



# Data-Driven Approaches for Prediction and Classification of Building Energy Consumption

# 2

Yixuan Wei, Xingxing Zhang,  
and Yong Shi

## Abstract

A recent surge of interest in the building energy consumption has generated a tremendous amount of energy data, which boosts the data-driven algorithms for broad application throughout industry. This chapter reviews the prevailing data-driven approaches used in building energy analysis under different archetypes and granularities including those for prediction (artificial neural networks, support vector machines, statistical regression, decision tree and genetic algorithm) and those for classification (K-mean clustering, self-organizing map and hierarchy clustering). To be specific, we introduce the fundamental concepts and major technical features of each approach, together summarizing its current R&D status and practical applications while

pointing out existing challenges in their development for prediction and classification of building energy consumption. The review results demonstrate that the data-driven approaches, although they are constructed based on less physical information, have well addressed a large variety of building energy related applications, such as load forecasting and prediction, energy pattern profiling, regional energy-consumption mapping, benchmarking for building stocks, global retrofit strategies and guideline making etc. Significantly, this review refines a few key tasks for modification of the data-driven approaches in the contexts of application to building energy analysis. The conclusions drawn in this review could facilitate future micro-scale changes of energy use for a particular dwelling through appropriate retrofit in building envelop and inclusion of renewable energy technologies. They also pave an avenue to explore potential in macro-scale energy-reduction with consideration of customer demands. All these will be useful to establish a better long-term strategy for urban sustainability.

---

Y. Wei (✉)  
School of Civil Engineering and Resources,  
University of Science & Technology Beijing,  
Beijing, China  
e-mail: [weiyixuan@ustb.edu.cn](mailto:weiyixuan@ustb.edu.cn)

X. Zhang  
Department of Energy and Community Buildings,  
Dalarna University, 79188 Falun, Sweden  
e-mail: [xza@du.se](mailto:xza@du.se)

Y. Shi  
Research Centre for Fluids and Thermal  
Engineering, University of Nottingham Ningbo  
China, Ningbo, China  
e-mail: [yong.shi@nottingham.edu.cn](mailto:yong.shi@nottingham.edu.cn)

## Keywords

Data driven approach · Building · Energy  
consumption · Prediction · Classification

## 2.1 Introduction

### 2.1.1 The Need for Energy Consumption Analysis

The global contribution from buildings towards energy consumption has steadily increased reaching figures between 20 and 40% in developed countries and about 1/3 of greenhouse gas emission. The case of China is particularly striking-the country only takes two decades to double its building energy consumption at an average growing rate of 3.7% (Pérez-Lombard et al. 2008; UNEP 2013). These facts demonstrate that to facilitate energy efficiency of building is a cost-effective resource for reducing energy consumption and carbon emission from building (Mathew et al. 2015). Also, large potential saving in economy has been anticipated by a large variety of previous studies. For instance, Nikolaidis et al. have shown that among various energy saving measures for common building types, isolation of roof constitutes the most superiority nearly €5000 economic benefit during 30 years (Nikolaidis et al. 2009). As the central approaches transmitting to energy efficiency, prediction and classification of energy consumption in building are significantly necessary with the aim to improve building performance, reduce environmental impact, and estimate economical potential for further energy conservation and renewable energy program (Zhao and Magoulès 2012).

Energy consumption in building has been heavily analyzed by substantial studies during the entire building lifecycle, with different focuses on identifying the sub-component energy use at the building level (Kang and Jin 2014; Bojić and Lukić 2000) or measuring energy performance in a nationwide analysis (Farahbakhsh et al. 1998; Huang 2000; Shimoda et al. 2004). This comprehensive set of analyses on different levels could help us not only optimize the energy use of a particular dwelling through appropriate retrofit in building envelop or inclusion of state-of-the-art renewable energy technologies (at the microscale), but also explore possible energy

reduction opportunities and establish better urban-sustainability strategies (at the macroscale).

### 2.1.2 Advantage and Motivation

However, it is recognized that realization of a precise energy consumption analysis is a formidable task at the current stage. As an alternative, great efforts have been paid to developing models to predict and classify approximately energy consumption. Generally, these models possess some prominent functions. They includes (1) taking measures for energy conservation based on accurate prediction, (2) implementing demand-side management (DSM) after profiling electricity consumption, (3) outlining/mapping energy on the urban level, (4) establishing benchmark database of multi-scale building communities, and (5) integrating the processes of design, operation, retrofit of contemporary building (Perino et al. 2015; Hong et al. 2014). The results simulated by these models can not only offer essential information about energy footprint in regional building stocks, but also facilitate estimations of financial return on investment. It is thus not surprising energy simulation has become a favourable tool for stakeholders throughout building industry including policymakers, building owners, investors, operators and engineers (Mathew et al. 2015).

### 2.1.3 Usage of Building Energy and Performance Data

Management and optimization of building energy consumption call for a full understanding of building performance, which should first identify energy resources and major end-uses of a building. Energy resources in a building usually refer to electricity, natural gas and district heating supply. The corresponding major end-uses include heating, ventilation and air-conditioning (HVAC) system, domestic hot water, lighting, plug-loads, elevators, kitchen

equipment, ancillary equipment and appliances. Figure 2.1 illustrates a representative classification of building energy use adopted in ISO Standard 12655:2013 (ISO 2013). Note that on top of the above building energy resources and major end-uses, HVAC operation schedule and indoor/outdoor conditions are also two important contributing factors to be considered in a building performance analysis.

Generally, reliability of a building performance analysis relies heavily on the datasets in use, which should contain sufficient energy consumption information of the buildings under investigation. Utility bills for electricity and natural gas from power supply companies are the common type of databases of building energy consumption. Facility managers or research institutes also collect information via survey and questionnaire for large-scale buildings, such as

the residential sector (Residential Energy Consumption Survey (RECS), EIA 2009) and commercial buildings (Commercial Building Energy Consumption Survey (CBECS), EIA 2012) (Hong et al. 2014). In addition, in today’s building performance analyses, virtual building database (VBD) developed from simulation software (e.g. TRNSYS and EnergyPlus) and energy disclosure laws (Mathew et al. 2015; Nikolaou et al. 2012) (e.g., US Energy Information Administration database) are the other two possible data resources. It is particularly worth mentioning that the empirical datasets taking advantage of smart meters and building energy system have emerged in recent years. These databases substantially improved accuracy and reliability of the related analyses (Mathew et al. 2015) despite their expensive costs and technical complexity involved for many practical commercial-uses.

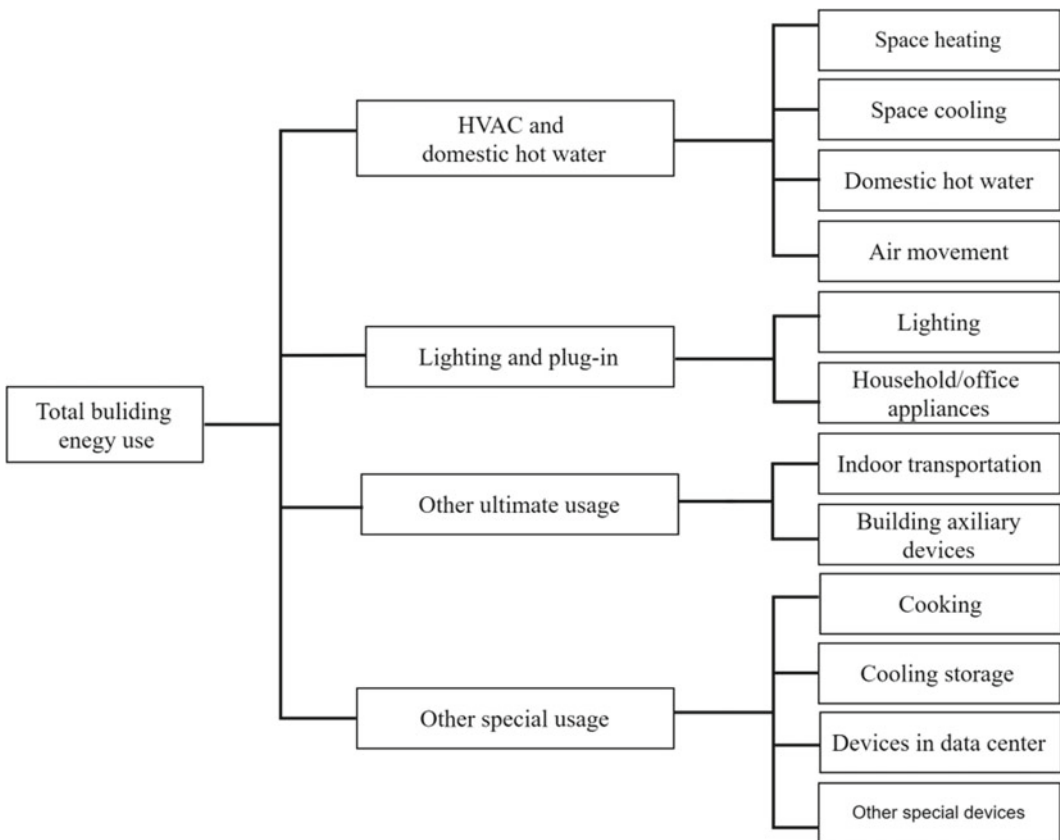


Fig. 2.1 The usage of energy in buildings (ISO 2013)

### 2.1.4 Proposed Methodologies for Building Energy Consumption

It is a challenging task to precisely describe energy consumption in a building as such an energy performance depends on a wide range of factors, such as weather condition, thermal properties of building envelope, occupancy behaviour, sub-level components' (lighting, HVAC and plug equipment) performance and schedules (Zhao and Magoulès 2012). A large number of efforts have been paid in the literature to ascertain the complexity pertinent to building energy consumption and strive to a precise depiction of building energy performance. Currently, these approaches used for building energy simulation are categorized roughly as: (1) white-box based approaches, (2) grey-box based approaches and (3) black-box based approaches, whose main features are summarized in Table 2.1.

White-box based approaches are physical-based approaches, which require detailed information of complex building phenomena. This basic characteristic determines their simulations will be rather computationally expensive. Recently, a series of attempts have made to simplify the white-box based approaches. However, these simplifications are error-prone and usually overestimate energy-saving of buildings

(Al-Homoud 2001; Barnaby and Spittler 2005). Grey-box based approaches are a modification of these white-box based approaches through use of statistical methods combining the simplified physical information with historical data to simulate building energy. One primary issue in current grey-box version is computational inefficiency as the approaches involve uncertain inputs and complex interactions among elements and stochastic occupant behaviours (Paudel et al. 2015; Li et al. 2014). To circumvent the above shortfalls of white- and grey box based approaches, black-box based approaches are developed which are able to conduct a building energy consumption analysis only based on historical data without the detailed knowledge of on-site physical information. This essential change enable black-box based approaches fast calculations in high accuracy in comparison to their white- and grey-box counterparts (Zhao and Magoulès 2012). In many practical scenarios, the black-box based approaches are also called as data-driven approaches due to the statistical algorithm structures and a large amount of data in use. We will follow this convention and use the data-driven approaches throughout the following discussion in this review.

The remainder in this review is organized as follows: we will introduce various mainstream data-driven approaches and summarize their

**Table 2.1** Comparison among white-box, grey-box and black-box approaches for building energy consumption

Approaches building simulation		Inputs needed	Typical software methods	Easy to use	Running speed	Accuracy
White-box based	Elaborated simulation	Detailed physical information	DOE-2, EnergyPlus, TRYSYS, ESP-r	No	Low	High
	Simplified simulation		Degree day method, temperature frequency method, residential load factor method	Yes	High	Fairly high
Data-driven models	Grey-box based	Physical information and historical data	RC network	No	Low	Fairly high
	Black-box based	Historical data	ANNs, SVMs, statistical regression, GA, cluster algorithms	No	High except SVM	High except regression

applications in prediction and classification of building energy consumption in Sects. 11.2.2 and 11.2.3, respectively. In Sect. 11.2.4, a few promising future directions in data-driven approaches with applications to building energy will be proposed. Finally, we draw our salient conclusions in Sect. 11.2.5.

## 2.1.5 Data-Driven Approaches

Data-driven models are constructed based on a group of datasets consisting of historical data records. These historical data will be used as benchmarks to justify the model's performance and guide its algorithm design. To be specific, all the parameters in a data-driven model will be carefully selected and modified through systematic comparisons between the model outputs and the historical data. This is the so-called learning process and only when the output errors fall within the required threshold, the corresponding data-driven models are deemed to be qualified for practical applications with fresh input data. Currently, the data-driven models is very prevailing in medical diagnosis (Kuo et al. 2001), political campaigns (Sides 2014) and commerce (Alhamazani et al. 2015) because of their low costs with no need of expensive equipment and audit activity. As to the building energy consumption studies, data-driven models are widely applied to either estimate the building energy demands (i.e., data-driven prediction models) or profile the energy consumption patterns (i.e., data-driven classification models), which are grouped in Fig. 2.2.

## 2.1.6 Data-Driven Prediction Models

Among the most popular data-driven prediction models are artificial neural networks (ANNs), support vector machine (SVM), statistical regression, decision tree (DT) and genetic algorithm (GA). This subsection will introduce each of these models.

### 2.1.6.1 Artificial Neural Networks

ANNs are designed mimicking the basic architecture of human brain, whose basic element is called as processing unit modelling a biological neuron. The network consists of a large number of these process units arrayed in layers, and process units in different layers are connected with one another via connections, shown in Fig. 2.3.

Each process unit, say  $l$ , will deal with signals,  $x_{il}$  ( $i = 1, 2, \dots, m$ ), from units connected with it in the other layers. These signals are input through the incoming connections with a weight  $w_{il}$  ( $i = 1, 2, \dots, m$ ). The process unit then takes two basic operations on the input signals: summation and activation, and delivers an output  $y_l$  (Magoules and Zhao 2016).

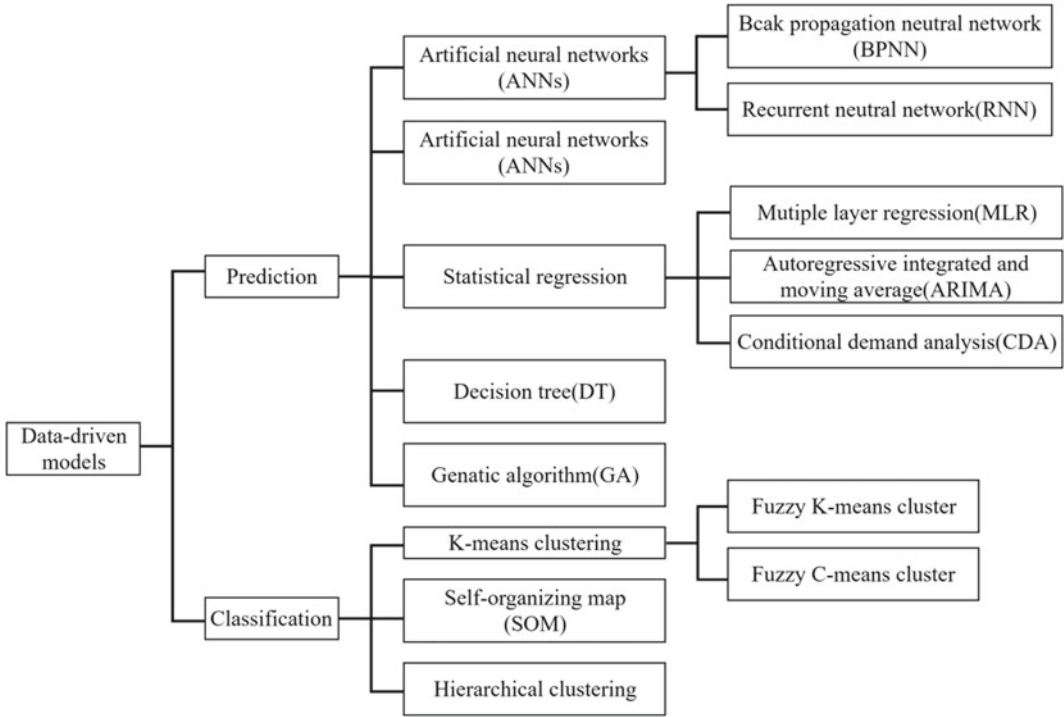
$$y_l = f\left(\sum_{i=1}^m w_{il}x_i + b_l\right) \quad (2.1)$$

where  $b_l$  is a bias set specifically for each process unit and  $f$  is the activation function, commonly defined as the sigmoid function (Magoules and Zhao 2016).

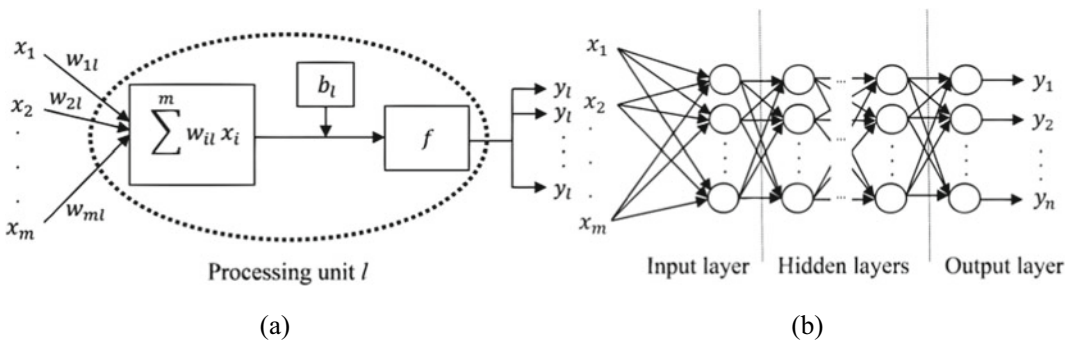
$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.2)$$

The output  $y_l$  will be used as an input signal for the process units in the next layer connecting to the process unit  $l$ .

As we discussed, all the process units in ANNs are arranged in a layer-structure and process units in different layers are interconnected based on a designed architecture. Figure 2.3b shows a simple example: feed-forward ANNs where process units are arrayed in the input, hidden and output layers and the information flows in one direction throughout these layers. In today's ANNs studies, ANNs models also take other architectures to more effectively approximate human brain activities. Two representative are back-propagation neural network (BPNN) and recurrent neural network (RNN), see Fig. 2.4. The former computes the error of output



**Fig. 2.2** Different data-driven models for building energy consumption



**Fig. 2.3** Schematic of ANN. **a** A single process unit; **b** Artificial neural networks

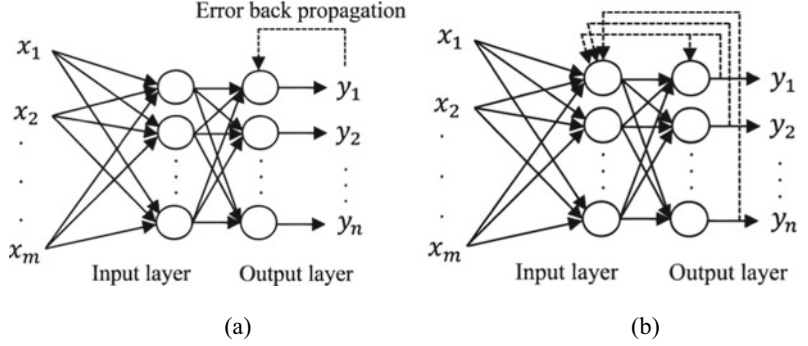
every time, and then propagates this information as a negative feedback to tune the incoming connection weight and bias. This manipulation offers flexibility to modify the output error to a minimum, and thus improving accuracy of ANN calculation. As to RNN, it involves the backward connections feeding back the outputs themselves as the inputs to the process units in the former-layer or even the current unit to capture tempore

behaviours. Such a recurrent design makes RNN deal with time series datasets without random data, which leads it to being particularly welcome for sequence events (Kalogirou and Bojic 2000).

No matter what kind of network architecture is in use, an ANNs model must experience a training (learning) process to specify all needed connection weights and biases before real



**Fig. 2.4** Schematic of **a** two-layer BPNN and **b** two-layer RNN. White circles: process units in different layers. Solid arrows: connections; dashed arrows: feedbacks



applications. This training process will take advantage of available historical data records, which will be used as benchmarks to cultivate the proper response of the ANNs model for given inputs. Therefore, ANNs are capable of learning the relationship among input signals, and capturing key information through a training process based on historical data records. On top of that, it also possesses a number of other advantages, such as fault tolerance, robustness and noise immunity. Thanks to these favourable features, ANNs have achieved great success in solving non-linear problems so far. On the other hand, meanwhile, it should be also pointed out that the architecture choice and learning-rate optimization in the current ANNs are still developed on an ad hoc base. This implies ANNs applications are usually case-dependent nonetheless. They have to be designed and validated for every and each time for different applications (Kalogirou 2001).

### 2.1.6.2 Support Vector Machine

Supported vector machine (SVM) is another popular artificial intelligent method (Vapnik et al. 1996), which deals with  $n$  data records, i.e.,  $\{(x_i, Y_i)\}_{i=1}^n$ , with the input  $x_i \in \mathbb{R}^N$  and the target  $Y_i \in \mathbb{R}$ . (Note that  $Y_i$  could also be in binary for some applications (Zhao and Magoulès 2010)). Nowadays, this method has been widely applied to solve regression problems to estimate an underlying relationship between the nonlinear inputs to the continuous real-valued target. The SVM used for regression is called as support vector regression (SVR), which has

become a particularly important data-driven approach for predicting building energy consumption.

The core task in SVR is to construct a decision function,  $F(x_i)$ , by use of a training process based on historical data. It is required that for a given input  $x_i$ , the result estimated by this function should not deviate from the actual target  $Y_i$  larger than the predefined threshold. In SVR, such a function is usually assumed in the form of

$$F(x_i) = \langle w, \varphi(x_i) \rangle + b \quad (2.3)$$

where the bias  $b \in \mathbb{R}$ .  $\langle \cdot, \cdot \rangle$  and  $w$  represent the dot product and weight defined in  $\mathbb{R}^N$ .  $\varphi(x_i)$  is a non-linear mapping of the input space to a high-dimensional feature space (Dong et al. 2005).  $w$  and  $b$  are two unknown in Eq. (2.3), and need to be estimated through minimizing the regularized risk function (Dong et al. 2005). In SVM theory, the latter is easily solved in its dual formulation by an introduction of a Lagrangian  $L$  (Magoules and Zhao 2016),

$$\begin{aligned} L := & \frac{1}{2} \|w\|^2 + c \sum_{i=1}^n (\xi_i + \xi_i^*) - \sum_{i=1}^n (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^n a_i (\varepsilon + \xi_i - y_i - \langle w, \varphi(x_i) \rangle - b) \\ & - \sum_{i=1}^n a_i^* (\varepsilon + \xi_i^* - y_i - \langle w, \varphi(x_i) \rangle - b) \end{aligned} \quad (2.4)$$

where  $\{a_i, a_i^*, \eta_i, \eta_i^* \geq 0\}$  are the Lagrange multiplier.  $\|w\|$  is the Euclidean norm.

$\{\xi_i, \xi_i^* \geq 0\}$  are two slack variables to copy with some infeasible optimization constraints. The constant  $c > 0$  is defined to determine the trade-off between the training error (over-fitting) and model flatness (under-fitting). It should be noted that the Lagrange multipliers are all independent. They are  $\eta_i = c - a_i$  and  $\eta_i^* = c - a_i^*$ , and  $\{a_i, a_i^*\}$  can be determined by the corresponding dual optimization (Dong et al. 2005),

$$\begin{aligned} \text{Maximize } W(a_i, a_i^*) &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (a_i - a_i^*) (a_j - a_j^*) (\varphi(x_i) \cdot \varphi(x_j)) \\ &\quad + \sum_{j=1}^n (a_i - a_i^*) y_j - \varepsilon \sum_{j=1}^n (a_i + a_i^*) \\ \text{subject to } &\begin{cases} \sum_{j=1}^n (a_i - a_i^*) = 0 \\ a_i, a_i^* \in [0, c] \end{cases} \end{aligned} \quad (2.5)$$

With the computed  $a_i, a_i^*$ , the weight  $w$  can be written a function of  $\{a_i, a_i^*, x_i\}_{i=1}^n$ . This gives rise to the decision function in SVR

$$F(x) = \sum_{x_i \in SV} (a_i - a_i^*) K(x, x_i) + b \quad (2.6)$$

where  $K(x, x_i) = \varphi(x) \cdot \varphi(x_i)$ . In SVR, this is called as the kernel function, having different formulas for various applications in the literature, e.g.,  $K(x, x_i) = \exp(-\gamma \|x - x_i^2\|)$ . It should be pointed out the sum in Eq. (2.6) does not cover all inputs. Instead, only those (i.e., support vectors  $x_i \in SV$ ) corresponding to  $(a_i - a_i^*) \neq 0$  are included. Moreover, the bias  $b$  in Eq. (2.6) is also computed by these support vectors

$$\begin{aligned} b &= \frac{1}{N_1} \left\{ \sum_{a_i \in (0, c)} \left[ Y_i - \sum_{x_j \in SV} (a_j - a_j^*) E(x_i, x_j) - \varepsilon \right] \right. \\ &\quad \left. + \sum_{a_i^* \in (0, c)} \left[ Y_i - \sum_{x_j \in SV} (a_j - a_j^*) K(x_i, x_j) + \varepsilon \right] \right\} \end{aligned} \quad (2.7)$$

Here,  $N_1$  is the number of support vectors with either  $\{a_i \in (0, c), a_i^* = 0\}$  or  $\{a_i = 0, a_i^* \in (0, c)\}$ . Once the decision function, i.e.,

Eq. (2.6), is fully specified by the training dataset, the SVR model can be used as a predicting tool for a new input  $x$ .

It is worth emphasizing that the superiority of SVR, or more generally SVM, to other models are that its framework is easily generalized for different problems and it can obtain globally optimal solutions. Its capability of dealing with nonlinear relations by transferring them into high-dimensional linear problem is also impressive for practical applications. Nonetheless, the method is rather time-consuming for large-scale problems (Zhao and Magoulès 2010; Li et al. 2009a). Recently, immerse efforts has been paid to developing possible ways to optimize its computational efficiency.

### 2.1.6.3 Statistical Regression

Prediction of building energy-consumption relies on a regression analysis to devise a relationship linking an output (i.e. response,  $Y_i, i = 1, 2, \dots, n$ ) to the contributing inputs (i.e., predictors,  $x_{i,j}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$ ). In the previous section, we have discussed a regression process based on the SVM theory-SVR. On top of that, there still exist other regression models, e.g., statistical regression, used for predicting building energy consumption. Statistical regression investigates the relationship among different variables in a probabilistic framework, which formulate the output as

$$\text{Multiple : } Y_i = \alpha_i + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_m x_{i,m} + \varepsilon_i \quad (2.8)$$

or

$$\text{Polynomial : } Y_i = \tilde{\alpha}_i + \tilde{\beta}_1 x_{i,1} + \tilde{\beta}_2 x_{i,2}^2 + \dots + \tilde{\beta}_m x_{i,m}^m + \varepsilon_i \quad (2.9)$$

where  $\varepsilon_i$  represents a random error assumed to be normally distributed, and  $\alpha_i, \tilde{\alpha}_i, \beta_j$  and  $\tilde{\beta}_j$  ( $j = 1, \dots, m$ ) are the parameters to be estimated. Note that both Eqs. (2.6) and (2.7) are linear with respect to these parameters whilst they are not necessarily linear with respect to the contributing predictors, as seen as Eq. (2.7). Like other data-driven approach for prediction, the



statistical regression equations make use of the finite number of historical data to estimate the involved parameters. For demonstration, we choose the multiple linear regression Eq. (2.6) as an example, in which the estimates of all parameters will be derived using the least squares (LS). To be specific, the sum of squared errors (SSE) is first defined

$$SSE = \sum_{i=1}^n (y_i - A_i - B_1 x_{i,1} - B_2 x_{i,2} - \dots - B_m x_{i,m})^2 \quad (2.10)$$

In Eq. (2.8)  $A_i$  and  $B_j$  ( $j = 1, \dots, m$ ) are the corresponding LS estimates of  $\alpha_i$ ,  $\beta_j$  ( $j = 1, \dots, m$ ) in Eq. (2.6). SSE is then minimized which gives rise to  $m + 1$  equations. Each of these equations includes one of partial derivatives of SSE with respect to  $A_i$  and  $B_j$  ( $j = 1, \dots, m$ ), to be set zero, respectively. It is these equations that are used to solve  $A_i$  and  $B_j$  ( $j = 1, \dots, m$ ) directly subject to the given historical dataset  $\{x_{i,j}, Y_i, i = 1, 2, \dots, n, j = 1, 2, \dots, m\}$ . Finally, the prediction equation with the estimated parameters in multiple linear regression is specified as

$$y_i = A_i + B_1 x_{i,1} + B_2 x_{i,2} + \dots + B_m x_{i,m} \quad (2.11)$$

In statistical regression, there is another variable introduced to quantify the goodness of fit of the regression line by Eq. (2.9), that is the coefficient of determination  $R^2$ ,

$$R^2 = 1 - \frac{SSE}{SS_{tot}} \quad (2.12)$$

where  $SS_{tot} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ , with the mean value  $\bar{Y} = \sum_{i=1}^n Y_i$ . Generally, a regress equation with a larger  $R^2$  indicates it can better fit the original data.

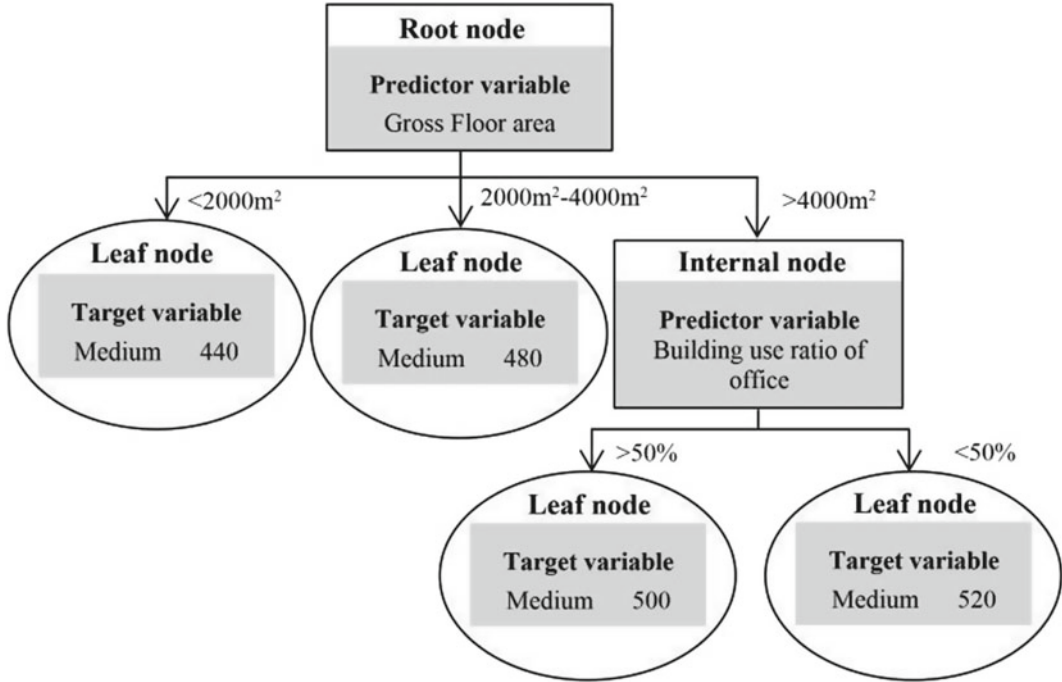
Based on the above discussion, it is seen that statistical regression is an easy-to-use approach

for predicting building energy consumption. In particular, it was popular to predict average consumption over a long period in the early studies. However, the regress models require a large number of historical data for training, and the resulting accuracy of a short-term prediction is yet poorer than that of other data-driven approaches, such as ANN or SVM. It is also challenging for statistical regression to select a set of plausible predictors and an appropriate time scale to well fit energy consumption for buildings under a wide range of environment and weather conditions. Worse, the selected predictors in some cases may not be literally independent. The unforeseen correlations among them would result in uncertain inaccuracy in the regression outputs (Swan and Ugursal 2009).

#### 2.1.6.4 Decision Tree

Decision tree (DT) is a technique to partition data into groups using a tree-like flowchart. In this sense, a DT model manifest itself as a graph consisting of a root node and a couple of branch nodes. A DT starts from the root node where the input data are split into different groups based on some predictor variables predefined as splitting criteria. These split data are then disseminated to sub-nodes as branches emanating from the root node. The data on sub-nodes will undergo either further or no splits. The former are the internal nodes where the subsequent data split is conducted to form new subgroups as son-branches emanated graphically at the next level. Whereas the latter are leaf nodes which treat the corresponding data group at the current level as their final outputs. Figure 2.5 illustrates a DT representation used for medium annual source energy consumption per unit floor ( $\text{kWh}/\text{m}^2/\text{yr}$ ) of a commercial building. In this case, the gross floor area and building use ratio are chosen as predictor variables in the root node and internal node, respectively, and a mixture of data about energy consumption has been purified into a hierarchy of groups.

Significantly, in a DT analysis the information entropy is an important concept used to quantify data group homogeneity. It is defined by



**Fig. 2.5** Decision tree illustration of a medium annual source energy consumption per unit floor of a commercial building

$$E = \sum_{i=1}^n -P_i \log_2 P_i \quad (2.13)$$

where  $E$  is the information entropy.  $n$  and  $P_i$  are the number of different target values and the probability of a dataset taking the  $i$ th target value, respectively. This entropy is used to calculate the information gain or gain ratio, based on which a DT structure linking the top root node to each branch node is specified. Readers can refer to Quinlan (1986) for detailed splitting procedure using the gain ratio or information gain.

In comparison to other data-driven approaches, DT's tree-like structure is easy to understand and its implementation does not involve complex computation knowledge. However, its deficiency is also evident—the targets used in a DT are primarily based on expectations. This usually leads to significant deviations of its predictions from the real results. The DT

architecture is also a restriction so as the method is unable to deal well with time-series and non-linear data.

### 2.1.6.5 Genetic Algorithms

Genetic algorithms (GAs) are stochastic optimization inspired by natural evolution based on the idea of “survival of the fittest” (Goldberg 1986). Many GAs in building energy prediction formulate three kinds of algebraic equations to compute the output (as solution) according to the given inputs:

$$\text{Linear : } y = w_1 x_1 + \cdots + w_m x_m, \quad (2.14)$$

$$\begin{aligned} \text{Quadratic : } y = & w_1^{(1)} x_1 + \cdots + w_m^{(1)} x_m + w_{1,2} x_1 x_2 \\ & + \cdots + w_{1,m} x_1 x_m + w_{2,3} x_2 x_3 + \cdots \\ & + w_{m-1,m} x_{m-1} x_m + w_1^{(2)} x_1^2 + \cdots + w_m^{(2)} x_m^2, \end{aligned} \quad (2.15)$$

$$\begin{aligned} \text{Exponential : } y = w_0 + w_1 x_1^{\tilde{w}_1} \\ + w_2 x_2^{\tilde{w}_2} + \dots + w_m x_m^{\tilde{w}_m}, \end{aligned} \quad (2.16)$$

where  $(x_1, x_2, \dots, x_m)$  are  $m$  independent inputs contributing to the output,  $y$ , and  $w_i, w_i^{(1,2)}$  and  $\tilde{w}_i$  are the real-valued weights. In GAs, different sets of weights compose a search space where a point represents a feasible solution to the problem under investigation. The core task of a GA is to model an evolution process to identify the best among all feasible solutions in this space. In implementation, a GA first randomly chooses  $n$  sets of weights and encode each weight as a 1 bit

binary string, e.g.  $w_i = \overbrace{100\dots01}^1$ . In so doing, a set of weights is then represented as a chromosome  $X_j = \overbrace{100\dots01}^{w_1} \overbrace{000\dots11}^{w_2} \dots \overbrace{100\dots10}^{w_m}$ , and the  $n$  chromosomes form an initial population  $r$ . Importantly, every chromosome  $X_j$  in the population  $r$  is mapped to a fitness  $h(X_j)$  (a real value) and assigned a probability  $P_j$ . In most cases, these two variables are defined by

$$h(X_j) = \left( y(x_1, x_2, \dots, x_m | X_j) - Y \right) \quad (2.17)$$

and

$$P_j = \frac{h(X_j)}{\sum_r h(X_k)} \quad (2.18)$$

where  $Y$  is the targeted output from historical datasets and the Greek letter “ $\Sigma$ ” denotes a sum of the fitness of all chromosomes in the population  $r$ . Next, pairs of chromosomes are selected as parents to reproduce the offspring (still chromosomes). Generally, the better fitness the chromosomes have, the more possible they are selected. The chosen parents then proceed crossover and mutation. One simple crossover operation is to randomly choose a crossover point and exchange the alleles up to this point of the two parent chromosomes. As to mutation, a few of bits in the chromosome after crossover,

again chosen randomly, are switched between 0 and 1 (e.g.  $1000\underline{1} \rightarrow 100\underline{1}1$ ). Selection, crossover and mutation will be repeated to generate sufficient new offspring to form a new population,  $r'$ , at the next level. It should be pointed out that the fitness of all offspring chromosomes in this new generated population will be computed and compared with the user's requirements. Generally, a GA will continue further runs of the above evolution process unless a chromosome (i.e., a set of weights) with satisfactory fitness is reproduced.

The aforementioned introduction of GAs indicates this method is a powerful optimization tool in dealing with complex multi-modal problems (Beyer 2000). The algorithms can obtain suitable solutions based on either the objective functions or subjective judgements when large and sophisticated input data are given. Meanwhile, two major deficiencies in the current GAs are also noted—non-unique results and large computation time. In the literature, attempts to combine a GA with other data-driven approaches (e.g. ANN) have been made to mitigate the negative impacts arisen from the deficiencies.

## 2.1.7 Data-Driven Classification Approaches

Besides great success in predicting building energy consumption, data-driven approaches have been extensively used to attack building energy classification over the last several decades, among which K-means algorithm, self-organizing map (SOM), hierarchical clustering and regression are the most popular choices.

### 2.1.7.1 K-Means Cluster

The K-means clustering algorithm is a classification approach quite popular in building load analysis. Technically, this algorithm partitions a set of data into a number of non-hierarchical groups of similar data points, i.e., clusters. The similarity among data points is quantified by the Euclidean distance, based on which a K-mean clustering procedure includes the following

steps. A data set ( $x_i, i = 1, 2, \dots, n$ ) is first input with the cluster centers ( $\mu_j, j = 1, 2, \dots, K$ ) being specified randomly. The Euclidean distances between each data point and each cluster center are then computed. A datum  $x_i$  is set to belong to a cluster  $C_j$  if its distance to the cluster center  $\mu_j$  is shorter than those to any other center. As a consequence, this classification forms  $K$  clusters in the input dataset, and the center of each cluster is re-calculated as a mean based on new data grouping. The  $K$  mean clustering algorithm will repeat the above distance computation, data classification and center relocation till all the  $K$  cluster centers do not move their locations with further iterations (Magoules and Zhao 2016). In many cases, a squared error function  $J$  is introduced to characterize this convergence,

$$J = \sum_{j=1}^K \sum_{x_i^{(j)} \in C_j} (x_i^{(j)} - \mu_j)^2 \quad (2.19)$$

where  $x_i^{(j)}$  represents a data point belonging to the cluster  $C_j$  (Panapakidis et al. 2014). In the  $K$  mean clustering algorithm, a priori specifications of the cluster number  $K$  and initial positions of the cluster centers are required. This results in the algorithm has to be conducted several times in practice with these parameters with different values. Only the best results after comparison will be deemed as the algorithm's ultimate outcomes.

It is worth mentioning to improve its feasibility, the  $K$ -means clustering algorithm has been modified using the fuzzy methods. The modified version, in contrast to the aforementioned discussion, allows soft clustering, i.e., every data point can potentially belong to multiple clusters and a degree of membership is defined to characterize such relationships (Dunn 1973). Nikolaou et al. (2012) discusses one widely-used fuzzing cluster approach in building energy projects, i.e., fuzzy  $C$ -means (FCM) cluster. Interested readers can refer to it for more details.

### 2.1.7.2 Self-organizing Map

Self-organizing map (SOM) is developed from ANNs which transfers an incoming signal pattern in arbitrary dimensions into a one- or two- or multi-dimensional topographic map (Magoules and Zhao 2016). The method is trained by an unsupervised learning process and capable of classifying new inputs into clusters with different features in a neurobiological-like manner. Figure 2.6 illustrates a frequently-used network architecture of SOM consisting of a one-dimensional input layer and a two-dimensional computational layer. In this computational layer, a number of process units, i.e., neurons ( $j = 1, 2, \dots, m$ ), are arranged in rows and columns, each of which connects all input signals ( $x_i, i = 1, 2, \dots, n$ ) with connection weights  $w_{ij}$ . The output of the neuron  $j$  is sometimes given

$$\text{by } y_j = \sum_{i=1}^n w_{ij} x_i.$$

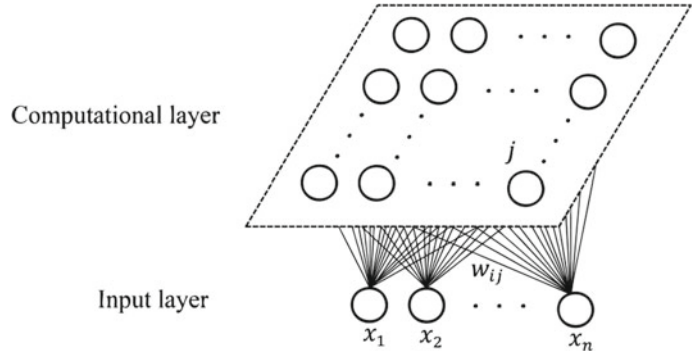
In SOM, a squared Euclidean distance between all the input signals and connection weights pertinent to every neuron is computed

$$d_j = \sum_{i=1}^n (x_i - w_{ij})^2 \quad (j = 1, 2, \dots, m) \quad (2.20)$$

This distance is termed as the discriminant function, and the neuron with the smallest discriminant function is designated as the winner for a given set of input signals. Typically, a SOM iteration starts from initializing all connection weights with small random numbers and choosing a set of input signals from historical database at random to form the input layer. Computation of the discriminant function for each neuron in the computational layer is then performed. Only the neuron with the smallest discriminant function is identified as the winner at this iterative level. Immediate to this, a topological neighborhood centered at the selected winner is defined, in which the connection weights linking every neuron to the input signals are adjusted subject to

$$w_{ij}(n+1) = w_{ij}(n) + g_j [x_i - w_{ij}(n)] \quad (2.21)$$

**Fig. 2.6** Schematic of SOM.  
 White cycles: process units;  
 Solid lines: connections



where  $n$  represents the current iterative level,  $g_j$  is the learning rate depending on  $n$  and the distance between the winner and the neighboring neuron  $j$ . The next iteration at  $n + 1$  will be conducted with these adjusted correction weights and the new randomly-chosen input signals. Note that while the SOM iteration proceeds, both the learning rate and the size of the winner’s neighborhood will decrease. The whole iteration will terminate once a threshold is met, e.g.,  $g_j \leq g_{j,min}$  or only the winner itself or none being included in the neighborhood. After training, a particular neuron (i.e., winner) in SOM will be activated the most for a particular type of input signals. This correspondence ensures SOM to be effective means used for clustering new input signals.

In sum, SOM can effectively reduce the dimensions of a high-dimensional signal pattern to a feature map in which the similarities and differences among input objects are easily discerned. Moreover, its outputs can be directly followed by further classification using other clustering algorithms. This will lead to more mutually exclusive and well-separated groups. On the other hand, it is also noted that SOM clustering suffers from oscillation if a rambling dataset without any pretreatments is used as the input. Importantly, its computational cost will dramatically increase with the increasing dimension of the data. Therefore, a good SOM should be equipped with a well-designed tuning process and a clear parametric analysis on the impacts of different parameters. These

parameters usually include the learning rate, neighborhood function, number of process units, and et al.

### 2.1.7.3 Hierarchical Clustering

Hierarchical clustering in building energy consumption commonly uses the bottom-up fashion to organize data points into a tree-like hierarchy of clusters (Nikolaou et al. 2012). Such clustering is known as the agglomerative algorithm starting with  $n$  data points. ( $x_i, i = 1, 2, \dots, n$ ), each of which is treated as a singleton cluster. To characterize the inter-cluster similarity, the distances among different clusters are computed, and form a  $n \times n$  matrix

$$H = \begin{bmatrix} 0 & \dots & D(C_n, C_1) \\ \vdots & \ddots & \vdots \\ D(C_1, C_n) & \dots & 0 \end{bmatrix} \quad (2.22)$$

In the above matrix, the distance between two clusters  $D(C_i, C_j)$  is defined by

$$D(C_i, C_j) = \min d(x_i, x_j), \text{ with } x_i \in C_i \text{ and } x_j \in C_j,$$

where  $d(x_i, x_j)$  is the distance (i.e., Euclidean distance) between two data points in these two cluster and  $D(C_i, C_j) = 0$  when  $i = j$  (Nikolaou et al. 2012). In the literature, there are the other ways to define the distance between two clusters. Interested readers can refer to Vesanto and Alhoniemi (2000) for more details. After computing the inter-cluster distances, the next step is

to merge two closest clusters having the minimal  $D(C_i, C_j)$ , and then update the corresponding distance matrix. This merging manipulation will proceed iteratively till all data points have been included in a single cluster.

In hierarchical clustering, merging can be conducted in different ways and terminated at different levels provided the similarity criterion requires. Figure 2.7 illustrates an example where two distinct sets of three clusters are obtained in different merging routes based on different merging criteria. In building energy studies, hierarchical clustering has been proven that it can reveal the data internal structure and generate useful knowledge about energy consumption in a building (Magoules and Zhao 2016).

## 2.2 Practical Application of Data-Driven Approaches

### 2.2.1 R & D Works and Practical Applications

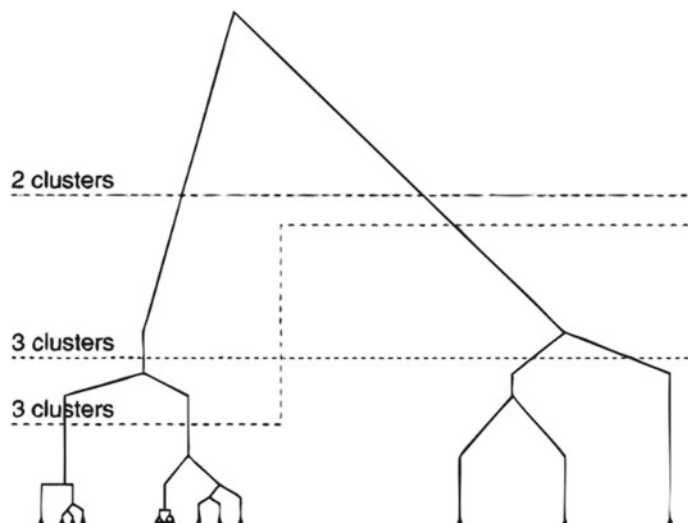
All the aforementioned data-driven approaches are widely applied to a large variety of prediction or classification applications of load prediction, energy pattern profile of specific use-cases, regional energy consumption mapping, energy benchmark for building stock, retrofit strategies

and guideline making, see a summary in Table 2.2. This broad range of applications covers micro-scale and macro-scale studies that provide useful information and instructive suggestions for different stakeholders, including government, investors, engineers and occupants throughout the building life cycle from the early planning/design stage to later operation/retrofit stage.

#### 2.2.1.1 Prediction

Originally, many data-driven approaches were established to predict the energy consumption of building, in particular electricity usage. It is well recognized that estimations of energy usage in the long-, medium- and short-term (i.e., annual, monthly and daily) are of importance for energy market planning and investments. Especially, a very short-term (hours or minutes ahead) estimation of electricity usage can exert a vital influence on the final dispatch for national electricity market (Setiawan et al. 2009). Therefore, a precise prediction in these scenarios would lead to more efficient energy management and direct to considerable reduction in operational cost for both energy suppliers and end-users in buildings (Setiawan et al. 2009; Mathieu et al. 2011; Neto and Fiorelli 2008). At the current stage, ANN and SVM are the two favourable data-driven approaches used for prediction of building energy consumption.

**Fig. 2.7** Schematic of hierarchical clustering algorithm. The partitive clusters can be obtained at different levels of similarity (Vesanto and Alhoniemi 2000)





**Table 2.2** Summary of data-driven approach for applications in building energy consumption

Data-driven approaches	Applications prediction classification	
ANN	(Kalogirou 2001; Kalogirou and Bojic 2000; Neto and Fiorelli 2008; Sözen and Arcaklioglu 2007; Yang et al. 2005; Canyurt et al. 2005; An et al. 2013; Yezioro et al. 2008; Yan and Yao 2010; Yokoyama et al. 2009; Olofsson and Andersson 2002; Aydinalp-Koksal and Ugursal 2008; Aydinalp et al. 2002; Yalcintas 2008; Asadi et al. 2014)	(Yalcintas 2006; Yalcintas and Ozturk 2006)
SVM	(Zhao and Magoulès 2012, 2010; Magoules and Zhao 2016; Dong et al. 2005; Li et al. 2009a, 2010, 2009b; Setiawan et al. 2009)	N/A
Regression	(Zhao and Magoulès 2012; Amjady 2001; Mejri et al. 2011; Wauman et al. 2013; Tiedemann 2007; Aydinalp-Koksal and Ugursal 2008; Larivière and Lafrance 1999; Mastrucci et al. 2014; Howard et al. 2011; Nikolaou et al. 2011; Chung et al. 2005; Yalcintas and Ozturk 2006)	N/A
DT	(Tso and Yau 2007; Yu et al. 2010)	(Park et al. 2016)
GA	(Canyurt et al. 2005; Sadeghi et al. 2011; Azadeh et al. 2007; Li and Su 2010; Juan, et al. 2009; Asadi et al. 2014)	N/A
K-means cluster	N/A	(Nikolaou et al. 2012; Panapakidis et al. 2014; Tsekouras et al. 2007; Heidarinejad et al. 2014; Arambula Lara et al. 2014; Jones et al. 2007; Fonseca and Schlueter 2015; Santamouris et al. 2007; Wang 2015; Lannon et al. 2050)
SOM	N/A	(Nikolaou et al. 2012; Panapakidis et al. 2014)
Hierarchical cluster	N/A	(Tsekouras et al. 2007; Xiao et al. 2012; Wang 2015)

### Prediction Application of ANNs

ANNs have been extensively used as a prediction means in diverse areas (Kalogirou 2001). In building sector, ANNs excels in predicting building energy consumption, electricity demand, heating/cooling loads, important energy parameters and even assessment of software etc. Table 2.3 has centrally summarized these applications of ANNs in the literature.

In terms of energy consumption, ANNs are the popular candidate for both the short-term and long-term prediction. Kalogirou and Bojic (2000) used ANNs to predict energy consumption in a holiday passive solar building, where engineers working in the HVAC field were not included. In their study,

the RNN model based on the back-propagation architect was applied for the training process. In so doing, such a model could detect features in the raw data of previous knowledge, e.g., the changing rules of operating conditions along different time epochs. In addition, Sözen and Arcaklioglu (2007) even derived an ANN model to shed light on causality link behind economic indicators, population and net energy consumption. Their study suggested economic indicators (e.g. gross national product (GNP) and gross domestic product (GDP) etc.), rather than conventional energy indicators (e.g. gross generation, installed capacity and years), are playing a more important role for an accurate prediction of energy consumption.

**Table 2.3** Summary of ANNs in predicting building energy consumption

Type of house	Scale	Inputs	Output	Data source	Measure length	Algorithm
Holiday passive house (Kalogirou and Bojic 2000)	Single	Season, insulation function, wall thickness, heat transfer coefficient, time of day	Energy consumption	Measured data: ZigBee Input Device (ZID)	Two seasons	RNN combined with BPNN
Multiple (Sözen and Arcaklioglu 2007)	National	Economic indicators (GNP and GDP), population	Net energy consumption	World Energy Council	37 years (1968–2005)	BPNN
Office building (Yang et al. 2005)	Single	Outdoor dry-bulb temperature, outdoor humidity, water temperature of chiller, compressor status etc	Dynamic chiller electric demand	Simulated data (DOE 2.1E) and measured data	1 year	Sliding window ANN and accumulative ANN
Office building (Canyurt et al. 2005)	Single	Previous load, temperatures of previous day, occupancy condition, sin and cosine of the hour	One day ahead electric power consumption	Great building energy predictor shootout I and measured data	1 year and a half	BPNN
Multiple An et al. 2013)	Reginal	Previous electricity consumption	Half-hour ahead electricity demand	Australian Energy Market Operator	9 weeks	Multi-output BPNN
Residential (Yan and Yao 2010)	7 buildidngs	18 building envelope parameters, heating degree day, cooling degree day	Heating and cooling energy consumption	Simulated data (DeST)	1 year	BPNN
Commercial (Yokoyama et al. 2009)	Single	Previous cooling demand, air temperature and relative humidity	Cooling demand	Measured data	45 weekdays	BPNN

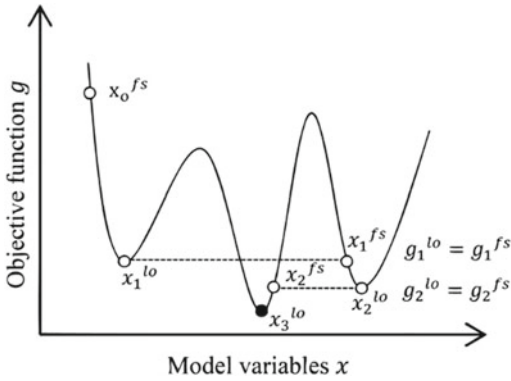
(continued)

**Table 2.3** (continued)

Type of house	Scale	Inputs	Output	Data source	Measure length	Algorithm
Residential (Olofsson and Andersson 2002)	7 single family-building	Supplied heating demand, electricity domestic demand, flag parameter	Indoor-outdoor temperature difference	Measured data	2 years	BPNN
Solar house (Yezioro et al. 2008)	Single	Outdoor temperature, relative humidity, set point temperature, occupancy schedule	Heating/cooling consumption	Measured data	2 days	BPNN
Office building (Neto and Fiorelli 2008)	Single	Outdoor dry-bulb temperature, day type (working day or weekend)	Daily total consumption	Measured data: energy demand measurement system	54 days	BPNN

As to electricity demand, the majority of ANN models focus on dynamic and short-term predictions, which require careful selection and pre-treatment of input data. One example is Yang et al. (2005) where an on-line chiller electricity prediction model was established through use of both the simulated data and measured data. Their results recommends the sliding-window ANN, which constantly drops the oldest data and adds new measurements during training process, showed better performance than the accumulative ANN based on measured data. Besides, Karatasou et al. has reported one-day ahead prediction of electricity consumption, called a 24-steps predictor, in Canyon et al. (2005). The predictor used previous energy consumption data records with time delays larger than 24 h as inputs to train the network to perform next day's prediction. Interestingly, An et al. (2013) further developed an (EMD)-based signal filtering which is able to forecast half-hour electricity demand ahead. Such an EMD-based signal filtering can decompose an incoming signal into a series of pure modes and residues. The results revealed that the EMD-based filter a critically-functioned

component in the ANNs prediction model. In fact, ANNs also play an important role in prediction of heating/cooling loads. In this particular type of applications, the ANN models usually require to input detailed climate information, envelop parameters and occupancy schedules (Yezioro et al. 2008; Yan and Yao 2010). Besides reliable input data, algorithm optimization is the other way to promote the prediction accuracy. To minimize the drawback of BPNN (e.g. local optimization of model parameters in training process), a global optimization called "Modal Trimming Method" was proposed by Yokoyama et al. (2009). This method was composed of two steps, shown as Fig. 2.8: (1) search for local optimal solution of input variables in an objective function ( $x_0^{fs} \rightarrow x_1^{lo}$ ). Normally, the objective function is defined as calculation error between predicted and measured values; (2) search for another feasible solution of the same objective function value with previous local optimization ( $x_1^{lo} \rightarrow x_1^{fs}$ ). These two steps were repeated ( $x_1^{fs} \rightarrow x_2^{lo} \rightarrow x_2^{fs}$ ) until tentative global optimal one  $x_3^{lo}$  is found. They validated



**Fig. 2.8** Concept of modal trimming method

this method and concluded that significant error of predicted cooling demand from measured data was reduced compared to traditional local optimization method.

On top of the above applications, ANNs' application is also extended to predicting the key parameters of energy performance of building. For instance, Olofsson and Andersson (2002) proposed a use of the BPNN model to estimate the total heat loss coefficient (HLC) and domestic energy gain factor of inhabited single-family buildings. Here, the total HLC characterizes heat loss resulted from transmission and air-flow while the domestic energy gain factor focuses on the gain of heating or cooling from inside sources. In this kind of ANN model, flag parameter of each measured case was introduced to distinguish non-linear dependences among various predictors, instead of average dependency from previous experience.

It is worth mentioning that ANNs are sometimes used as tools to assess simulation software for building performance. Neto and Fiorelli (2008) compared the BPNN against EnergyPlus by using both to predict building energy consumption. The latter is recognized a mainstream simulator in building sector which can deliver much more accurate results than Energy\_10, Green Building Studio web tool, and eQuest (Yezioro et al. 2008). Interestingly, Neto et al. found that when building and climate data were just briefly described, the used BPNN model works much better in daily energy demands

prediction than EnergyPlus does. Importantly, especially for hourly prediction, all these simulation tools in current market give rather poor results in comparison to ANNs. This finding equips ANNs a new function as a benchmark to test accuracy of commercial software for estimating building energy performance.

### Prediction Application of SVM

Prediction is also a primary function of SVM use in building energy simulation. Table 2.4 lists the up-to-date studies on SVM-prediction applications. Generally speaking, SVM works in high accuracy in the medium-term (Dong et al. 2005) and short-term (Setiawan et al. 2009) prediction. Significantly, the method only requires a few model-parameters to implement its calculation. On the other hand, however, computing speed of SVM is slower than that of other approaches, such as linear regression and the ANNs. Currently, how to optimize SVM algorithm is regarded as the core task for its future development.

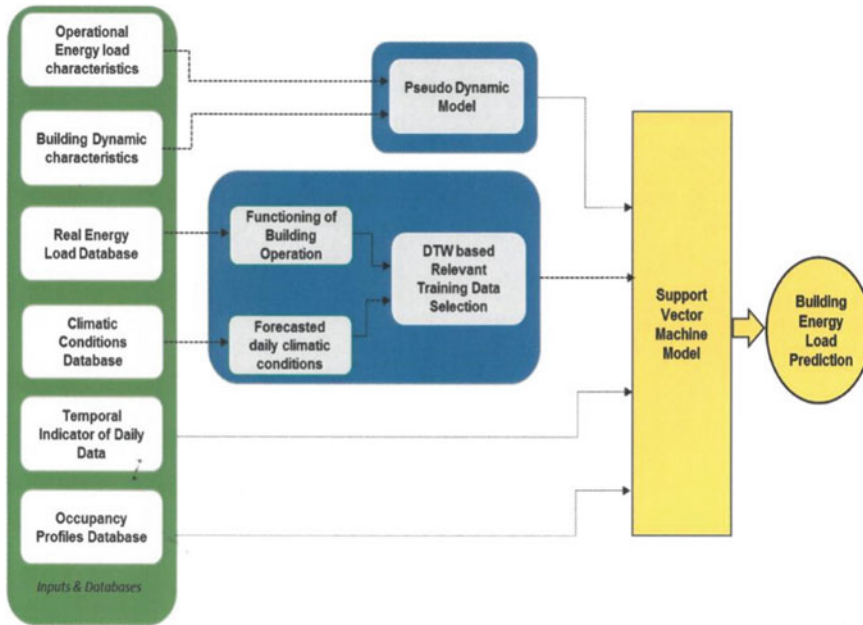
Many efforts were actually made on SVM optimization in recent years. To save the computer memory and expedite the time-consuming training process, Zhao and Magoulès (2010) proposed targeted solutions for dual optimization process (see Eq. (2.5)) and Kernel function calculation. The main idea was to divide the entire dual optimization problem into sub-problems and calculate them in parallel. Then, the Kernel function matrix would be updated for each sub-problem calculation. This parallelized training process could be stopped until convergence. The modified SVM gains a capability of dealing with a large amount of data to predict energy consumption of multiple buildings. Another possible optimization solution is to develop a hybrid SVM. For example, Li et al. (2010) presented a hybrid approach combining SVM and FCM clustering algorithm to forecast building cooling loads. In this research, FCM was first employed to extract valid data records from the pool of raw data, and then the SVM followed with a training procedure based on the extracted valid data records. Clearly, such a pretreatment of data records effectively reduce the noise of inputs for SVM calculation. It should be pointed out the SVM is compatible with diverse input

**Table 2.4** Summary of SVM in predicting building energy consumption

Type of house	Scale	Inputs	Output	Data source	Measure length	Algorithm
Commercial (Dong et al. 2005)	4 single buildings	Outdoor temperature, relative humidity, global solar radiation, previous electricity consumption	Building energy consumption per month	Survey: monthly utility bill, National Environment Agency	3 years	SVM
Multiple (Setiawan et al. 2009)	Regional	Historical electricity consumption data	5-min ahead electricity load	Australian electricity operator	3 years	SVM, statistical regression, and BPNN
Office buildings (Zhao and Magoulès 2010)	100 buildings	Heating consumption, electrical consumption	Heating demand, electrical load	Simulated data (EnergyPlus)	5 months	Parallel SVM
Campus building (Li et al. 2010)	Single	Cooling load	Cooling load	Measured data	4 months	Fuzzy SVM combined FCM clustering
Office buildings (Zhao and Magoulès 2012)	Single	Previous energy load, building dynamic characteristics, outdoor temperature, occupancy schedule	Building energy demand	Measured data: data acquisition system	7 months	SVM with pseudo dynamic approach
Office buildings (Li et al. 2009b)	Single	Previous cooling load, air temperature, relative humidity, solar radiation intensity	Hourly cooling load	Simulated data (DeST)	Half year	SVM, BPNN

information. Besides the conventionally-used energy loads and climate conditions, Zhao and Magoulès (2012) also used energy load characteristics and hidden inertial effects of building as their SVM inputs. The energy load characteristics were described as operation level of HVAC system and the occupancy profile, the hidden inertial effects were provided as fluctuation of internal temperature. The block diagram of such a model for predicting building energy consumption is shown in Fig. 2.9. As we can see, partial selection of input data called dynamic time warping

(DTW) was adopted during prediction process, which measures on the outdoor temperature difference between training days and prediction days. The minimal difference between two time series was chosen as optimal path for solution. Similarly, the previous energy load database was also partially selected by DTW as inputs to consider the most recent data rather than whole data. The result showed that the designed training leads to higher accuracy and better computational efficiency in comparison to that based on the whole input data.



**Fig. 2.9** Block diagram of SVM in prediction of energy demand using pseudo dynamic approach (Zhao and Magoulès 2012)

### Prediction Application of Statistical Regression

Statistical regression is well treated as a simple tool for prediction for a long time (Zhao and Magoulès 2012). However, this approach suffers from low-accuracy in its prediction results, and such a deficiency has greatly limited its applications in building energy consumption analysis. This motivates a great deal of modification and optimization in statistical regression, which are briefly illustrated in Table 2.5. Among various modifications are multiple linear regression (MLR) proposed by Li and Huang (2013) for short-term prediction. This model utilized not only climate data, room temperature set point, but also the cooling loads of previous four hours as its inputs. The obtained prediction results achieved very impressive accuracy higher than that of conventional ANN models. Moreover, autoregressive, integrated and moving average (ARIMA) model under the statistical regression framework was designed to correlate time-series data. Amjady's study (2001) has well examined the exactitude of ARIMA model for predicting

daily peak and hourly load based on national power net. He further extended ARIMA model with use of the estimated electricity load as an extra input. The accuracy of his model reach a higher level even compared to original ARIMA and ANNs.

In most cases, statistical regression models are adopted to estimate important parameters characterizing energy performance. For instance, Mejri's et al. (2011) investigated statistical regression modelling for predicting indoor air temperature. In their study, they analyzed the similarity in dynamic behaviours among different thermal zones for HVAC system design. Another example goes to Wauman's et al. (2013), where they used statistical regression to explore correlation between heat balance ratio and heat gain factor of some school buildings exemplified in their research. These obtained correlations are regarded of crucial significance for designing, tracing and analyzing building thermal behaviours. They are also important supportive materials for drafting heating control strategy for energy saving.



**Table 2.5** Summary of statistic regression, DT and GA in predicting building energy consumption

Type of house	Scale	Inputs	Output	Data source	Measure length	Algorithm
Office building (Li and Huang 2013)	Single	Dry bulb outdoor air temperature, solar horizontal radiation, and room temperature set point, cooling load of previous 4 h	Cooling load	Simulated data from TRNSYS	60 measured case	MLR, ANN, grey-box approach
Multiple (Amjady 2001)	National	Previous load, estimated current load, temperature	Hourly electricity load and daily peak	National dispatching center	1 year	ARIMA
Office building (Mejri et al. 2011)	Single	Indoor temperature of four rooms	Room temperature	Measured data	2 months	Statistical regression
Multiple (Tso and Yau 2007)	Groups	Power rating of appliance, consumption time	Average weekly electricity consumption	Survey	Two seasons	BPNN, Least-squares regression, DT
Residential (Yu et al. 2010)	80 buildings	Outdoor temperature, building characteristics, appliance energy source and usage (10 inputs)	Energy use intensity	Survey and research committee	3 years	DT
residential-commercial (Canyurt et al. 2005)	National	GDP, population, import, export, house production, basic house appliance consumption figures	Future energy demand (2003–2030)	World Energy Council and State Statistics Institute	8 years	GA
Residential (Sadeghi et al. 2011)	National	GDP, real price of electricity and natural gas in residential sector	Future per-capita consumption of electricity (2009–2025)	Iran Statistics Center, Central Bank of Iran	39 years	GA
Hotel (Li and Su 2010)	Single	Outdoor temperature of past 2 days, air conditioning consumption of past 3 days	Daily air conditioning consumption	Measured	7 months	GA-HANFIS

### Prediction Application of Decision Tree

In the large family of data-driven approaches for building energy consumption prediction, DT is a

relatively new member, but involves much simple techniques. Tso and Yau (2007) compared statistic regression method, BPNN and DT by

predicting the electricity consumption in summer and winter periods. Results showed that DT used in their study performed as well as BPNN, both of which deliver accurate results than statistical regression did. Yu et al. (2010) also applied the DT approach to predict energy use intensity (EUI) of residential buildings. They designed ten predictor variables concerning indoor temperature, building envelop, appliance types and occupant number in the DT framework. Their result clearly demonstrated that DT is able to well predict building energy consumption level as high/medium/low. The significances of these predictor variables were ranked in terms of degree of closeness to the outdoor temperature (predictor variable of root node), which is the most important determinant of EUI. The results showed that several building parameters, e.g. heat loss coefficient and equivalent leak area, deserve more attention at early design stage and benefit energy conservation in retrofit.

### Prediction Application of Genetic Algorithms

GA has been regarded as a powerful prediction approach in building energy consumption. As shown in Table 2.5, most applications of GA models are national analysis. One typical example is prediction model of energy consumption for residential-commercial building section in Canyurt et al. (2005). Three different scenarios were proposed in order to find out the best fit solution. The result showed that GA model, which considers residential housing production, house appliances of washing machine, television, vacuum cleaner and refrigerator as the input parameters, can obtain the most accurate quadratic prediction model of energy consumption. Sadeghi et al. (2011) developed prediction model of electricity consumption using GA on national level. It was found out that exponential equation had the more accurate results compared to linear and quadratic forms.

Hybrid methods of GA and ANNs are widely used in electricity prediction application (Azadeh et al. 2007). Li and Su (2010) predicted the daily air-conditioning consumption by using the genetic algorithm-hierarchical adaptive network-based fuzzy inference system (GA-HANFIS).

Before developing prediction model, clustering algorithm was applied to identify the nature groups and qualities of a large data set, and GA was used to optimize the unknown cluster-parameters through minimizing the error of predicting result. Figure 2.10 shows the architecture of GA-HANFIS, in which the outdoor temperature of predicted day  $T(k)$ , the air-conditioning consumption of past two days  $y(k-1)$  and  $y(k-2)$  were identified as more significant inputs of network layer 1. These less significant variables  $T(k-1)$ ,  $y(k-3)$ ,  $T(k-2)$  and  $T(k-3)$  were selected as inputs of network layer 2 and layer 3. Output  $y(k)$  was air-conditioning consumption of predicted day. The rule base of each layer contained two if-then rules; readers can refer to Tsekouras et al. (2007) for more details. Moreover, the calculation rules were different according to different clusters. This hybrid method outperformed regular BPNN in prediction accuracy.

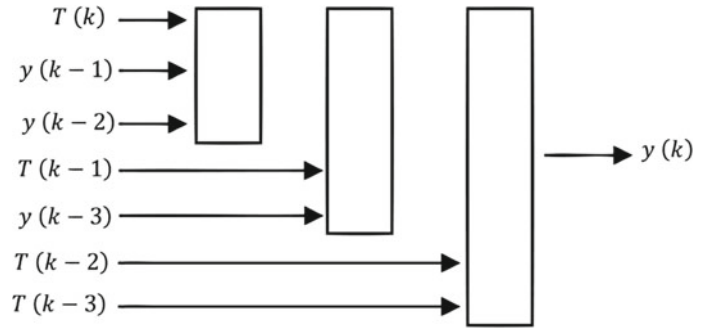
### 2.2.1.2 Profile

The energy consumption profile in building is to quantify the total consumption contribution to sub-components, or further distinguish the usage characteristics. Regarding the positive influence for end-users, the capability of profiling the energy use as the feedback can educate the occupants on how to consume and change the consumption behaviours to certain extent. As for utility companies, DSM measures are implied after extracting load profiles in order to reach a proper load-shape objective, i.e. “peak clipping”, “valley filling”, “strategic conservation”, “flexible load shape”, “load building” and “load shifting” (Panapakidis et al. 2014). The commonly-used methods for energy and electricity profiling are clustering based method, which is detailed in Table 2.6.

### Profile Application of Cluster Method

As one application of cluster method, analyzing electricity behaviour through pattern recognition and load curve classification has been investigated by massive researches. Tsekouras et al. (2007) developed a two-stage pattern recognition for customer’s classification. The first stage was

**Fig. 2.10** Architecture of GA-HANFIS model with 3 layers (Li and Su 2010)



**Table 2.6** Summary of data-driven approaches in building energy consumption profiling

Type of house	Scale	Inputs	Data source	Measure length	Algorithm
Multiple (Tsekouras et al. 2007)	94 buildings	Daily electricity consumption	Meters	10 months	K-means cluster, fuzzy K-means cluster, seven hierarchical cluster
University building (Panapakidis et al. 2014)	27 buildings	Daily electricity consumption	Meters	2 years	SOM combined with K means++
National office buildings (Xiao et al. 2012)	24 provinces	Annual electricity consumption	Survey	1 year	Hierarchical cluster
High performance buildings (Heidarinejad et al. 2014)	134 buildings	Energy end use	Simulated data	3 year	K-means cluster
Campus buildings (Arambula Lara et al. 2014)	85 buildings	Heating demand	School manage service	5 years	K-means cluster combined with MLR
Residential buildings (Tiedemann 2007)	791 customers	Weather data and energy consumption of residential appliances	Survey	2 years	Statistical regression
Residential buildings (Aydinalp-Koksal and Ugursal 2008)	8767 customers	Weather data and energy consumption of residential appliances	Survey	1 years	Statistical regression and ANN
Residential buildings (Aydinalp et al. 2002)	8767 customers	Appliance, lighting, cooling loads, space heating, domestic heat water	Survey	1 year	PBNN

to pattern load curves of each customer; the second stage was to cluster the customers according to pattern features. In their research, K-means cluster was proven by adequacy

measures as the most appropriate approach compared to other methods. The function of adequacy measures is to evaluate the within-group similarity and between-group dissimilarity,

in order to obtain a well-separated classification. Panapakidis et al. (2014) incorporated K-means++ cluster within SOM to reduce the number of centers and increase the accuracy. The data records including vast of load curves were aggregated from various buildings, SOM was thus an appropriate approach to map high-dimensional database into low-dimensional patterns. As the improvement of the basic K-means clustering algorithm, K-means++ algorithm tries to initialize the centroids that far from each other rather than random selection. The combination of SOM and K-means++ resulted in small errors in all cases.

Recently, cluster method becomes prevailing to profile EUI of buildings on large-scale. Xiao et al. (2012) conducted a study on EUI (excluding district heating) of business office buildings in China. Each data point was defined as  $(x_{1i}, x_{2i})$  in which  $x_{1i}$  and  $x_{2i}$  refer to EUI and gross floor area of corresponding building. Eventually, two clusters were formed by using hierarchical cluster and the frequency distribution of EUI is illustrated in Fig. 2.11. The cluster results revealed the unique “dual section distribution” pattern which is different from developed countries. Heidarinejad et al. (2014) used K-means cluster algorithm to classify the EUI of 134 U.S. high-performance buildings (HPBs) by the squared Euclidean distance. These HPBs were well separated into three clusters, as high/medium/low EUI. Studies showed that unregulated loads which include various equipment and uncategorized loads, accounted for 30–40% total energy consumption that should be reduced specifically through effective programs and modification. It can be found out that studies mentioned above that analyze building energy issues on large scale, are greatly dependent on the clustering methodology.

Clustering technology can be also applied for heating/cooling demand classification. K-means cluster analysis combined with MLR were proposed by Arambula et al. (2014) to analysis the heating demand of 85 high schools. In their model, MLR analysis was firstly conducted to select 6 significant building thermal indicators

according to  $R^2$  value Eq. (2.12). Three clusters were developed by K-means cluster analysis based on these 6 indicators, while later  $R^2$  was calculated for each cluster. The regression analysis showed that cluster 3 need to be further divided by clustering analysis since its low within-group similarity ( $R^2 < 0.5$ ). Finally, more reasonable classification results could be obtained after such twice MLR analysis and twice clustering analysis when comparing to the sole clustering.

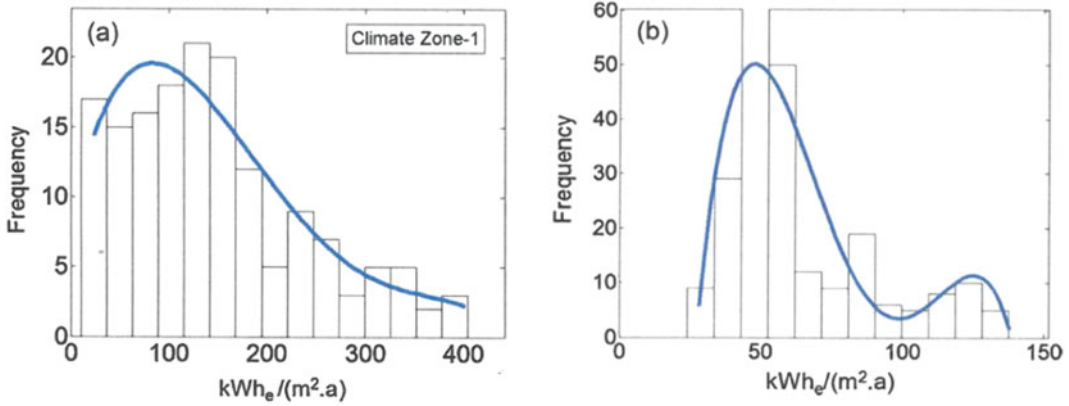
### Profile Application of Regression

One regression method specialized for profiling energy consumption of residential buildings is conditional demand analysis (CDA). The basic idea of the CDA model is that total household consumption is the sum of various end-use consumptions.

CDA is frequently used to profile building energy consumption at national level (Tiedemann 2007). Aydinalp-Koksal and Ugursal (2008) used CDA to profile residential end-use energy consumption at national level, large-scale database including the surveys from occupants, weather conditions as well as historical energy bills were used. Their CDA model adopts 6 electricity end-uses including main and supplementary space heating, domestic heating water, space cooling, lighting, major and minor appliances. Meanwhile, they also developed neutral network model for comparison purpose (Aydinalp et al. 2002). In their research, BPNN outperformed CDA model in evaluating the effects of socio-economic factors, such as income, dwelling ownership and area sizes of residence. Because these socio-economic factors were considered as input variables in BPNN while CDA cannot not include comprehensive variables due to the limitation of statistical regression.

#### 2.2.1.3 Energy Mapping

Energy mapping methods, usually based on the Geographic Information System (GIS) city building database, consider using data-driven technology for pre-and post-progressive



**Fig. 2.11** Frequency distribution and polynomial fitting plot of EIU in office buildings. **a** US climate zone-1. **b** Certain city of China (excluding district heating) (Xiao et al. 2012)

operation (Caputo et al. 2013). Thanks to the capabilities of GIS, immediate updating of energy evaluation and visual representation via maps are both permitted in a user-friendly model, to provide energy consumption distribution within the city. Among the massive technologies for energy mapping, statistical regression (MLR) and clustering algorithm are the mostly

utilized data-driven methods, as displayed in Table 2.7.

MLR is a traditional used approach in energy mapping of building section at zip-code level (Larivière and Lafrance 1999). Mastrucci et al. (2014) applied MLR model to map the energy consumption of dwellings in a city of Dutch. The contributing inputs included floor area, number

**Table 2.7** Summary of data-driven approaches in energy mapping

Type of house	Scale	Energy consumption Inputs	Data source	Measure length	Algorithm
Multiple (Larivière and Lafrance 1999)	45 cities	Annual electricity, population density, age of inhabitants, weather condition, living expenditure	Organizations involved in city affairs	1 year	MLR
Multiple (Howard et al. 2011)	City	Annual electricity consumption, natural gas, steam, and fuel oil consumption	Utility company, RECS, CBECS, georectified database	1 year	MLR
Residential buildings (Mastrucci et al. 2014)	City	Yearly consumption of natural gas and electricity	GIS database, Royal Netherlands Meteorological Institute, metering	Nearly 50 years	MLR
Residential buildings (Jones et al. 2007)	City	Energy consumption and carbon dioxide emission	GIS database, electricity map, surveys	18 months	Clustering algorithm
Multiple (Fonseca and Schlueter 2015)	City	Electricity load, thermal loads,	GIS database, simulated and measured data, building standards	1 year	K-means cluster

of occupants and type of house defined for each combination of type of dwelling and period of construction during 50 years. The predicted natural gas consumption was apportioned into space heating, domestic hot water and cooking. The results observed that space heating is the biggest contribution (average 50%) in energy consumption. After 50 years tracking from 1965, they assumed the percentage of energy reduction is nearly zero for dwellings after 2005. Besides single building function of residential family, energy mapping model has been expended to profile the building energy consumption of multiple functions. For example, Howard et al. (2011) calculated the annual EUI in New York City through MLR analysis, both the tax lot designations and building area categories were used to place the buildings into  $n$  building functions (e.g. residential family, office, warehouse, education and et al.). The MLR analysis is explained in Eq. (2.16),  $y_i$  is the energy consumption of  $i$ th zip-code,  $x_{in}$  is the total building area of each building function in  $i$ th zip-code.  $\beta_i$  is the coefficient need to be determined in MLR. On top of that, it was found out extra contributing inputs are needed in some regions to distinct the unique characteristics of energy consumption. However, the research excluded energy consumption for cooking, electrical heating and other end-uses, which inevitably causes errors in energy mapping.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} \quad (2.23)$$

Clustering algorithm is typically used as subsidiary approach for mapping the energy consumption at urban scale. In Jones et al. (2007) research, cluster analysis technique was adopted to classify 55,000 dwellings with similar energy consumption and carbon dioxide emission in a Local Authority of UK. The energy rating results and carbon dioxide emission results were profiled on the regional map for further retrofit purpose. Clustering algorithm is not limited to classification of energy consumption, also utilized to develop geographical clusters. Instance, Yamaguchi et al. (2007) proposed a district clustering model for commercial buildings in

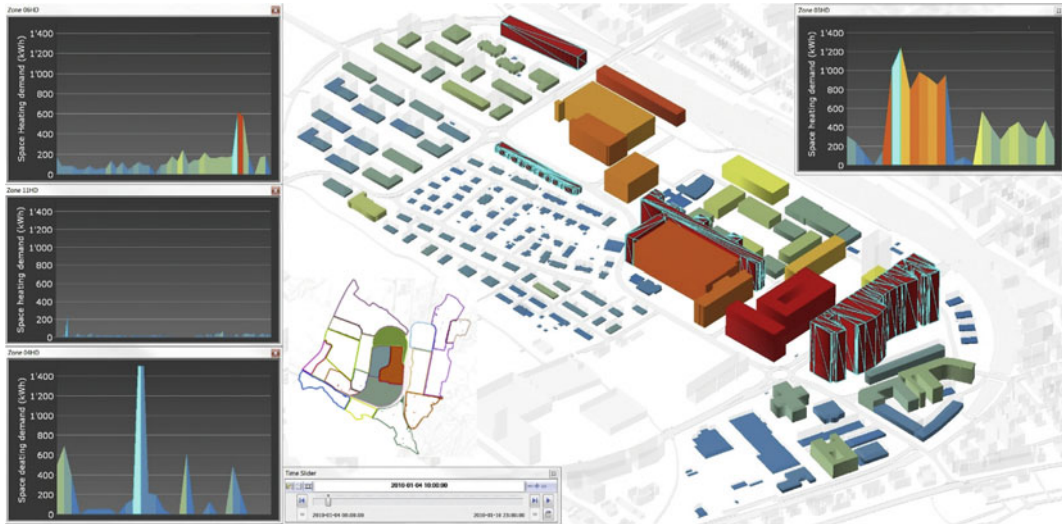
Osaka city. Firstly, clustering of district were presented by small grid cells, each of them was classified to certain representative building-type category. Then, EUI was used as evaluation for the district typology. Fonseca and Schlueter (2015) proposed a model for mapping the spatiotemporal building energy consumption in a city district of Switzerland. The model involved K-means cluster for spatial grouping in the band of 50–200 m, where spatial association of every variable of interesting buildings was strongly persistent (e.g. infrastructure types and temperature requirements). Two significant variables were used to measure the intensity of spatial clusters and similarity of groups. GIS framework gathered overall results and enabled 4D visualization that provides understandable display. The peak space heating demand of buildings in four zones at 10–11 am (April 1st, 2010) is presented as Fig. 2.12. The height and color code of buildings represent the demand level in relation to their associated zones.

#### 2.2.1.4 Benchmarking of Buildings

Different from individual building energy analysis, benchmarking was used to address large-scale building energy related issues. Two fundamental issues in benchmarking are: (1) ascertaining the current energy performance of certain building (good, average or poor) compared to same types of building stock; (2) identifying the previous/current energy performance for energy saving potential and retrofit changes (Nikolaou et al. 2011). Regression based model, ANNs, cluster algorithms and DT are the typical data-driven techniques for building energy benchmarking. Table 2.8 provides the benchmarking pilots that usually adopt EUI as the single benchmarking index.

Regression technique is one popular method in building energy performance benchmarking. Chung et al. (2005) benchmarked the EUI of 30 supermarkets in Hong Kong. MLR model was established to calculate EUI based on nine significant variables. By using bootstrapping function (Efron and Tibshirani 1993) for the empirical sample  $\{EUI_{(1)}, EUI_{(2)} \dots EUI_{(30)}\}$ , they





**Fig. 2.12** Spatio-temporal energy map of space heating demand of a city district in Switzerland (Yamaguchi et al. 2007)

obtained the estimation of EUI cumulative distribution as percentiles  $\{EUI_{10}, EUI_{20}, \dots, EUI_{90}\}$ . Although conducting on small-scale samples, they formed a benchmarking table through the percentiles. The results showed that average value of energy consumption is greater than UK energy benchmarking. They also raised the suggestions that only unmanageable factors (e.g. building thermal characteristics) should be considered during benchmarking process while all manageable variables (e.g. occupancy behaviour) were set into average values, in order to present clearer improvement suggestions for government.

ANNs method in energy benchmarking was initially presented by Yalcintas (2006). He developed three sub-models to predict EUIs as output for the plug load, lighting and HVAC components over 60 mix-used buildings. The information from questionnaire includes lighting types, floor area, equipment types and hours were used as inputs. The elaborated ANNs model could identify the EUI if new data is entered. The most outstanding advantage of ANNs benchmarking method is to renew the algorithm itself rather than manual update. Yalcintas and Ozturk (2006) also developed a national energy benchmarking model for commercial buildings based on ANNs. Different from abovementioned ANN

model which included continuous value of inputs/output, both input variables and output EUI were standardized into categorical forms for classification purpose in this model. In order to avoid inappropriate benchmarking results, database was firstly divided into 9 geographic regions. The results showed that ANN model provides more accurate EUI estimation and reasonable benchmarking result than MLR model in all cases except one.

Fuzzy cluster algorithm is a frequently-used methodology for energy benchmarking for buildings. Santamouris et al. (2007) proposed an energy rating system for 340 schools based on fuzzy clustering technology. Five classes of total and thermal energy consumption had been defined. Compared to frequency distribution rating system, fuzzy clustering rating system is more reasonable to avoid unbalanced classification, such as too small or too large range. Apart from building energy consumption benchmarking, thermal comfort rating system was also proposed by Nikolaou et al. (2012) based on FCM cluster. The predicted mean vote index, which represents mean response about thermal comfort from a larger group of people, was used as thermal comfort indicator. In their study, the thermal comfort of each climate zone was

**Table 2.8** Summary of data-driven approach in building energy benchmarking

Type of house	Scale	Benchmarking variables	Data source	Measure length	Algorithm
Supermarkets (Chung et al. 2005)	30 buildings	Building age, occupancy condition, indoor temperature, energy system type and et al.	Survey	45 year	MLR
mix-used buildings (Yalcintas 2006)	60 buildings	plug load, lighting, HVAC	Questionnaire	1 year	BPNN
Commercial buildings (Yalcintas and Ozturk 2006)	National	building-operation hours, age category, building-area, cooling category, lighting category, CDD, number of floors category	CBECS	1 year	ANNs, MLR
School buildings (Santamouris et al. 2007)	340 buildings	Heating demand, electricity demand, total energy consumption	Energy bills	3 years	Fuzzy cluster
office buildings (Nikolaou et al. 2012)	30,000 buildings	Heating load, cooling load, thermal comfort	Simulated data: VBD	1 year	K-means cluster, SOM, FCM cluster
Residential buildings (Wang 2015)	324 buildings	EUI, CDD efficiency, HDD efficiency, bath room oriented, total room oriented efficiency and etc	Panel dataset	3 years	TOPSIS, PCA, K-means cluster
Commercial buildings (Park et al. 2016)	1072 buildings	EUI, gross floor area, building use ratio	Official building register, Korea Appraisal Board	3 years	DT

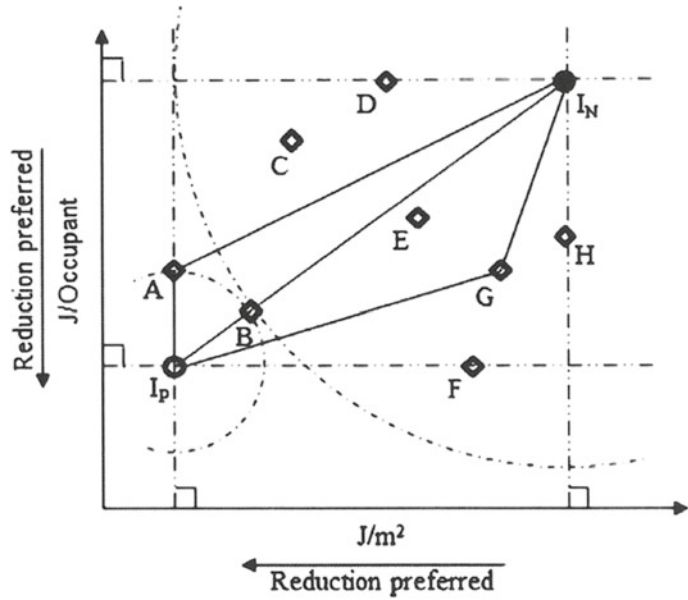
classified as three clusters, respectively. The majority commercial buildings in Greek were belong to class 2, while “best practice office buildings” were belong to class 1.

Another energy benchmarking method for improving energy efficiency of office building is DT. Park et al. (2016) developed DT model to benchmark the energy consumption of 1072 office buildings in South Korea. Gross floor area and building use ratio were identified as two significant predictor variables by correlation analysis, source EUI was defined as target variable. As the result, six rating groups of EUI were developed for each type of building use. After establishment of benchmarking model, analysis of variance was utilized to test the difference

among groups. DT model was believed to improve the conventional baseline benchmarking system via a more reasonable and fair classification.

Although most benchmarking projects are developed based on single EUI indicator, there is much effort for the multi-criteria benchmarking indicators. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) based energy efficiency benchmarking approach using seven indicators was developed by Wang et al. (Wang 2015). The illustration of TOPSIS of two indicators (energy use per occupant and EUI) is showed as Fig. 2.13. When building A and building B have the same distances to the most energy efficient condition  $I_p$ , distances to the

**Fig. 2.13** Illustrative example of TOPSIS for building energy benchmarking (Wang 2015)



least energy efficient condition  $I_N$  were used to evaluate A and B. Since MLR cannot easily produce reliable weights among highly correlated indicators, principle component analysis (PCA) was adopted to weight the importance of seven energy indicators. PCA can transform a high-dimensional dataset consisting of possibly correlated variables into a less number of their linear combinations. Finally, K-means cluster was adopted to classify the TOPSIS space into six categories as benchmarking table. Without a doubt, the benefits were obvious compared to single-criteria benchmarking which is observed with collision during evaluation process.

### 2.2.1.5 Retrofit of Buildings

Retrofit is based on the knowledge of energy profiling and benchmarking on existing buildings, presenting the largest potential of incorporation of renewable energy technology and energy conservation after efficiency retrofit measures. ANNs and GA are the main data-driven approaches in building retrofit projects, within a brief introduction in Table 2.9.

ANNs are usually applied to predict energy-saving potential for single retrofit project. Yalcintas et al. (2008) developed BPNN model for two hotel equipment-retrofit projects. Energy

usage data, weather data and occupancy data of post-retrofit period were used to train the neural network model. It then estimated the energy consumption of pre-retrofit equipment as output. The difference between recorded and predicted energy consumption was regarded as the energy saving.

As a powerful optimization algorithm, GA has been frequently adopted as the evaluation tool in building retrofit project. Juan et al. (2009) presented a GA-based on-line decision support system to offer residents a series of optimal refurbishment actions considering two objectives, cost and quality. In GA, each chromosome represented a set of retrofit solutions, the distance between chromosome and trade-off curve of cost and quality was used as fitness function to select the parents for generation. With the process of evolution, the trade-off curve would gradually converge to the best retrofit solutions with higher quality and acceptable cost.

Developed based on two-objective optimization, multi-objective optimization model was conducted by Asadi et al. (2014). They adopted GA associated ANNs to study the interaction between three main conflicting target variables, including energy consumption (EC), retrofit cost (EC), thermal discomfort hours (TDH) and

**Table 2.9** Summary of data-driven approach in building retrofit

Type of house	Retrofit measures	Algorithm
Hotels (Yalcintas 2008)	Install energy management systems and Variable Frequency Drives (VFDs) on the air-handling units. New cooling towers and VFDs on motor fans	BPNN
Residential buildings (Juan et al. 2009)	More than twenty retrofit measures under six main criterion, including safety, usage, convenient, comfortable, utility and health	GA
School (Asadi et al. 2014)	External wall insulation materials, roof insulation materials, the windows type, solar collector type, the HVAC systems	ANNs and GA
Residential buildings (Lannon et al. 2050)	Lower the carbon dioxide emission of grid, renewable resources, improve energy efficiency, and change occupancy behaviour	K-means cluster algorithm

assess their trade-offs in school retrofit project. First, the database was created in simulation tool for training and validating ANN model. BPNN model adopted in this study was composed of input layer representing different retrofit measures, one hidden layers and one output layer of energy consumption and thermal discomfort indicator. Then, the GA tool was used for minimize these three target variables as Eq. (2.17) and provide optimal combinations of retrofit measures.

$$\min y_1 = EC(X) \quad (2.24a)$$

$$\min y_2 = RC(X) \quad (2.24b)$$

$$\min y_3 = TDH(X) \quad (2.24c)$$

$X = \{x_{WALL}, x_{ROOF}, x_{WINDOW}, x_{COLLECTOR}, x_{HVAC}\}$

where  $x$  represent different materials/types of alternative retrofit choices. The treat-off curves of multi-objective optimization could be available on 3D visualization. The proposed approach presented variety of recommendations with high computation efficiency. However, simultaneous optimizations of conflicting variables gave large diversity of retrofit choices, which are difficult to understand the impact of each retrofit action at whole level.

Cluster algorithm is usually adopted to make a distinction of retrofit measures among different buildings on large scale. Lannon et al. (2050) developed model of 55,000 houses over 50-year performance via cluster analysis, aiming to

investigate the retrofit pathways to UK government's ambitious target of 80% reduction greenhouse gases emission by 2050. 100 clusters were developed to identify the dwelling with similar energy consumption and built age. Different combinations of retrofit measures were proposed and analyzed in the simulation tool. Overall, challenges and barriers in aggregate are still difficulties for individual family house.

## 2.2.2 Analyses of the Review Works

Data-driven approaches for predicting and classifying building energy consumption typically focus on total energy consumption, electricity demand, heating/cooling load and important energy parameters. The scopes of these researches are from sub-system level to single building level or even to national level.

Substantial up-to-date mythologies are proposed in order to enhance the accuracy and reliability of data-driven models, such as algorithm optimization and data pretreatment. As for algorithm optimization, micro-scale researches based on individual buildings are proposed with considerations to develop variants of basic algorithms and hybrids of several approaches (Zhao and Magoulès 2012; Kalogirou and Bojic 2000; Yang et al. 2005; Li et al. 2010; Li and Su 2010). The improvements of macro-scale analyses of building energy performance are invested to increase calculation efficiency when the raw data is large and chaos (Zhao and Magoulès

2010; An et al. 2013; Tsekouras et al. 2007; Wang 2015). In addition to algorithm optimization, data pretreatment is another focus for many researches (Yokoyama et al. 2009; Li et al. 2010; Amjady 2001; Arambula Lara et al. 2014). Appropriate pretreatment layered on the top of data-driven approach is the premise of accurate results and high computation efficiency. In short, high similarity between training and testing dataset is important for establishing a good model.

Meanwhile, substantial studies applied the simulated database to test model performance rather than the measured data. The analysis results of these models cannot be regarded persuasive enough since simulated data records are less fluctuant than real situation. In these scenarios, the question arises for reliability of simulated data again with no clear answer.

So far, the researches on residential buildings are not elaborated as researches of commercial buildings. The main reasons are including (1) lack of energy-use database from family-houses; (2) more freedom of occupancy behaviour in residential buildings. Hence, most researches on residential buildings are at low granularity, such as roughly profile energy consumption on regional level (Yu et al. 2010; Sadeghi et al. 2011; Mastrucci et al. 2014; Wang 2015; Lannon et al. 2050).

---

### 2.3 Opportunities for Further Works

As effective and useful techniques providing profound insights and possible strategic solutions in policy and management of building energy consumption, data-driven approaches have been deemed as favorable means for facilitate future in-depth studies on building energy performance. In this section, we tailor a few promising research directions of data-driven approaches applied in building energy.

The first direction suggested is to modify the framework of the current data-driven algorithms in the contexts of building performance to better fulfill the calculation of building energy

consumption. This necessitates special optimization of the data-driven approaches responding to unique requirements of prediction and classification of building energy consumption. It also calls for reliable and sufficient data source to yield training in high quality. Well-designed pretreatments particularly compatible with data collection in buildings and high-granularity measurement from smart meters are strongly recommended.

Another direction is to enrich the applications of the current data-driven versions, enabling them to be viable for building energy characterization at different scales and various climate conditions. In the literature, a great number of efforts in this direction are limited to small-scale building stocks. Future focuses should be placed on energy-use evaluation of a large amount of building population. To be specific, the next generation of data-driven approaches in building industries should cover more building thermal characteristics and climate conditions, and are able to perform energy mapping and benchmarking at the macro-scale.

Furthermore, it has been widely noted that most building energy consumption models at the present stage only offer short-term analyses. They fail to discuss the long-term profiles and predictions, nonetheless. Therefore, strong drive exists to extend data-driven approaches with applications to building energy consumption with climate changes in a large temporal scale. This will play an important role in future building retrofitting, where long-term energy consumption is a crucial contributing factor for new HPB design when used to replace old buildings.

As pointed out above, building performance evaluation nowadays strives for high efficiency in building energy consumption. However, other factors, such as indoor air quality, occupant thermal comfort, occupancy behaviour interaction and equipment energy-performance coefficient, are also equally important. This indicates an ideal data-driven model should make use of multiple indexes to provide a comprehensive analysis of building performance, instead of the current single output of energy consumption or heating/cooling loads. Significantly, apart from

the basic functions (i.e., prediction and classification) discussed in this review, future outlook of data-driven techniques targets decision-making machine, such as occupancy behaviour recommender and equipment operation instructor. These data-driven based developments in building industries would offer real-time on-site information for thermo-comfortable accommodation with minimum energy consumption.

---

## 2.4 Conclusion

Given that the growing concerns about natural resource shortages and eco-environment deterioration, today's building industry necessitates strongly innovative techniques to better optimize energy consumption in buildings. Data-driven approaches are one of the most effective advanced means to fulfill this goal in function of more accurate prediction and clearer classification.

This chapter reviewed the mainstream data-driven techniques currently applied in building energy studies. A comprehensive summary of each approach, including their basic thought, algorithm structure, strengths and weaknesses, was elaborated. In the category of data-driven approaches for prediction, ANNs gains a primary position in a large number of applications ranging from load forecasting, retrofit potential estimation. The challenges in this method are to choose an appropriate architecture and learning rate. As to SVM, the method has a simple training process requiring a few inputs whereas low calculation efficiency has become a roadblock in its application, especially for large-scale building energy analysis and high granularity. To shorten its calculation time, recent tremendous efforts have made to optimize SVM structure and develop hybrid models combing the method with clustering algorithms, DTW and et al. Statistical regression plays an important role in significant parameters evaluation, energy mapping and benchmarking at the urban scale. However, inaccuracy in short-term prediction and possible unforeseen correlations among the selected predictors greatly undermine the effectiveness of the regression models in practical applications. As to

DT, it is regarded as a feasible technique for energy consumption prediction and building energy benchmarking provided that these cases do not involve sequence and nonlinear data. The major concern is DT results are primarily based on expectations, and thus validity of many DT models being questioned. GA excels in coping with sophisticated data and can provide optimal and multi-objective solutions, but GA suffers from low computational efficiency and non-unique deliveries. As far as data-driven approaches for classification are concerned, K-means clustering, self-organizing map and hierarchy clustering were exemplified with focuses on pattern recognition of load curves, energy mapping, benchmarking and retrofit at macro-scale. The K-means clustering algorithm achieves highly accurate data partition in many applications. A fundamental deficiency in this classification approach is its framework lacks a priori to specify the number of clusters and centroids yet when the classifying features are unavailable. Self-organizing map can effectively reduce the dimensions of input data to a feature map, but its accuracy relies on a careful parameter selection and an appropriate design of its algorithmic structure. As to hierarchical clustering, it is usually implemented by a number of diverse merging and terminating ways, which could lead to multiple clustering results even subject to the same similarity criterion. This greatly influences its accuracy and effectiveness when performed for many clustering problems.

Based on the above analysis, this review points out several future directions for data-driven approaches to better underpin building energy analysis. They include (1) modifying the frameworks of different data-driven approaches in the context building performance features to more accurately respond specific demands from building energy calculation; (2) enriching applications of data-driven approaches to cover building energy uses at different scales under a wide spectrum of weather conditions; (3) extending data-driven approaches serving building energy consumption investigation and in both the short- and long-term; (4) integrating multiple target indices in the data-driven framework to



deliver a more balanced evaluation on building energy performance. It is believed that data-driven techniques with progress along these routes will offer more efficient and reliable support for energy management and optimization in future building industries.

## References

- Alhamazani K, Ranjan R, Mitra K et al (2015) An overview of the commercial cloud monitoring tools: research dimensions, design issues, and state-of-the-art. *Computing* 97(4):357–377
- Al-Homoud MS (2001) Computer-aided building energy analysis techniques. *Build Environ* 36(4):421–433
- Amjady N (2001) Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *IEEE Trans Power Syst* 16(3):498–505
- An N, Zhao W, Wang J et al (2013) Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting. *Energy* 49(1):279–288
- Arambula Lara R, Cappelletti F, Romagnoni P et al (2014) Selection of representative buildings through preliminary cluster analysis
- Asadi E, Silva MGD, Antunes CH et al (2014) Multi-objective optimization for building retrofit: A model using genetic algorithm and artificial neural network and an application. *Energy Build* 81(na):na
- Aydinalp M, Ugursal VI, Fung AS (2002) Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks. *Appl Energy* 71(2):87–110
- Aydinalp-Koksal M, Ugursal VI (2008) Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector. *Appl Energy* 85(4):271–296
- Azadeh A, Ghaderi SF, Tarverdi S et al (2007) Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. *Appl Math Comput* 186(2):1731–1741
- Barnaby CS, Spitler JD (2005) Development of the residential load factor method for heating and cooling load calculations. *ASHRAE Trans* 111:291–307
- Beyer HG (2000) Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice. *Comput Methods Appl Mech Eng* 186(2–4):239–267
- Bojić M, Lukić N (2000) Numerical evaluation of solar-energy use through passive heating of weekend houses in Yugoslavia. *Renew Energy* 20(2):207–222
- Canyurt OE, Ozturk HK, Hepbasli A et al (2005) Estimating the Turkish residential-commercial energy output based on genetic algorithm (GA) approaches. *Energy Policy* 33(8):1011–1019
- Caputo P, Costa G, Ferrari S (2013) A supporting method for defining energy strategies in the building sector at urban scale. *Energy Policy* 55(249):261–270
- Chung W, Hui YV, Lam YM (2005) Benchmarking the energy efficiency of commercial buildings. *Appl Energy* 83(1):1–14
- Dong B, Cao C, Lee SE (2005) Applying support vector machines to predict building energy consumption in tropical region. *Energy Build* 37(5):545–553
- Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern* 3(3):32–57
- Efron B, Tibshirani R (1993) An introduction to bootstrap. Chapman & Hall, New York
- Farahbakhsh H, Ugursal VI, Fung AS (1998) A residential end-use energy consumption model for Canada. *Int J Energy Res* 22(13):1133–1143
- Fonseca JA, Schlueter A (2015) Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. *Appl Energy* 142:247–265
- Goldberg DE (1986) The genetic algorithm approach: why, how, and what next? In: Adaptive and learning systems. Springer US, pp 247–253.
- Heidarinejad M, Dahlhausen M, McMahon S et al (2014) Cluster analysis of simulated energy use for LEED certified U.S. office buildings. *Energy Build* 85:86–97
- Hong T, Le Y, Hill D et al (2014) Data and analytics to inform energy retrofit of high performance buildings. *Appl Energy* 126(C):90–106
- Howard B, Parshall L, Thompson J et al (2011) Spatial distribution of urban building energy consumption by end use. *Energy Build* 45:141–151
- Huang YJ (2000) A bottom-up engineering estimate of the aggregate heating and cooling loads of the entire US building stock. Escholarship University of California
- ISO (2013) ISO Standard 12655: energy performance of buildings—presentation of real energy use of buildings
- Jones P, Patterson J, Lannon S (2007) Modelling the built environment at an urban scale—energy and health impacts in relation to housing. *Landsc Urban Plan* 83(1):39–49
- Juan YK et al (2009) GA-based decision support system for housing condition assessment and refurbishment strategies. *Autom Construct* 18(4):394–401
- Kalogirou SA (2001) Artificial neural networks in renewable energy systems applications: a review. *Renew Sustain Energy Rev* 5(4):373–401
- Kalogirou SA, Bojic M (2000) Artificial neural networks for the prediction of the energy consumption of a passive solar building. *Energy* 25(5):479–491
- Kang Z, Jin M, Spanos CJ (2014) Modeling of end-use energy profile: an appliance-data-driven stochastic approach. *Statistics* 5382–5388
- Kuo WJ, Chang RF, Chen DR et al (2001) Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast Cancer Res Treat* 66(1):51

- Lannon S, Georgakaki A, Macdonald S (2013) Modelling urban scale retrofit, pathways to 2050 low carbon residential building stock. *Ibpsa*
- Larivière I, Lafrance G (1999) Modelling the electricity consumption of cities: effect of urban density. *Energy Econ* 21(1):53–66
- Li K, Su H (2010) Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system. *Energy Build* 42(11):2070–2076
- Li Q, Meng Q, Cai J et al (2009a) Predicting hourly cooling load in the building: a comparison of support vector machine and different artificial neural networks. *Energy Convers Manage* 50(1):90–96
- Li Q, Meng Q, Cai J et al (2009b) Applying support vector machine to predict hourly cooling load in the building. *Appl Energy* 86(10):2249–2256
- Li Z, Han Y, Xu P (2014) Methods for benchmarking building energy consumption against its past or intended performance: an overview. *Appl Energy* 124(7):325–334
- Li Z, Huang G (2013) Re-evaluation of building cooling load prediction models for use in humid subtropical area. *Energy Build* 62(3):442–449
- Li X, Deng Y, Ding L et al (2010) Building cooling load forecasting using fuzzy support vector machine and fuzzy C-mean clustering[C]. In: International conference on computer and communication technologies in agriculture engineering (CCTAE 2010), vol 2010, pp 438–441
- Magoules F, Zhao HX (2016) Data mining and machine learning in building energy analysis. Wiley
- Mastrucci A, Baume O, Stazi F et al (2014) Estimating energy savings for the residential building stock of an entire city: a GIS-based statistical downscaling approach applied to Rotterdam. *Energy Build* 75(2):358–367
- Mathew PA, Dunn LN, Sohn MD et al (2015) Big-data for building energy performance: lessons from assembling a very large national database of building energy use. *Appl Energy* 140:85–93
- Mathieu JL, Price PN, Kiliccote S, Piette MA (2011) Quantifying changes in building electricity use, with application to demand response. *IEEE Trans Smart Grid* 2:507–518
- Mejri O, Barrio EPD, Ghrab-Morcós N (2011) Energy performance assessment of occupied buildings using model identification techniques. *Energy Build* 43(2):285–299
- Neto AH, Fiorelli FAS (2008) Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy Build* 40(12):2169–2176
- Nikolaïdis Y, Pilavachi PA, Chletsis A (2009) Economic evaluation of energy saving measures in a common type of Greek building. *Appl Energy* 86(12):2550–2559
- Nikolaou T, Kolokotsa D, Stavrakakis G (2011) Review on methodologies for energy benchmarking, rating and classification of buildings. *Adv Build Energy Res* 5(1):53–70
- Nikolaou TG, Kolokotsa DS, Stavrakakis GS et al (2012) On the application of clustering techniques for office buildings' energy and thermal comfort classification. *IEEE Trans Smart Grid* 3(4):2196–2210
- Olofsson T, Andersson S (2002) Overall heat loss coefficient and domestic energy gain factor for single-family buildings. *Build Environ* 37(11):1019–1026
- Panapakidis IP, Papadopoulos TA, Christoforidis GC et al (2014) Pattern recognition algorithms for electricity load curve analysis of buildings. *Energy Build* 73(2):137–145
- Park HS, Lee M, Kang H et al (2016) Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. *Appl Energy* 173:225–237
- Paudel S, Nguyen PH, Kling WL et al (2015) Support vector machine in prediction of building energy demand using pseudo dynamic approach. *Comput Sci*
- Pérez-Lombard L, Ortiz J, Pout C (2008) A review on buildings energy consumption information
- Perino M, Tardioli G, Kerrigan R et al (2015) Data driven approaches for prediction of building energy consumption at urban level. *Energy Proc* 78:3378–3383
- Quinlan JR (1986) Induction of decision trees machine learning. In: Data: goals and general description of the IN L.EN System, pp 257–264
- Sadeghi H, Zolfaghari M, Heydarizade M (2011) Estimation of electricity demand in residential sector using genetic algorithm approach
- Santamouris M, Mihalakakou G, Patargias P et al (2007) Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy Build* 39(1):45–51
- Setiawan A, Koprinska I, Agelidis VG (2009) Very short-term electricity load demand forecasting using support vector regression. In: International joint conference on neural networks, IJCNN 2009, Atlanta, Georgia, USA, 14–19 June. DBLP, pp 2888–2894
- Shimoda Y, Fujii T, Morikawa T et al (2004) Residential end-use energy simulation at city scale. *Build Environ* 39(8):959–967
- Sides J (2014) The victory lab: the secret science of winning campaigns. *Public Opin Q* 78(S1):363–364
- Sözen A, Arcaklioglu E (2007) Prediction of net energy consumption based on economic indicators (GNP and GDP) in Turkey. *Energy Policy* 35(10):4981–4992
- Swan LG, Ugursal VI (2009) Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. *Renew Sustain Energy Rev* 13(8):1819–1835
- Tiedemann KH (2007) Using conditional demand analysis to estimate residential energy use and energy savings. In: Proceedings of the Cdeee
- Tsekouras GJ, Hatzizargyriou ND, Dialynas EN (2007) Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Trans Power Syst* 22(3):1120–1128
- Tso GKF, Yau KKW (2007) Predicting electricity energy consumption: a comparison of regression analysis,

- decision tree and neural networks. *Energy* 32 (9):1761–1768
- UNEP (2013) Energy efficiency for buildings. <http://www.studiocollantin.eu/pdf/UNEP%20Info%20sheet%20-%20EE%20Buildings.pdf>
- Vapnik V, Golowich SE, Smola A (1996) Support Vector method for function approximation, regression estimation, and signal processing. *Adv Neural Inf Process Syst* 9:281–287
- Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map. *IEEE Trans Neural Networks* 11 (3):586
- Wang E (2015) Benchmarking whole-building energy performance with multi-criteria technique for order preference by similarity to ideal solution using a selective objective-weighting approach. *Appl Energy* 146:92–103
- Wautman B, Breesch H, Saelens D (2013) Evaluation of the accuracy of the implementation of dynamic effects in the quasi steady-state calculation method for school buildings. *Energy Build* 65(10):173–184
- Xiao H, Wei Q, Jiang Y (2012) The reality and statistical distribution of energy consumption in office buildings in China. *Energy Build* 50(50):259–265
- Yalcintas M (2006) An energy benchmarking model based on artificial neural network method with a case example for tropical climates. *Int J Energy Res* 30 (14):1158–1174
- Yalcintas M (2008) Energy-savings predictions for building-equipment retrofits. *Energy Build* 40 (12):2111–2120
- Yalcintas M, Ozturk UA (2006) An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database. *Int J Energy Res* 412–421
- Yamaguchi Y, Shimoda Y, Mizuno M (2007) Proposal of a modeling approach considering urban form for evaluation of city level energy management. *Energy Build* 39(5):580–592
- Yan CW, Yao J (2010) Application of ANN for the prediction of building energy consumption at different climate zones with HDD and CDD. In: International conference on future computer and communication. IEEE, pp V3-286–V3-289
- Yang J, Rivard H, Zmeureanu R (2005) On-line building energy prediction using adaptive artificial neural networks. *Energy Build* 37(12):1250–1259
- Yezioro A, Dong B, Leite F (2008) An applied artificial intelligence approach towards assessing building performance simulation tools. *Energy Build* 40(4):612–620
- Yokoyama R, Wakui T, Satake R (2009) Prediction of energy demands using neural network with model identification by global optimization. *Energy Convers Manage* 50(2):319–327
- Yu Z, Haghghat F, Fung BCM et al (2010) A decision tree method for building energy demand modeling. *Energy Build* 42(10):1637–1646
- Zhao HX, Magoulès F (2010) Parallel support vector machines applied to the prediction of multiple buildings energy consumption. *J Algorithms Comput Technol* 4(2):231–250
- Zhao HX, Magoulès F (2012) A review on the prediction of building energy consumption. *Renew Sustain Energy Rev* 16(6):3586–3592