gungor et al. (2019) four years of household electricity consumption data are used to compare various machine learning methods, including the random forest, K-nearest neighbors (KNN), stochastic gradient descent, logistic regression and SVM approaches. The results indicated that the random forest method had the lowest prediction error.

In almost all of these studies, an efficient method for building energy consumption prediction is proposed according to the results

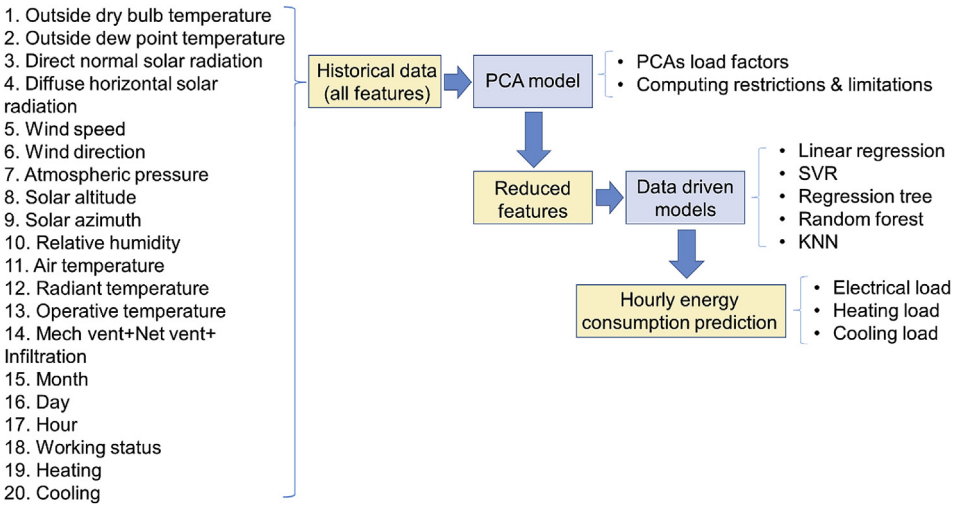


Fig. 1. Data flow diagram of the PCA-based energy consumption prediction method.

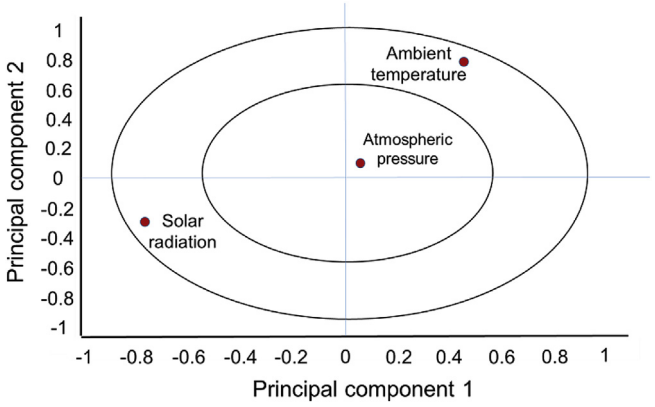


Fig. 2. A PCA loading plot as an example.

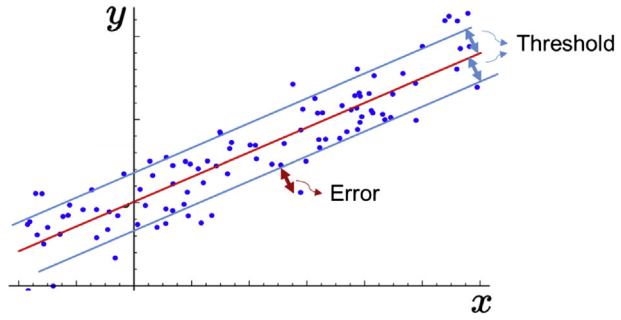


Fig. 4. Threshold and error definition in the SVR method for an example scatter plot.

of a case study. However, prediction methods are highly dependent on the historical data of the buildings studied, and it is not possible to propose a general method applicable to all buildings of the same type (commercial, residential, etc.). In our study, the dependence of the method's performance on the energy consumption data pattern is clarified.

In addition, most of the reviewed literature has focused on using data-driven methods to predict energy consumption. To develop an accurate model, most of the building features should be considered, including the time, outside dry bulb temperature, outside dew point temperature, direct normal solar radiation, diffuse horizontal solar radiation, wind speed, wind direction, atmospheric pressure, solar azimuth, relative humidity, air temperature, radiant temperature and operative temperature. Technology that can accurately monitor, collect and store the vast amount of data involved in this process is now available. Recent literature on this topic indicates that while the available methods are able to predict energy consumption, there is still room for improvement. Specifically, as the amount of data increases over time, the available methods tend to predict energy consumption with lower accuracy or much higher execution times. As a result, preprocessing techniques should be utilized to aid data processing in prediction models (Lin et al., 2020). In this study, the principal component analysis (PCA) method was used to identify the features with the strongest effect on energy consumption. Principal component analysis is a dimensionality-reduction method that is often used to reduce the dimensionality of large datasets by transforming a large set of variables into a smaller one that still contains most of the information of the large set. This method was used to find the factors with the greatest impact on energy consumption prediction.

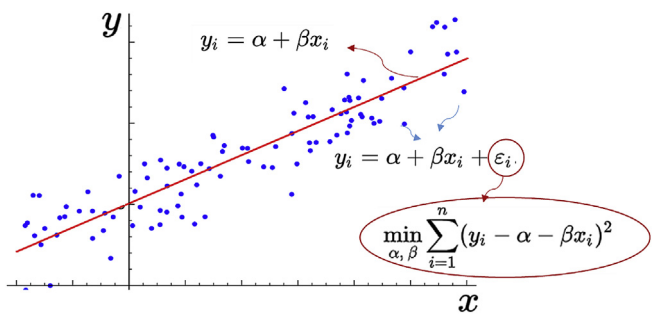
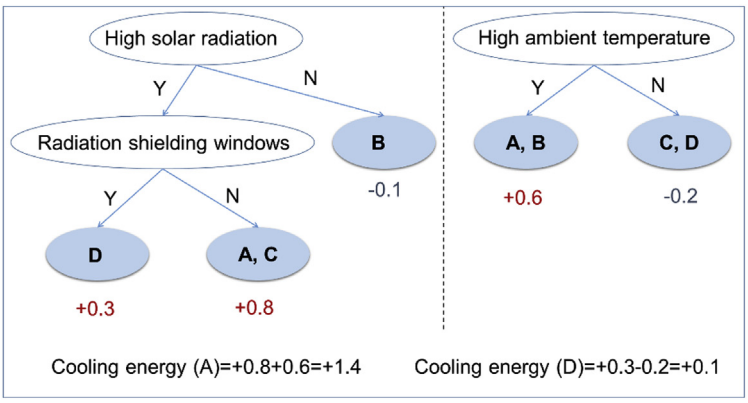


Fig. 3. A simple linear regression plot for an example scatter plot.

This study utilizes the PCA as a preprocessing method to assist five energy consumption prediction models, widely used in this field. The integrated PCA based prediction models are applied to four types of energy data patterns, and the results are compared. Results show that in all cases, the PCA method helps to identify the most important features for energy consumption prediction and improves prediction model performance. The main contributions of this research could be summarized as follows:



Buildings description

Building	A	B	C	D
Building Characteristic (Shielding window)	No	-	No	Yes
Ambient temperature	Hot	Hot	Cold	Cold
Solar radiation	Yes	No	Yes	Yes

Fig. 5. A regression tree example for cooling energy consumption prediction.

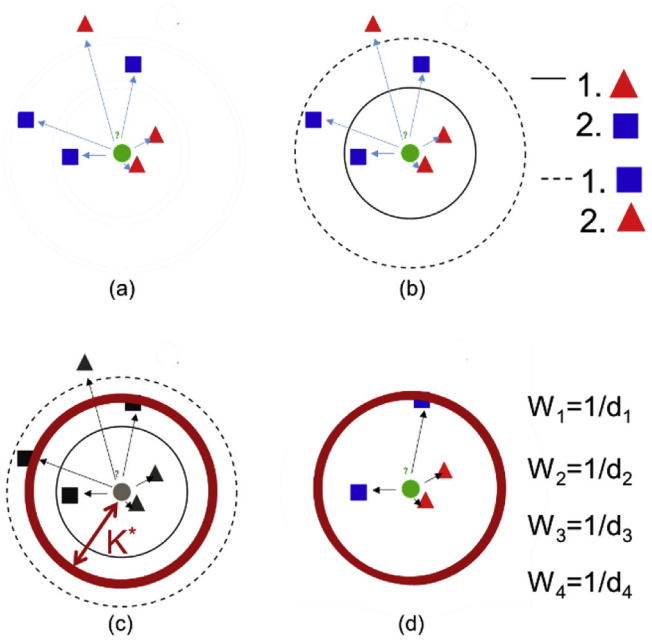


Fig. 6. Four main steps of the KNN algorithm.

1. In this study, the dependence of prediction models' performance on the historical data, in addition to the building type, is illustrated. In other words, this study demonstrates that in addition to the building type, the climate zone should be considered in selecting the prediction model. This finding could assist in the development of hybrid methods in online prediction models that could be updated over time based on a building's historical data (i.e., the energy prediction model of the building could be continuously updated and changed in accordance with the data gathered over time).
2. As mentioned, there is a growing need for energy prediction models that can monitor and analyze building energy performance. However, the continuously growing complexity of energy systems and the ever-increasing amount of data make this process difficult. This study assessed the effectiveness of PCA data reduction method on energy consumption prediction models. It is shown that PCA method could be used as a powerful feature reduction method in energy prediction models. This method reduces the time and storage space required, while it improves the performance of energy consumption prediction models.
3. Many prediction methods can be used to estimate building energy consumption. The five most effective methods, according to the reviewed literature, are selected and compared in relation

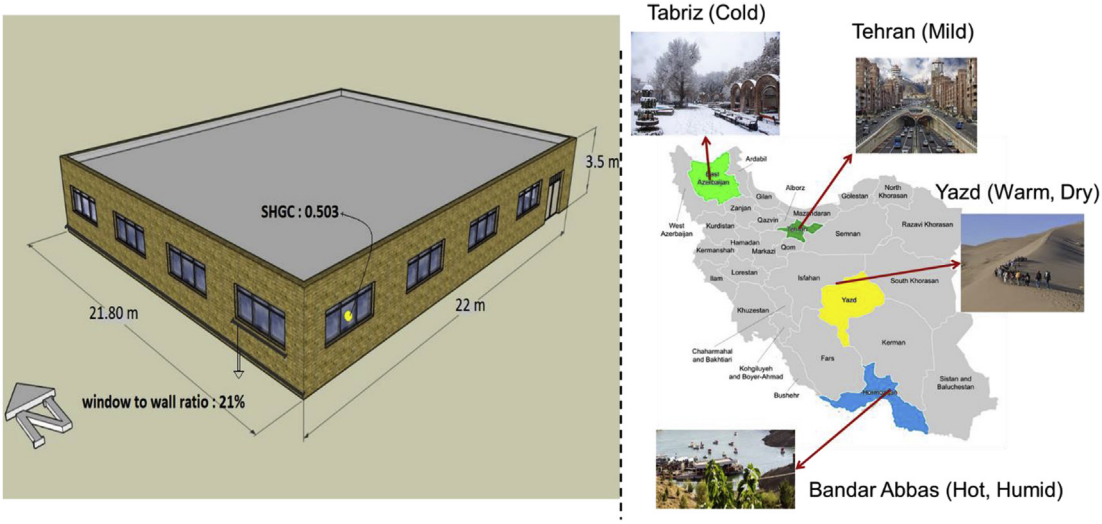


Fig. 7. Schematic of the studied building and climate zones.

Table 1
The construction properties of the studied building.

Component	Construction properties
Wall	Gypsum (13 mm) + concrete block (100 mm) + polystyrene (20 mm) + brickwork outer (100 mm)
Roof	Gypsum (13 mm) + air gap (25 mm) + concrete (300 mm) + wool + asphalt (10 mm)
Window	Generic clear (6 mm) + air (6 mm) + generic blue (6 mm)

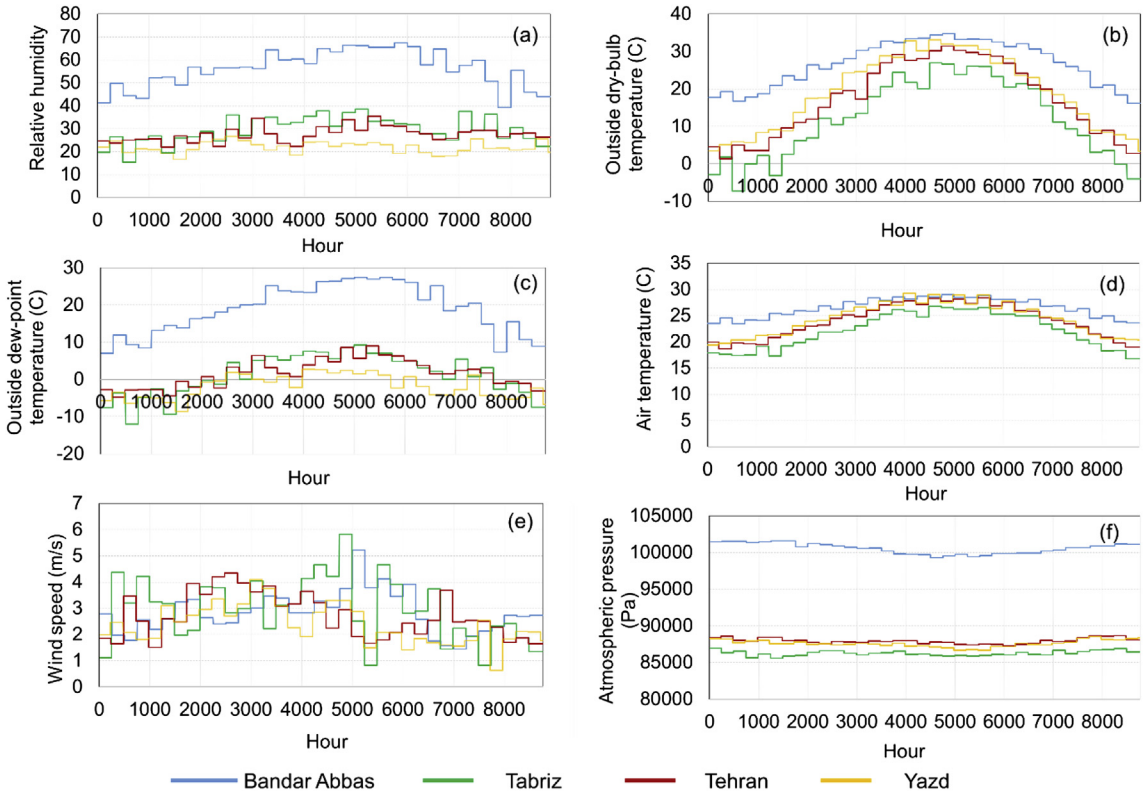


Fig. 8. Relative humidity (a), dry bulb temperature (b), dew point temperature (c), ambient temperature (d), wind speed (e) and atmospheric pressure (f) over a year in the studied climate zones.

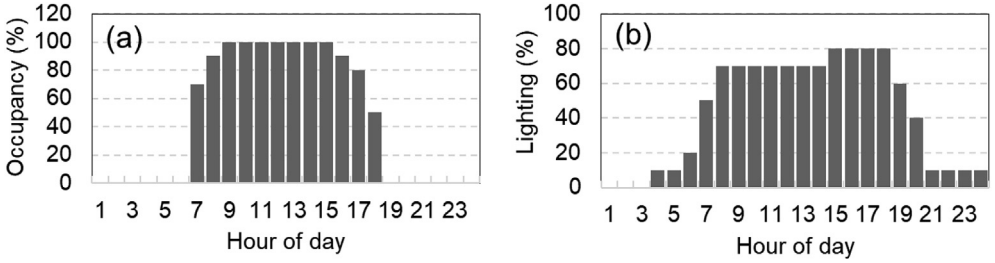


Fig. 9. Occupancy (a) and lighting (b) daily profiles.

to four different climate zones. On that basis, the most efficient PCA-based energy prediction method for each climate zone is proposed.

2. Research framework and methodology

Fig. 1 presents the data flow diagram of the PCA-based energy consumption prediction method, displaying its two main steps. In the first step, the historical data of a building are taken as inputs for the PCA model. The historical data consist of 20 features, as presented in Fig. 1. In the PCA model, the most influential features are selected based on the model's load factors and problem restrictions. In the second step, the reduced features are utilized to predict energy consumption using multiple prediction models, including linear regression, SVR, regression tree, random forest and KNN models. In this step, the most efficient prediction model is selected based on factors such as the mean square error (MSE) and R^2 values. Finally, the selected model is utilized to predict the hourly energy consumption of the building.

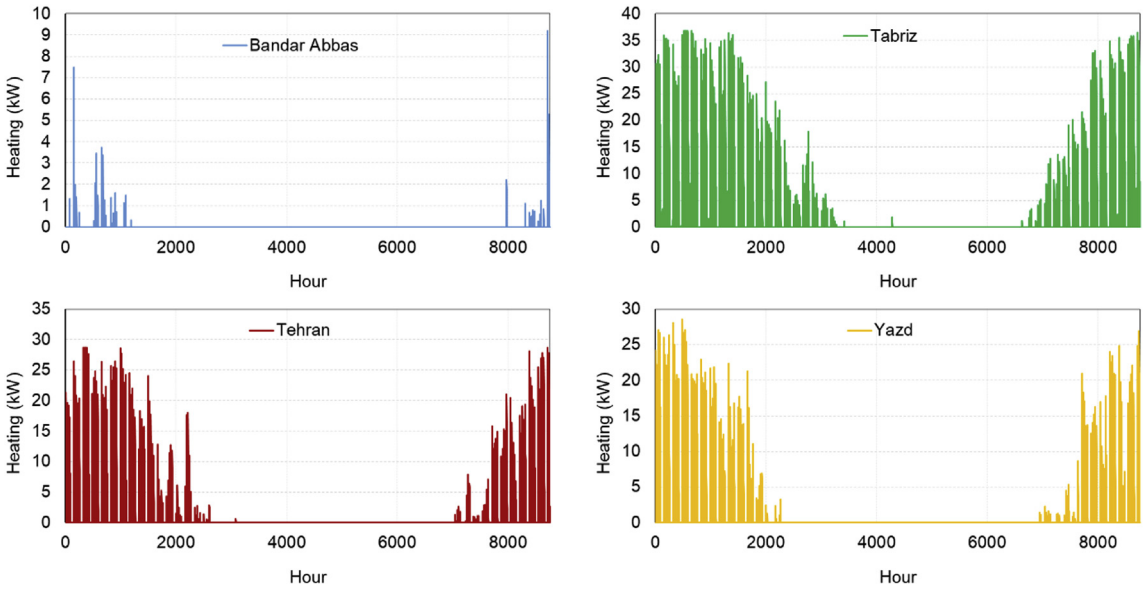


Fig. 10. Hourly heating energy consumption in the case studies.

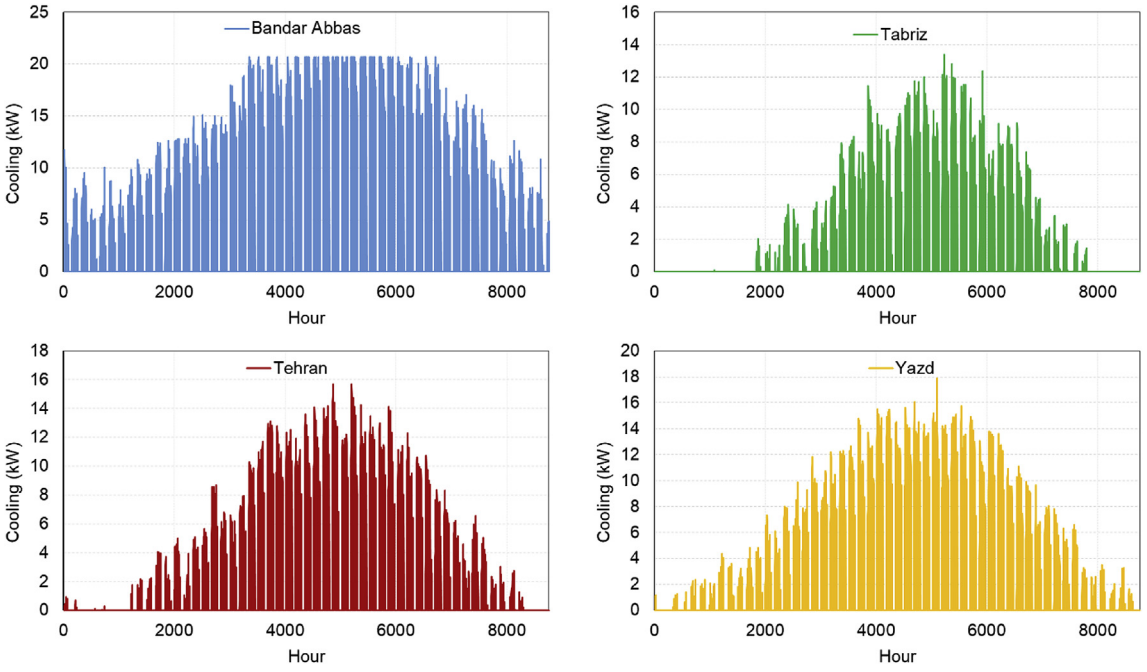


Fig. 11. Hourly cooling energy consumption in the case studies.

2.1. The principal component analysis (PCA) model

As shown in Fig. 1, the features of the historical meteorological data were extracted using statistical methods. These data are 20 features of a building's energy consumption, as presented in Fig. 1. The historical values of these features were used as inputs in the PCA model.

Principal component analysis is a mathematical procedure that transforms a number of (possibly) correlated variables into a smaller set of uncorrelated variables called principal components (Skjærvold et al., 2006). As one of the effective factor analysis method, PCA is widely used to reduce the number of variables under study, and consequently the ranking and analysis of decision-making units (Ghaderi et al., 2006).

Primarily, PCA decomposes data matrix X into a structure and a noise part. As presented in Eq. (1), data matrix $X\,(n \times k)$ is split into a modelled part M_X $(n \times k)$ and a residual error part E $(n \times k)$ (Huang et al., 2019).

$$X = M_X + E \tag{1}$$

The modelled part of X is expressed as a subspace with dimensionality A, where A represents the number of principal components. Consequently, when the chosen model dimensionality A is changed, the error content also varies.

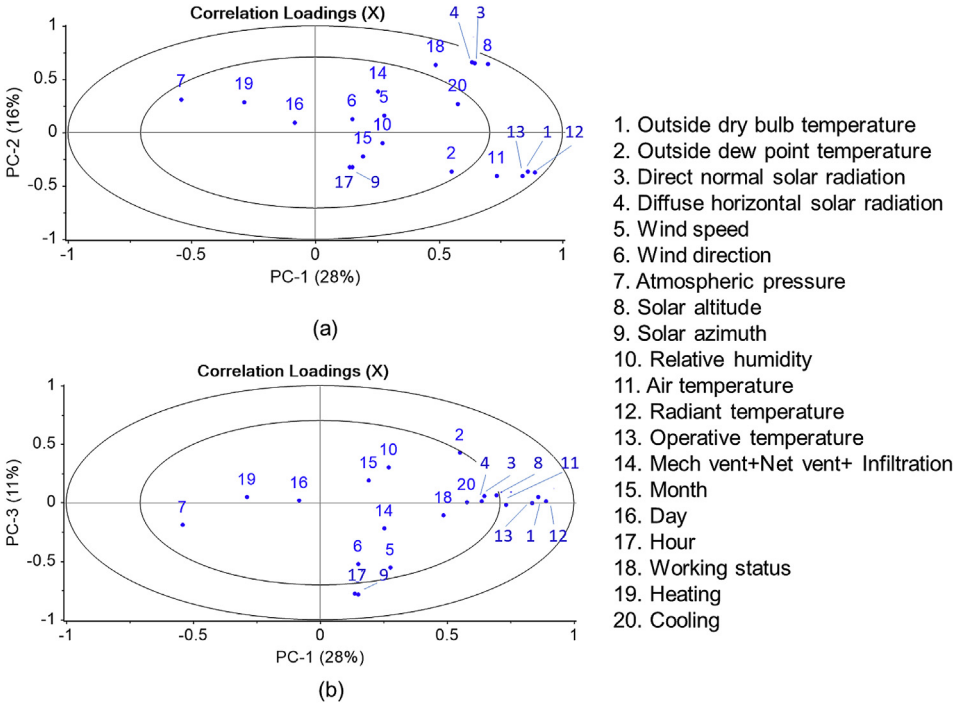


Fig. 12. Loading plot for the first and second components (a) and for the first and third components (b) of the Tehran dataset.

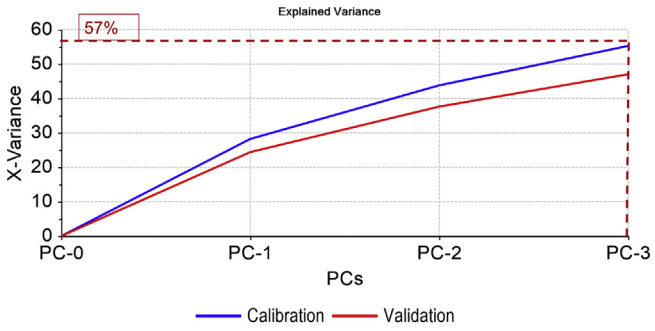


Fig. 13. Cumulative explained variance versus the number of principle components.

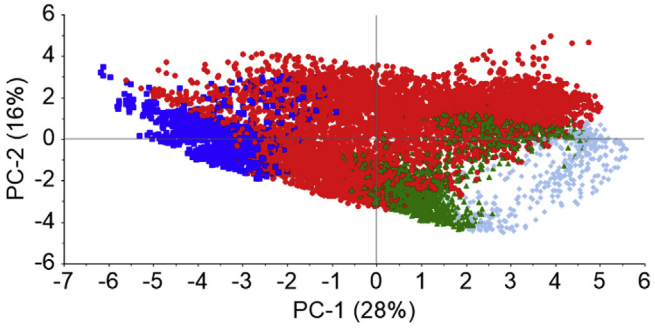


Fig. 14. The scatter score plot of PC-2 vs PC-1 for Tehran dataset.

$$X = M_X,\ A + EA \tag{2}$$

The historical data determine how many principal components are needed to have an accurate prediction model. The principal components are used to reduce features to a smaller number. The reduction process can be performed using a PCA loading plot of principal components (Gao et al., 2013). A loading plot shows how strongly each characteristic influences a principal component.

Fig. 2 presents a loading plot for a case as an example. Loadings can range from −1 to 1. Loadings close to −1 or 1 indicate that the variable strongly influences the component. Loadings close to 0 indicate that the variable has a weak influence on the component. Evaluating the loadings can also help one to characterize each component in terms of the variables.

In this example, the ambient temperature has a large positive loading and the solar radiation has a large negative loading on components 1 and 2. As a result, these features are more critical than the atmospheric pressure, which has a low loading value. Hence, atmospheric pressure can be eliminated in case of data reduction.

2.2. Linear regression

Linear regression is a basic and commonly used predictive model. Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. Fig. 3 shows a scatter plot of sample data. The red diagonal line is the regression line and consists of the predicted *Y* value for each possible value of *X*. As shown, *Y* can be predicted as a function of *X*. More precisely, each data point can be represented by using this equation plus an error. The distance between the data and the line represents the prediction error. The linear regression algorithm finds the constant values of the linear equation by minimizing these errors for all sample data (Harrell, 2015).

2.3. Support vector regression (SVR)

As the name suggests, SVR is a modified regression algorithm. In simple regression, we try to minimize the error rate, while with SVR we try to fit the error within a certain threshold. According to the regression model, SVR could be linear or nonlinear. Fig. 4

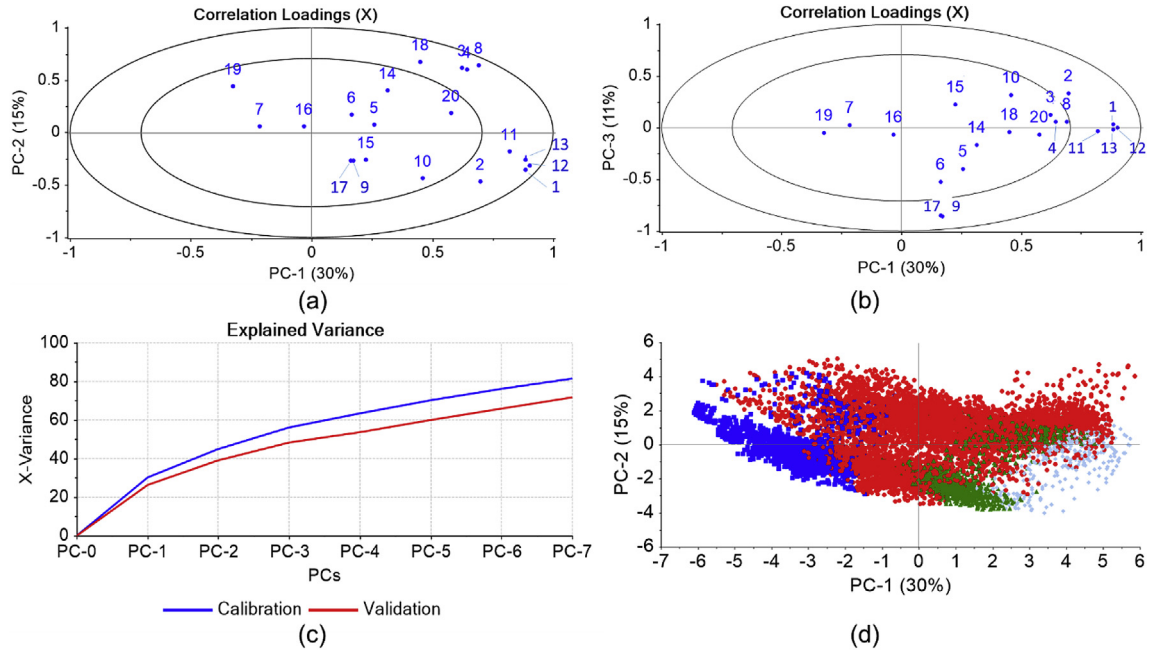


Fig. 15. Loading (a, b), cumulative explained variance (c) and scatter score (d) plots for the Tabriz dataset.

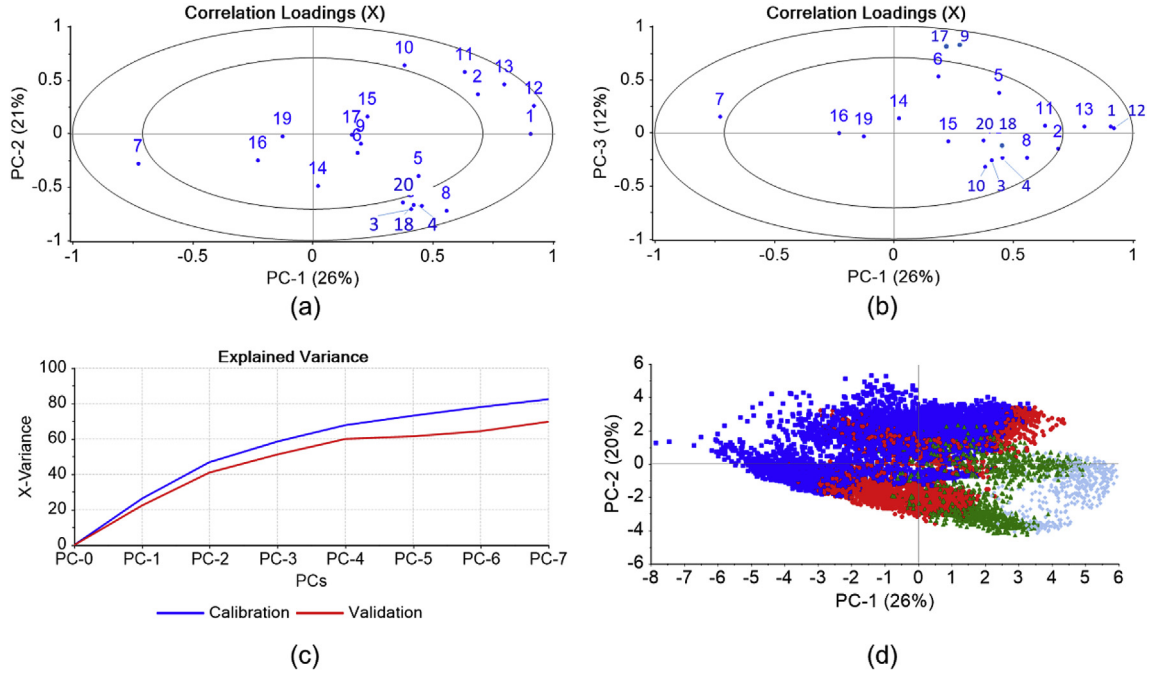


Fig. 16. Loading (a, b), cumulative explained variance (c) and scatter score (d) plots for the Bandar Abbas dataset.

presents a linear SVR example. As can be seen, the error is calculated from the threshold lines. This is the main difference between the regression method and SVR, but there are other rules that differentiate these two methods (Smola and Schölkopf, 2004).

2.4. Regression tree

A regression tree builds a regression model in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets, while at the same time incrementally developing an associated decision tree. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., energy consumption) has two or more branches (e.g., low, medium and high), each representing values for the attribute tested. A leaf node represents a decision on the numerical target (Loh, 2008).

The regression tree algorithm has two main parts. First, a decision tree is developed; in this step, data are branched according to some indexes. The next step is a regression model that is developed for each branch of the decision tree. This step mostly follows regression model rules. Fig. 5 presents a simple example of a

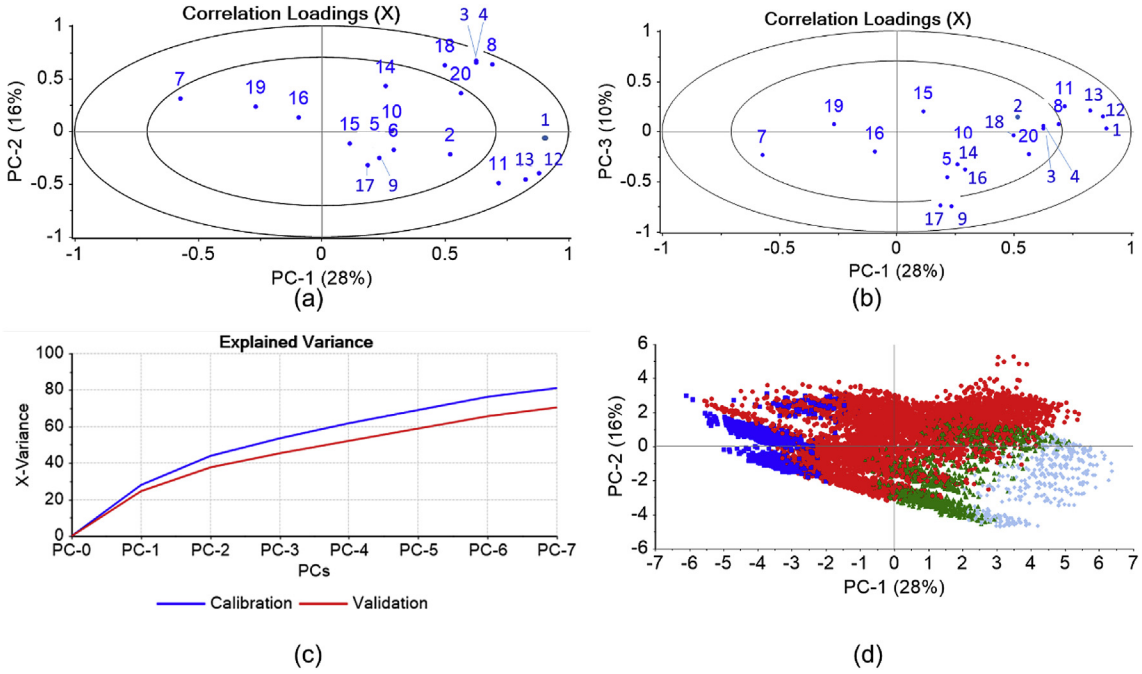


Fig. 17. Loading (a, b), cumulative explained variance (c) and scatter score (d) plots for the Yazd dataset.

cooling energy consumption prediction model that follows the regression tree method. This example consists of four buildings with different characteristics (with and without radiation shielding windows) and different climate zones (high and low ambient temperature, with and without high solar radiation). The energy consumption of buildings A and D is presented in the figure.

2.5. Random forest

Random forest is an ensemble classifier developed by constructing a multitude of decision tree models during the training phase from a randomly chosen subset of the training set to obtain a better predictive performance. Ensemble models combine the results from diverse models. The results from an ensemble model are usually better than the results from an individual model. This model then aggregates the votes from different decision trees to make the final decision for the test object. A subset of data is attributed to each tree in the random forest algorithm (Liaw and Wiener, 2002). For example, if 2000 rows (sample) and 50 columns (feature) are created, 200 rows and 20 columns, which are a subset of the data, are assigned to each tree. By using this data subset, these trees can decide and generate the training model.

2.6. K-nearest neighbors (KNN) algorithm

The KNN algorithm is a non-parametric method used for data prediction. In this algorithm, data are labeled (Soucy and Mineau, 2001) and prediction is performed based on the labels. Fig. 6 presents a simple example of the KNN algorithm. At the first step, the Euclidean or Mahalanobis distance from the query example to the labeled examples is calculated. In the figure, these distances are shown by light blue arrows. Next, the labeled examples are ordered by increasing distance (for instance, the label represented by triangles in the solid line circle). By increasing the distance (K), the order changes as the number of squares in the dashed-line circle is higher. In the third step, a heuristically optimal number K of nearest neighbors based on the *RMSE* should be found. This process is carried out using cross-validation. Last, an inverse distance weighted average with the k-nearest multivariate neighbors is calculated, and the unknown data are labeled accordingly.

3. Case study

To study the effectiveness of the proposed method, we applied the proposed method to four datasets. As a result, the sensitivity of the method to different input datasets is detected. The selected four datasets are weather and energy data of buildings in four different climate zones.

Datasets include hourly weather (relative humidity, dry bulb temperature, dew point temperature, ambient temperature, wind speed and atmospheric pressure) and hourly energy consumption (cooling and heating) data. In this section, the building characteristics, climate conditions of the studied zones and energy consumption of the office buildings located in these zones are presented.

3.1. Building information

An office building in four different climate zones, as shown in Fig. 7, was considered as a case study to evaluate the performance of the proposed methodology.

The size of the building is 21.80 m × 22 m × 3.5 m. The solar heat gain coefficient (SHGC) and glazing U-value are 0.503 and 3.094 ($W/m^2.K$), respectively. The window-to-wall ratio is 21%. The occupancy density, lighting power density and equipment power density are 18.6 (W/m^2), 19.5 (W/m^2) and 10.8 (W/m^2), respectively. The HVAC system of the building is a four-pipe fan coil unit. Four-pipe systems have separate heating and cooling fan coil units, which means that hot or chilled water is always available, enabling the system to immediately change over from heating to cooling mode. The construction properties of the building are summarized in Table 1.

Table 2
The least influential variables presented in PCA loading plots.

Climate zone	PC1 and PC2	PC1 and PC3	Common
Tehran	Atmospheric pressure	Atmospheric pressure	Atmospheric pressure
	Mech vent + net vent	Mech vent + net vent	Mech vent + net vent
	Outside dew point temp	Month	Day
	Day	Relative humidity	Wind direction
	Wind direction	Day	Month
	Hour	Wind direction	Relative humidity
	Solar azimuth	Wind speed	Wind speed
	Month	Working status	
	Relative humidity	Diffuse horizontal solar	
	Wind speed	Direct normal solar	
		Solar altitude	
Tabriz	Atmospheric pressure	Atmospheric pressure	Atmospheric pressure
	Day	Day	Day
	Hour	Wind direction	Month
	Solar azimuth	Wind speed	Relative humidity
	Month	Mech vent + net vent	Wind speed
	Relative humidity	Working status	Wind direction
	Wind speed	Month	Mech vent + net vent
	Wind direction	Relative humidity	
	Mech vent + net vent	Direct normal solar	
		Diffuse horizontal solar	
		Solar altitude	
Bandar Abbas	Day	Day	Day
	Mech vent + net vent	Mech vent + net vent	Mech vent + net vent
	Month	Wind direction	Month
	Wind speed	Wind speed	Wind speed
	Hour	Air temperature	Wind direction
	Solar azimuth	Month	
	Wind direction	Working status	
		Relative humidity	
		Direct normal solar	
		Diffuse horizontal solar	
		Solar altitude	
		Dew point temp	
Yazd	Atmospheric pressure	Atmospheric pressure	Atmospheric pressure
	Day	Day	Day
	Mech vent + net vent	Month	Mech vent + net vent Relative humidity
	Relative humidity	Outside dew point temp	Month
	Month	Direct normal solar	Wind speed
	Wind speed	Diffuse horizontal solar	Wind direction
	Wind direction	Solar altitude	Outside dew point temp
	Solar azimuth	Working status	
	Hour	Relative humidity	
	Outside dew point temp	Mech vent + net vent	
		Wind direction	
		Wind speed	

Table 3
Comparison of five prediction models for the Tehran dataset.

Prediction model/metric	$MSE\ (kW)^2$		R^2		Time(s)
	Train	Test	Train	Test	
Linear regression	11.56	12.69	0.68	0.64	1.66
SVR	9.02	12.59	0.75	0.64	10.62
Regression tree	0.67	1.18	0.98	0.97	0.94
Random forest	0.15	0.39	1.00[a]	1.00[b]	6.45
KNN	10.20	15.60	0.72	0.56	1.18

[a] The actual value is 0.99712 that is rounded up to 1.
[b] The actual value is 0.99601 that is rounded up to 1.

Table 4
Comparison of five prediction models for the Tabriz dataset.

Prediction model/metric	$MSE\ (kW)^2$		R^2		Time(s)
	Train	Test	Train	Test	
Linear regression	18.96	20.59	0.67	0.64	1.66
SVR	13.68	20.12	0.76	0.65	10.98
Regression tree	0.93	2.41	0.98	0.96	0.93
Random forest	0.24	1.10	0.99	0.98	7.13
KNN	15.62	24.45	0.72	0.57	1.20

3.2. Weather data

Four climate zones are considered to compare the effectiveness of the method. The hourly climate data (relative humidity, dry bulb temperature, dew point temperature, ambient temperature, wind speed and atmospheric pressure) of these four zones are derived from a weather forecast website [1] Fig. 8(a) displays the hourly relative humidity variation over a year in different cities. The highest humidity level is seen in Bandar Abbas in warm seasons. In general, as demonstrated, while Bandar Abbas is the most humid city among those studied, Yazd normally experiences the driest weather conditions during the year. Moreover, the range of relative humidity change is more visible in Bandar Abbas and Tabriz than in the other cities. However, the variation of relative humidity in Bandar Abbas is clearly visible compared to the other cities.

The hourly dry bulb, dew point and air temperature variations over a year in different cities are shown in Fig. 8(b), (c) and (d).

[1] https://www.weather-forecast.com/countries/Iran.

Table 5
Comparison of five prediction models for the Bandar Abbas dataset.

Prediction model/metric	$MSE\ (kW)^2$		R^2		Time(s)
	Train	Test	Train	Test	
Linear regression	5.05	5.64	0.93	0.92	2.01
SVR	11.63	15.91	0.86	0.79	10.33
Regression tree	0.48	0.68	0.99	0.99	1.01
Random forest	0.13	0.28	1.00[a]	1.00[b]	6.25
KNN	15.66	24.03	0.79	0.68	1.26

[a] The actual value is 0.99830 that is rounded up to 1.
[b] The actual value is 0.99691 that is rounded up to 1.

Table 6
Comparison of five prediction models for the Yazd dataset.

Prediction model/metric	$MSE\ (kW)^2$		R^2		Time(s)
	Train	Test	Train	Test	
Linear regression	10.17	11.00	0.75	0.72	1.99
SVR	9.68	14.57	0.76	0.66	10.94
Regression tree	0.48	1.03	0.99	0.97	1.15
Random forest	0.13	1.00	0.28	0.99	6.24
KNN	10.93	17.78	0.73	0.55	1.33

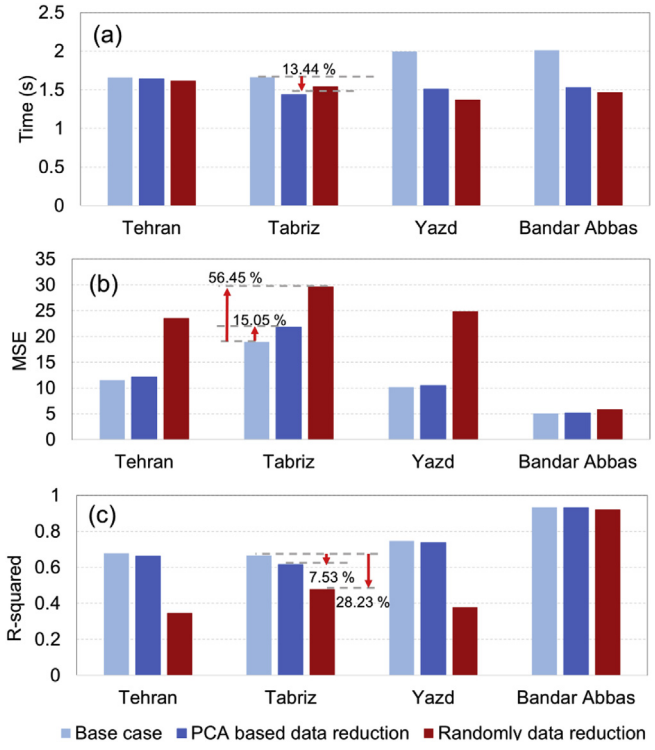


Fig. 18. The execution time (a), *MSE* (b) and R^2 (c) of the linear regression model for the three datasets and four climate zones.

Bandar Abbas has the highest dry bulb temperature, followed by Yazd, Tehran and Tabriz. While the warmest time of the year is seen in Bandar Abbas in July, the coldest time of year among the selected cities is found in Tabriz in January. It should be noted that in addition to the ambient temperature, seasonal temperature differences also influence the total energy consumption of buildings. The largest seasonal temperature difference is observed in Tabriz, with a 27 °C temperature difference between winter and summer. Bandar Abbas experiences the lowest seasonal temperature difference, with an almost 18 °C temperature difference between winter and summer. However, the temperature range changes in Tehran and Yazd are approximately identical.

As can be seen in Fig. 8(e), wind speed, unlike other environmental parameters, does not follow a specific pattern. For instance, in Tehran, the greatest variation in wind speed occurs in February to August, whereas in Yazd, it occurs between January and June. This phenomenon is explained by the geographic location of Iran.

3.3. Behavioral patterns

The real office building daily profiles including occupancy and lighting are used to set the schedules that influence the energy consumption of the building. The specific behavioral patterns for the occupancy (Duarte et al., 2013) and lighting (Li and Lam, 2001; Jiang et al., 2018) profiles are shown in Fig. 9(a) and (b).

3.4. Building energy data

The monthly electricity and gas consumption data of the buildings are collected from the available electricity and gas bills for the past 4 years. Working days, number of people in each building, and the type of cooling/heating systems are determined using the questionnaires that was filled by building's user personal. Lastly, the collected energy consumption data are simply converted from monthly to an hourly basis.

Figs. 10 and 11 show the hourly energy consumption for heating and cooling in the studied climate zones. As indicated in the figures, climatic conditions have a significant impact on building energy consumption. In this study, four cities with different climatic features were selected to demonstrate the impact of climate and building characteristics on energy consumption prediction models.

Bandar Abbas consumes the least heating and the most cooling energy among all studied cities because of its hot and humid climate. In contrast, as Tabriz is a cold and dry city, it consumes the most heating energy and the least cooling energy. In addition, all cities follow the same pattern of cooling and heating energy consumption over a year.

4. Results

4.1. Principal components identification

As indicated, the goal of the PCA model is to find a subset of reduced size with new variables in which the projected individuals retain their initial structures with the least possible distortion (Jiang et al., 2018). The PCA algorithm in this paper was used to identify the main components in the energy consumption of buildings based on available historical data. Fig. 12 demonstrates the loading plots for the Tehran data. As noted above, the loading plot shows the highly correlated parameters and the most influential factors for energy consumption. As the loading factors of a parameter increase, it shows that the parameter contributes more to principle components (PCs); as a result, this loading factors has a greater effect on the energy consumption and prediction model. Fig. 12(a) shows the load of all variables on two main principal components (PC1 and PC2). Fig. 12(b) shows PC3 versus PC1. It was assumed that these three principal components are the main components in energy consumption prediction. Although this assumption could reduce the accuracy of the modeling, the aim of this research was to demonstrate the effectiveness of the proposed methodology.

Fig. 13 shows that PC1, PC2 and PC3 cover 57% of the Tehran historical data. Although 60% is a low value to represent the behavior of the system, it is accurate enough to determine the most and least influential factors.

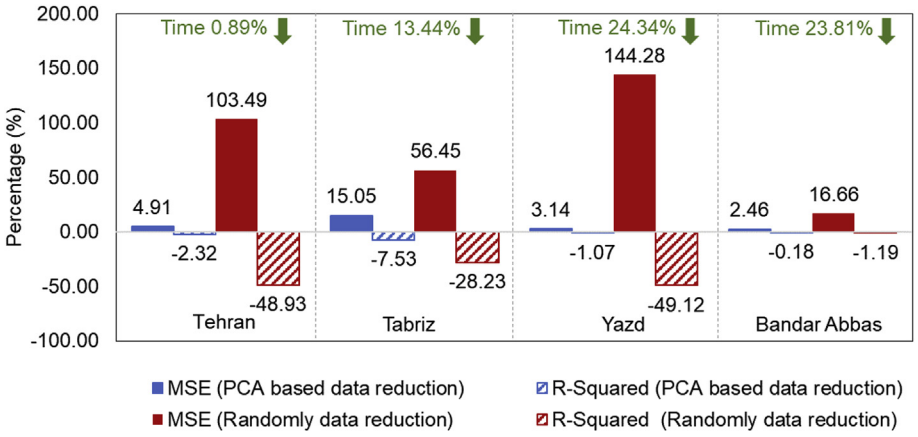


Fig. 19. Effectiveness of the data reduction methods as compared to the base case for the linear regression model.

The scatter score plot of PC-2 vs PC-1 for Tehran dataset is presented in Fig. 14. The scattered score plot could presents certain groupings in the data. For instance, in the presented figure, the dataset could be split into three or four distinct groups. This suggests that it is possible to classify the data based on these groups. This possibility of using PCA for classification forms the basis for a classification method called soft independent modelling of class analogies (SIMCA). In SIMCA method, for each data group (represented by different colors on the scattered score graph), a specific energy consumption prediction model could be developed that has a lower residual in comparison to a prediction model applied to the entire dataset.

Figs. 15–17 show the loading, cumulative explained variance and scatter score plots for Tabriz, Bandar Abbas and Yazd, respectively. The point labels in all loading plots are the same as in Fig. 12.

The least influential parameters of all climate zones were derived from the loading plots and are presented in Table 2. The first column shows the least influential parameters in the PC1 and PC2 loading plot; the second column shows the least influential factors in the PC1 and PC3 loading plot; and the last column shows the common parameters, presented in both column 1 and column 2. As these columns show the least influential factors of PC1, PC2 and PC3, these parameters can be eliminated in energy prediction models to improve the execution time.

4.2. Comparison of building energy consumption prediction models

Energy consumption was forecasted using five prediction models, namely linear regression, SVR, regression tree, random forest and KNN models. The observed data of each city is split into a training (80%) and a test (20%) datasets. The test dataset is used to provide an unbiased evaluation of the prediction models' performance. In order to study the effectiveness of the model in all weather conditions, 20% of the observed data for each city during each month is randomly selected to be used in the test dataset. The rest of the data of each month (80%), hence 80% of all observed data, are used as a training dataset.

Tables 3–6 present the comparison of these methods for the Tehran, Tabriz, Bandar Abbas and Yazd datasets. The comparison is based on two factors, the R^2 *and MSE*. R^2 is the proportion of the variance in the dependent variable that is predictable based on the independent variable, and it is a dimensionless quantity. In some cases, high R2 values could happen in time series datasets where dependent and independent variables both have trends over time. As our data are gathered over time with trends in some of the features (e.g. ambient temperature), it could result in high R^2 values in some of the prediction models. Therefore, MSE has been calculated as an additional index for quantifying the performance of the models. MSE is a measure of the average squared difference between the estimated energy consumption values and the actual value. The unit of energy consumption prediction in each hour is expressed in kW. Therefore, the MSE unit is presented in $(kW)^2$.

As can be seen, the prediction models perform differently depending on the climate zone. These differences show that the selection of a prediction model significantly depends on the historical data pattern. For instance, a highly efficient prediction model for a residential building cannot be generalized to other buildings, as the climate and other historical data are different. In most of the reviewed literature, a prediction model that fits well with a case study is generalized to a building type, such as commercial or residential. However, in addition to the building type, prediction model accuracy depends on the climate and the historical data of the building. As a result, different prediction models should be utilized to select the most efficient model for each case study. This approach significantly increases the execution time, as it requires running a different prediction model with a large historical dataset for every case under study. In this study, the PCA method was proposed to reduce the execution time. The related results are presented in the following sections.

Another point that could be seen in Table 6 is that random forest method fits better on the test set than the training set. In most cases, test sets have a higher error in comparison with training sets; however, it is totally possible that test sets have a higher R^2 value than training sets, as is presented in Table 6 for random forest method. This usually happens when a model is generalized well and/or when a training set is large, but the test set is small. The Yazd data set has a wide range in comparison to other cities that could be inferred from Figs. 8, 10 and 11. Therefore, the random forest model that is trained based on this dataset is well generalized, i.e., it has the ability to adapt properly to new, previously unseen data, drawn from the same distribution. As a result, the test data set fits better than the training set.

4.3. Performance validation of the proposed methodology

To evaluate the performance of the PCA method, we undertook a comparison between using and not using this method for data reduction. For this purpose, three datasets were selected:

1 Original building dataset (base case): Building raw data without any preprocessing were considered. These data cover all 20 building features and climate conditions presented in Fig. 1.

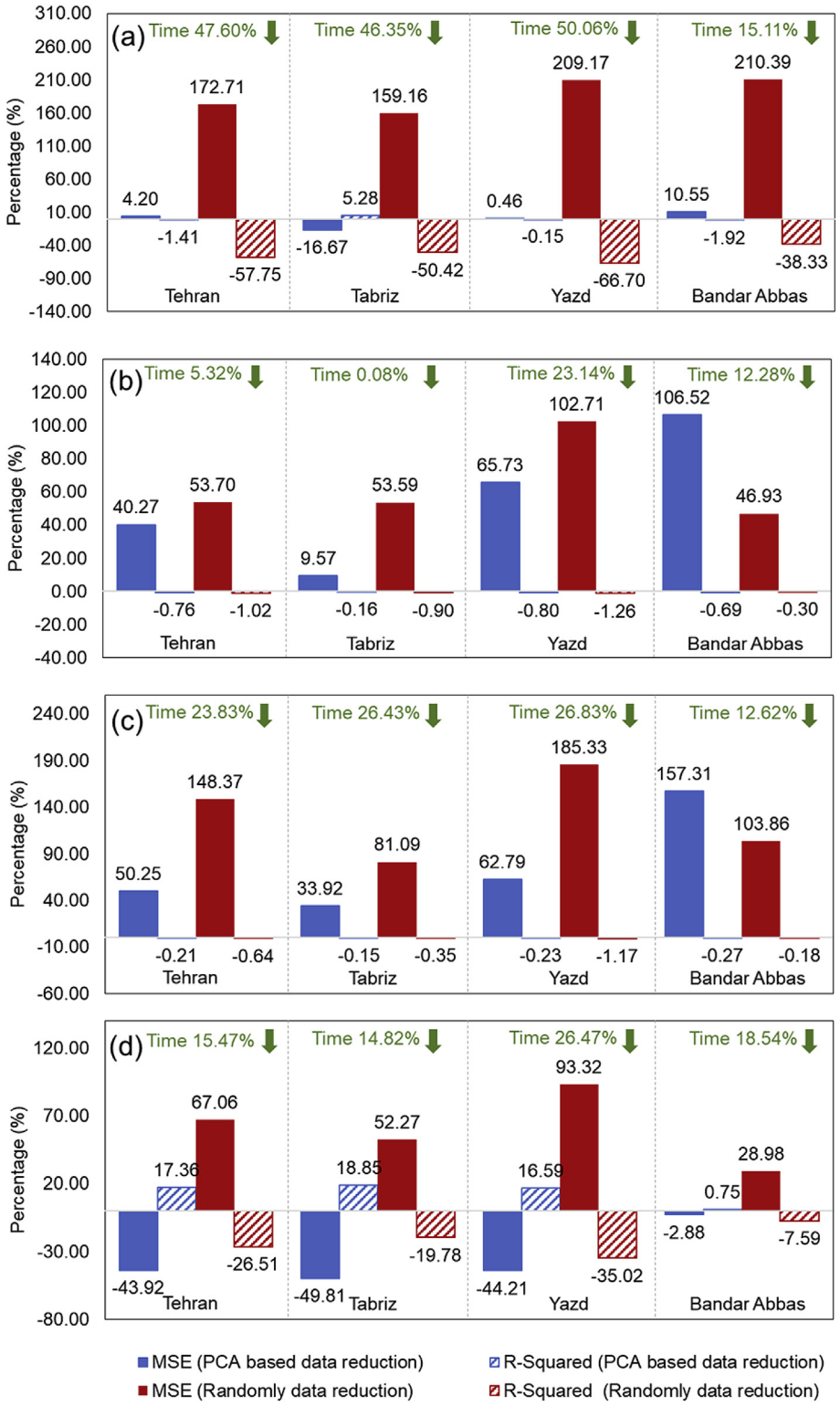


Fig. 20. Effectiveness of the data reduction methods as compared to the base case for the SVR (a), regression tree (b), random forest (c) and KNN (d) models.

2 PCA-based reduced dataset: The least influential factors were eliminated. This dataset includes all features except the common features presented in Tables 3–6.
3 Randomly reduced dataset: This dataset has the same number of features as the PCA-based reduced dataset, but the features were selected randomly. For instance, according to Table 2, seven features were eliminated for the PCA-based data reduction for the Tehran dataset. More specifically, seven features were randomly eliminated to generate the randomly reduced dataset for Tehran.

Fig. 18(a), (b) and (c) present the *MSE*, R^2 and execution time of

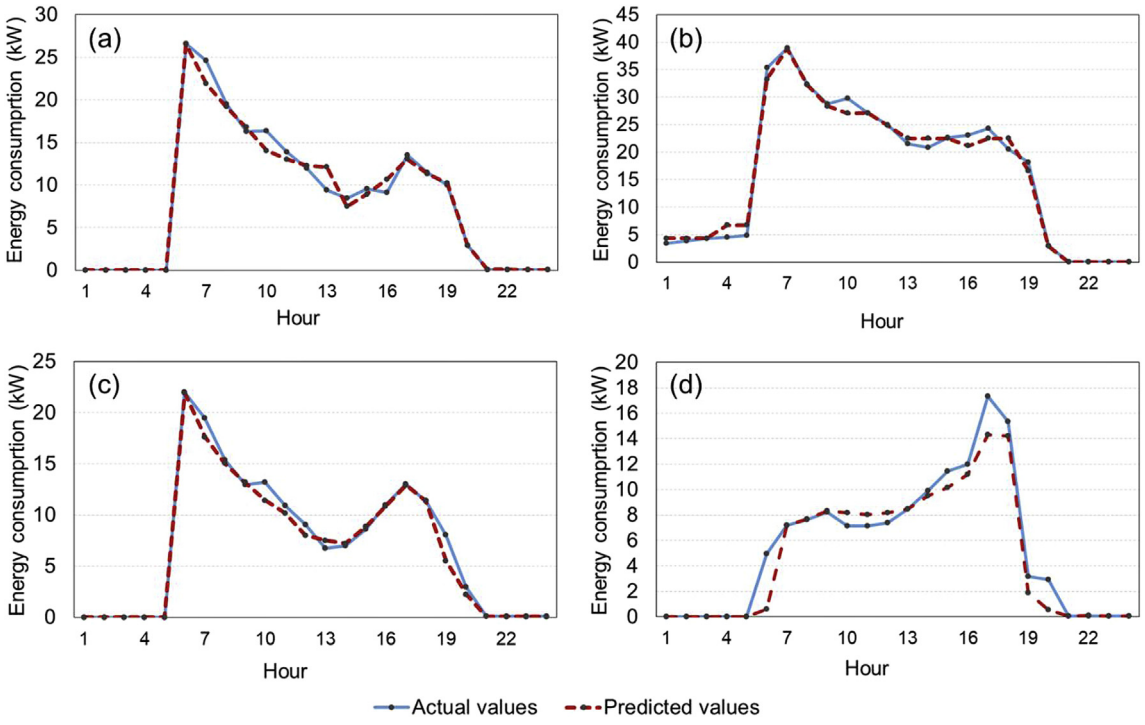


Fig. 21. One-day-ahead prediction of energy consumption in Tehran (a), Tabriz (b), Yazd (c) and Bandar Abbas (d) case studies.

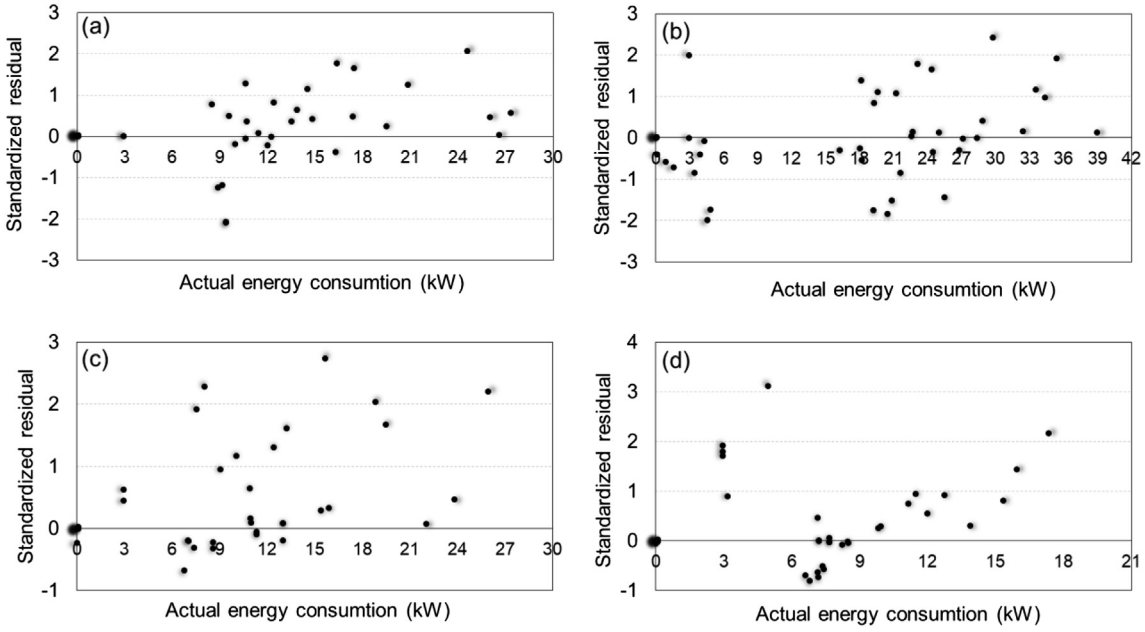


Fig. 22. Residual plot of energy consumption prediction models for Tehran (a), Tabriz (b), Yazd (c) and Bandar Abbas (d) case studies.

the linear regression model for the three datasets, respectively. As mentioned, the base case dataset involves 20 building features, including climate and building characteristics. The PCA-based reduced dataset considers all features except the common, less important features shown in Table 2.

The randomly reduced dataset contains the reduced data with exactly the same number of features as in the PCA-based dataset. In this dataset, the modeler selects the most important features based on an expert's judgment or randomly. There are multiple studies, such as (Li et al., 2009; Chou and Bui, 2014; Shao et al., 2020), that have selected limited number of features to predict/model energy consumption, and they have not performed any analysis to show how the features were selected and why these features are the most important ones. It is true that decreasing the number of features would result in better execution time; however, it decreases model accuracy as well.

The rationale behind using the randomly reduced dataset, is to emphasize that it is preferred not to remove features without

Table 7
The *MSE* and R^2 for the most efficient prediction method for all case studies.

Climate zone	Prediction model/metric	*MSE* $(kW)^2$	R^2	Time(s)
Tehran	Random forest	0.23	0.99	4.91
Tabriz	Regression tree	1.02	0.98	1.01
Yazd	Random forest	0.21	0.99	4.57
Bandar Abbas	Regression tree	1	0.99	0.89

performing any preprocessing data and feature reduction methods. The results of this section show the importance of dimensionality reduction methods.

The *MSE*, R^2 and execution time are the factors that illustrate the effectiveness of the data reduction process. Fig. 18(a) presents the execution time of the linear regression model for the three datasets and four climate zones. The reduction of the execution time is not the same for each climate zone. This finding demonstrates that, in addition to the prediction model, the data reduction method significantly depends on the data pattern. For instance, the execution time for all three Tehran datasets was almost identical, whereas for the Tabriz dataset the execution time was reduced by 13.44%, and for Yazd and Bandar Abbas, it decreased more significantly.

As indicated by Fig. 18(a), the data reduction methods reduced the execution time. However, data reduction also decreased the model performance. The performance of the prediction models is presented by the *MSE* and R^2. As shown in Fig. 18(b), the error increased significantly with the random data reduction method. However, this error slightly increased with the PCA-based data reduction method. The R^2 value followed the same pattern: its value was slightly lower for the PCA-based data reduction method and much higher for the random data reduction method. This result shows the effectiveness of data reduction based on the PCA method.

Fig. 19 summarizes the effectiveness of the PCA-based and random data reduction methods in comparison with the base case for a linear regression model. The blue columns represent PCA-based data reduction, and the red columns show the random data reduction results. As can be seen, the model performance was not the same for the different datasets (climate zones). Moreover, PCA-based data reduction had the strongest effect on the Yazd dataset when using a linear regression model, as it reduced the execution time by around 25%, with a 3% increase in the *MSE* and a 1% decrease in R^2. For the random data reduction method, the *MSE* was 144.28% higher and the R^2 was approximately 50% lower than with the base case.

Fig. 20 presents the same comparison of datasets discussed above for the other prediction models, namely the SVR, regression tree, random forest and KNN models. The results indicate that the effectiveness of data reduction was significantly dependent on the prediction method. According to Fig. 20(a), for the SVR model, PCA-based data reduction reduced the execution time by around 50% with a low effect on the *MSE* and R^2 values.

Fig. 20(b) shows that with the regression tree method, the execution time decreased slightly; however, the MSE increased dramatically. This result was due to the low value of the *MSE* with the regression tree method. As presented in Tables 3–6, the *MSE* for the regression tree and random forest models was significantly lower than for the linear regression, SVR and KNN models. As a result, a slight change in the *MSE* in the regression tree and random forest models could results in a high percentage value, as illustrated in Fig. 20(b) and (c). In contrast, the R^2 values did not change notably. Therefore, the results suggest that PCA-based data reduction could result in an accurate regression tree model. However, the execution time was only moderately reduced.

Fig. 20(c) shows the effectiveness of the PCA-based data reduction method for the random forest model. As can be seen, the *MSE* values increased significantly for the same reasons noted above for the regression tree model. However, for the random forest method, the execution time was reduced by around 25%, which is a remarkable value. Therefore, the PCA-based data reduction method could be an effective reduction method for regression tree models.

As illustrated in Fig. 20(d), the PCA-based data reduction method had a weak effect in terms of reducing the execution time for the KNN model. In addition, the *MSE* values increased, while the R^2 values decreased significantly. As a result, it can be inferred that PCA-based data reduction is not an effective solution for decreasing the execution time for KNN models. Overall, the results support the following points regarding PCA-based data reduction:

a) It reduced dataset has a shorter execution time compared to the original dataset due to the lower number of features.
b) It has a higher accuracy (lower MSE and higher R^2) compared to the datasets with the same number of features (and almost the same execution time).

4.4. Building energy consumption prediction

This section describes the prediction of the energy consumption of the case studies on a day-ahead basis. Data were reduced based on the PCA method, and the prediction model was selected according to the results in section 4-3. Fig. 21 presents the prediction of energy consumption for the last day of the year, represented with dashed red lines. The prediction is validated by results from the collected meteorological and energy data, represented by solid blue lines.

According to Fig. 21, the energy consumption patterns in Tehran and Yazd are almost the same. This result is due to the mild and moderate weather in both cities. Therefore, the most efficient prediction method for these two cases is the random forest method. The main difference between these two cities is the peak level of energy consumption that occurs in the early morning. The peak level is higher in Tehran due to the lower ambient temperature in this region.

However, in Tabriz and Bandar Abbas, the energy consumption patterns are completely different. Tabriz has a cold climate; therefore, energy consumption on the last day of December is higher in comparison to the other cities. This city has energy consumption during off time of the building, unlike other cities. This difference is mainly due to the low ambient temperature in this city, which requires that the heating system start working from midnight (hour 0 on the graph) to keep the building conditions in the acceptable temperature range during working hours. In contrast, Bandar Abbas has a hot climate, and the daytime ambient temperature is high. Therefore, energy consumption is low during the day, but increases during hours 16–18 when there is insufficient solar radiation and the ambient temperature is lower.

In addition, the energy consumption profiles in all the case studies follow the occupancy profile presented in Fig. 9(a). According to this occupancy profile, the office building is occupied from hours 7–18. Therefore, the electrical demands are highest during those hours. Furthermore, heating and ventilation systems should provide a comfort zone during working hours. To meet this criterion, heating and ventilation systems need to start working earlier, depending on the ambient conditions.

Fig. 22 presents the standardized residual plot of the energy consumption prediction models for all cities. The plot shows the standardized residual (test prediction error) as a function of energy consumption observation values. The standardized residual could

be calculated as Eq. (3).

$$Standardized\ residual = \frac{Residual\ (i)}{Standard\ deviation\ of\ residuals}$$

$$pt = \frac{Actual\ value\ (i) - Predicted\ value\ (i)}{\sqrt{\frac{(Residual\ (i) - Mean\ of\ residuals)^2}{Number\ of\ tests}}} \quad (3)$$

As can be seen, the pattern of energy consumption was predicted with high accuracy, and the values were predicted with acceptable accuracy. The MSE and R^2 metrics serve to evaluate the accuracy of the methods for all the case studies, as shown in Table 7. The most efficient prediction methods are the random forest and regression tree approaches, depending on the historical climate data. The reason is that these two prediction methods can perfectly predict datasets with a large number of features. In addition, the results imply that case studies with the same energy consumption patterns will have the same most efficient prediction methods.

5. Conclusion

In this paper, a hybrid PCA-based prediction method is proposed to predict building's energy consumption. As time passes, the historical meteorological and operational data of a building grow significantly. To develop an accurate energy consumption prediction model with a reasonable execution time, researchers should undertake data preprocessing. In this study, PCA is introduced as a data reduction method, and data preprocessing is performed for five prediction models including linear regression, SVR, regression tree, random forest and KNN. In addition, four types of datasets (four energy consumption patterns) are gathered to study the effect of the preprocessing method on the prediction models' performance. The results indicate that the PCA method can be a useful data reduction approach that significantly reduces the execution time of energy consumption prediction models.

In addition, according to the results, prediction model performance depends on the data pattern significantly, i.e., the best prediction model for cities with similar data patterns is the same. The PCA-based random forest model is proposed for Tehran and Yazd, and the best results are obtained from the regression tree for Tabriz and Bandar Abbas, with an MSE lower than 1 and an R^2 of around 0.99. The results verify that this approach could be applied to any other energy prediction models with large datasets, resulting in an accurate prediction with a significantly reduced execution time. This methodology could be beneficial in online performance monitoring, failure diagnosis and optimization systems that require highly efficient prediction models with low execution time.

CRediT authorship contribution statement

Tarannom Parhizkar: Conceptualization, Methodology, Writing - review & editing, Supervision. **Elham Rafieipour:** Data curation, Investigation, Modeling, Writing - original draft, preparation. **Aram Parhizkar:** Data curation, Investigation, Modeling.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Afroz, Z., Urmee, T., Shafiullah, G.M., Higgins, G., 2018. Real-time prediction model for indoor temperature in a commercial building. Appl. Energy 231, 29–53.

Alobaidi, M.H., Chebana, F., Meguid, M.A., 2018. Robust ensemble learning framework for day-ahead forecasting of household-based energy consumption. Appl. Energy 212, 997–1012.

Bagnasco, A., Fresi, F., Saviozzi, M., Silvestro, F., Vinci, A., 2015. Electrical consumption forecasting in hospital facilities: an application case. Energy Build. 103, 261–270.

Becerik-Gerber, B., Siddiqui, M.K., Brilakis, I., El-Anwar, O., El-Gohary, N., Mahfouz, T., Jog, G.M., Li, S.,., Kandil, A.A., 2013. Civil engineering grand challenges: opportunities for data sensing, information analysis, and knowledge discovery. J. Comput. Civ. Eng. 28 (4), 04014013.

Bui, D.K., Nguyen, T.N., Ngo, T.D., Nguyen-Xuan, H., 2020. An artificial neural network (ANN) expert system enhanced with the electromagnetism-based firefly algorithm (EFA) for predicting the energy consumption in buildings. Energy J. 190, 116370.

Candanedo, L.M., Feldheim, V., Deramaix, D., 2017. Data driven prediction models of energy use of appliances in a low-energy house. Energy Build. 140, 81–97.

Chen, Y., Tong, Z., Zheng, Y., Samuelson, H., Norford, L., 2020. Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings. J. Clean. Prod. 254, 119866.

Chou, J.S., Bui, D.K., 2014. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. Energy Build. 82, 437–446.

Duarte, C., Van Den Wymelenberg, K., Rieger, C., 2013. Revealing occupancy patterns in an office building through the use of occupancy sensor data. Energy Build. 67, 587–595.

Fan, C., Xiao, F., Wang, S., 2014. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl. Energy 127, 1–10.

Gan, V.J., Lo, I.M., Ma, J., Tse, K.T., Cheng, J.C., Chan, C.M., 2020. Simulation optimisation towards energy efficient green buildings: current status and future trends. J. Clean. Prod. 120012.

Gao, X., Malkawi, A.M., Yi, Y.K., 2013. A new method for predicting mixed-use building energy: the use of simulation to develop statistical models. In: Proceedings of the 13th Conference of International Building Performance Simulation Association, BS, pp. 2349–2356.

Gassar, A.A.A., Yun, G.Y., Kim, S., 2019. Data-driven approach to prediction of residential energy consumption at urban scales in London. Energy 115973.

Ghaderi, S.F., Azadeh, M.A., Omrani, H., 2006. August. An integrated DEA-COLS-PCA model for performance assessment and optimization of electricity distribution Units. In: Industrial Informatics, 2006 IEEE International Conference on. IEEE, pp. 236–241.

Guermoui, M., Melgani, F., Gairaa, K., Mekhalfi, M.L., 2020. A comprehensive review of hybrid models for solar radiation forecasting. J. Clean. Prod. 120357.

Güngör, O., Akşanlı, B., Aydoğan, R., 2019. Algorithm selection and combining multiple learners for residential energy prediction. Future Generat. Comput. Syst. 99, 391–400.

Guo, Y., Wang, J., Chen, H., Li, G., Liu, J., Xu, C., et al., 2018. Machine learning-based thermal response time ahead energy demand prediction for building heating systems. Appl. Energy 221, 16–27.

Harrell Jr., F.E., 2015. Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer.

Huang, Y., Shen, L., Liu, H., 2019. Grey relational analysis, principal component analysis and forecasting of carbon emissions based on long short-term memory in China. J. Clean. Prod. 209, 415–423.

Ilbeigi, M., Ghomeishi, M., Dehghanbanadaki, A., 2020. Prediction and optimization of energy consumption in an office building using artificial neural network and a genetic algorithm. Sustain. Cities Soc. 102325.

Iwafune, Y., Yagita, Y., Ikegami, T., Ogimoto, K., 2014. May. Short-term forecasting of residential building load for distributed energy management. In: 2014 IEEE International Energy Conference (ENERGYCON). IEEE, pp. 1197–1204.

Jiang, Q., Liu, Z., Liu, W., Li, T., Cong, W., Zhang, H., Shi, J., 2018. A principal component analysis based three-dimensional sustainability assessment model to evaluate corporate sustainable performance. J. Clean. Prod. 187, 625–637.

Li, D.H., Lam, J.C., 2001. Evaluation of lighting performance in office buildings with daylighting controls. Energy Build. 33 (8), 793–803.

Li, L.L., Liu, Y.W., Tseng, M.L., Lin, G.Q., Ali, M.H., 2020. Reducing environmental pollution and fuel consumption using optimization algorithm to develop combined cooling heating and power system operation strategies. J. Clean. Prod. 247, 119082.

Li, Q., Meng, Q., Cai, J., Yoshino, H., Mochida, A., 2009. Applying support vector machine to predict hourly cooling load in the building. Appl. Energy 86 (10), 2249–2256.

Liaw, A., Wiener, M., 2002. Classification and regression by random Forest. R. News 2 (3), 18–22.

Lin, G.Q., Li, L.L., Tseng, M.L., Liu, H.M., Yuan, D.D., Tan, R.R., 2020. An improved moth-flame optimization algorithm for support vector machine prediction of photovoltaic power generation. J. Clean. Prod. 119966.

Lin, W., Ren, H., Ma, Z., Yang, L., 2019. Using fuzzy clustering and weighted cumulative probability distribution techniques for optimal design of phase change material thermal energy storage. J. Clean. Prod. 233, 1259–1268.

Loh, W.Y., 2008. Classification and Regression Tree Methods. Encyclopedia of Statistics in Quality and Reliability, p. 1.

Luo, X.J., 2020. A novel clustering-enhanced adaptive artificial neural network model for predicting day-ahead building cooling demand. J. Build. Eng. 101504.

Ma, Z., Ye, C., Ma, W., 2019. Support vector regression for predicting building energy consumption in southern China. Energy Procedia 158, 3433–3438.

Moghadam, S.T., Delmastro, C., Corgnati, S.P., Lombardi, P., 2017. Urban energy planning procedure for sustainable development in the built environment: a review of available spatial approaches. J. Clean. Prod. 165, 811–827.
Parhizkar, T., Aramoun, F., Esbati, S., Saboohi, Y., 2019. Efficient performance monitoring of building central heating system using Bayesian Network method. J. Build. Eng. 26, 100835.
Parhizkar, T., Aramoun, F., Saboohi, Y., 2020. Efficient health monitoring of buildings using failure modes and effects analysis case study: air handling unit system. J. Build. Eng. 29, 101113.
Parhizkar, T., Hafeznezami, S., 2018. Degradation based operational optimization model to improve the productivity of energy systems, case study: solid oxide fuel cell stacks. Energy Convers. Manag. 158, 81–91.
Parhizkar, T., Mosleh, A., Roshandel, R., 2017. Aging based optimal scheduling framework for power plants using equivalent operating hour approach. Appl. Energy 205, 1345–1363.
Pham, A.D., Ngo, N.T., Truong, T.T.H., Huynh, N.T., Truong, N.S., 2020. Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. J. Clean. Prod. 121082.
Rahman, A., Srikumar, V., Smith, A.D., 2018. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Appl. Energy 212, 372–385.
Shao, M., Wang, X., Bu, Z., Chen, X., Wang, Y., 2020. Prediction of energy consumption in hotel buildings via support vector machines. Sustain. Cities Soc. 102128.
Skjærvold, N.K., Brovold, H., Rahmati, H., Martens, H., Tøndel, K., Cedersund, G., Munck, L.M., 2006. Multivariate Analyses and the Bridging of Biology's 'Math-Gap'. Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation, pp. 1–23.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.
Soucy, P., Mineau, G.W., 2001. November. A simple KNN algorithm for text categorization. In: Proceedings 2001 IEEE International Conference on Data Mining. IEEE, pp. 647–648.
Zhang, T., Liu, Y., Rao, Y., Li, X., Zhao, Q., 2020. Optimal design of building environment with hybrid genetic algorithm, artificial neural network, multivariate regression analysis and fuzzy logic controller. Build. Environ. 106810.
Zhou, L., Li, J., Li, F., Meng, Q., Li, J., Xu, X., 2016. Energy consumption model and energy efficiency of machine tools: a comprehensive literature review. J. Clean. Prod. 112, 3721–3734.