Energy and Buildings 138 (2017) 240–256

Contents lists available at ScienceDirect

Energy and Buildings

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

A relevant data selection method for energy consumption prediction of low energy building based on support vector machine

Subodh Paudel [a,b,c], Mohamed Elmitri [b], Stéphane Couturier [b], Phuong H. Nguyen [c], René Kamphuis [c,1], Bruno Lacarrière [a], Olivier Le Corre [a,*]

[a] Department of Energy Systems and Environment, Ecole des Mines de Nantes, GEPEA, CNRS, UMR 6144, Nantes, France
[b] Veolia Recherche et Innovation (VERI), Limay, France
[c] Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 24 March 2016
Received in revised form
29 September 2016
Accepted 8 November 2016
Available online 21 December 2016

Keywords:
Building energy consumption
Prediction
Low energy building
Support vector machine
Online and offline learning

ABSTRACT

Low energy buildings (LEBs) are being considered as a promising solution for the built environment to satisfy high-energy efficiency standards. The technology is based on lowering the overall heat transmission coefficient value (U-value) of the buildings envelope and increasing a heat capacity thus creating a higher thermal inertia. However, LEB introduces a large time constant compared to conventional building due to which it slows the rate of heat transfer between interior of building and outdoor environment and alters the indoor climate regardless of sudden changes in climatic conditions. Therefore, it is challenging to estimate and predict thermal energy demand for such LEBs.

This work focuses on artificial intelligence (AI) model to predict energy consumption of LEB. Two kinds of AI modeling approaches: "all data" and "relevant data" are considered. The "all data" uses all available training data and "relevant data" uses a small representative day dataset and addresses the complexity of building non-linear dynamics by introducing past day climatic impacts behavior. This extraction is based on dynamic time warping pattern recognition methods. The case study consists of a French residential LEB. The numerical results showed that "relevant data" modeling approach that relies on small representative data selection has higher accuracy ($R^2 = 0.98$; RMSE = 3.4) than "all data" modeling approach ($R^2 = 0.93$; RMSE = 7.1) to predict heating energy load.

1. Introduction

The energy efficiency of buildings has drawn significant attention in the recent years as a tool to reach a reduction of energy consumption. According to the European Union (EU), the building sector represents 40% of the total energy consumption resulting in 36% of CO_2 gas emissions [1]. It is estimated that residential buildings account for 25% of the final energy consumption in the EU [1]. This energy consumption varies depending on the materials used in the walls and roofs of buildings affecting the heat transfer mechanisms. Low heat transfer materials in the building's envelope help to improve the energy efficiency resulting in low energy building (LEB) or passive house, by lowering the heat transfer coefficient (U-value).

In the case of a LEB, the high insulation level increases the importance of heat gains from lighting and solar radiation. Due to lower U-value materials in LEBs, it dampens the indoor temperature fluctuations throughout the day resulting in an equilibrium indoor climate. In addition to this, the lower U-value increases the thermal resistance of the building by resulting in a slower heat transfer between the walls and indoor, and introduces a large time constant. Because of a large time constant as well as large heat capacity in LEBs compared with conventional buildings, it retains thermal gain from past climatic changes. Therefore, an estimation and prediction of LEB's thermal energy demand is quite challenging.

There are various prediction models based on physical, semi-physical and data-driven methods available to estimate and predict the thermal energy demand for different energy standard buildings. Physical methods estimate the energy demand of a building from known parameters, i.e., detailed geometrical information and thermal properties of the building. Several physical building sim-

* Corresponding author.
E-mail addresses: Olivier.Lecorre@mines-nantes.fr, lecorre@mines-nantes.fr, ollecorre@gmail.com (O. Le Corre).
[1] R. Kamphuis works at TUe, a part of VSNU.

http://dx.doi.org/10.1016/j.enbuild.2016.11.009
0378-7788/

ulation tools such as TRNsys [2], ESP-r [3], EnergyPlus [4] etc. are available to the estimate energy demand of buildings. These types of methods are suitable for an early stage of building design and energy consumption estimation. In order to reduce model equations in physical methods, semi-physical methods, for example, response factor method, transfer function method, frequency analysis method and lumped electrical analogy, i.e., resistance and capacitance method exist today [5]. These semi-physical methods also require detailed understanding of building thermal dynamics but overcome the limitation of physical method due to less complexity in physical parameter.

In case of data-driven models, they rely on data from empirical thermal behavior of buildings or data acquired from dynamic thermal energy simulations under various environmental conditions. For such models, physical properties or thermal performance of buildings or systems are known, and then important parameters are identified by statistical analysis. Simple statistical and regression methods seem more feasible to estimate the energy demand, but they are not quite as accurate as to represent non-linear behavior of building dynamics particularly LEB. In recent years, machine learning based artificial intelligence (AI) models like artificial neural network [6–9] and support vector machine [6–8,10] are used to estimate energy demand. These AI methods have some drawbacks compared to physical methods. For instance, AI methods cannot be built for an early stage of new building, and hence they miss the physical understanding. However, these AI methods have several advantages over physical and semi-physical methods. Firstly, both physical and semi-physical methods are highly parameterized due to their interactions between systems on various modes of heat transfer requiring more input information. All the physical thermal properties of building are not always known hence impracticable for Energy Services Company (ESCOs) and/or Building Energy Management System (BEMS) for planning and control use. Secondly, they are good in learning the response of a building energy system. Finally, they have a strong capability of being adaptive to update the model parameter to take into account dynamic environment of future conditions.

Before implementing an AI-based method, there are more essential research questions: how to introduce such past climatic impacts behavior and how to consider the amount of data. Accordingly, machine learning based AI models can be built with two kinds of approaches: "**all data**" and "**relevant data**". The approach is defined "**all data**" if all the available data (measurement or empirical behavior of building data) are used for model training to determine the parameters of model. For such model, the parameters are fixed for the considered building independently of the prediction day and future environment conditions. On the other hand, the approach is defined "**relevant data**" if the pre-selection of data is done initially for model training based on prediction day conditions. For such model, the data used for model training are reduced based on the relevance, and parameters of model are changed for the considered building by each prediction day conditions. This type of approach is also named as **"few representative data**" since it selects small data to build a model. There are three major reasons to consider "**relevant data**" compared to "**all data**". Firstly, all data used for model training contain similarities and dissimilarities of input patterns behaviors and some of the information might be redundant. Secondly, a predictive model takes a lot of time for model training when all the data are used. Finally, with the adaptability of growing this model in the future, the newest environment and climatic conditions have probably more useful information, which is not considered in "**all data**" approach due to its computational complexities. The effect of this new information is neglected to update the model parameter. In order to update the model parameters in "**all data**" approach, the initial learning algorithm should be modified to complex learning algorithm. In this paper, learning mechanism in "**all data**" approach is called "**offline learning**" since model parameters are not updated with new datasets. On the contrary, the "**relevant data**" approach uses both "**offline**" and "**online" learning**: The "**offline learning**" selects few representative data from all fixed available data whereas "**online learning**" selects few representative data from all updated available data so that it updates the model parameters with the new dataset and adapt to changing environment.

For instance, "**relevant data**" approaches based on similar trends of prediction day climatic conditions were used by various authors [11–15] to select small representative data for electricity energy consumption. All these methods determine the similarity of selected individual variables based on the Euclidean distance between prediction day with training day. They are then further multiplied by weight factors of each selected variable to select representative day for model training. The weight factors of selected variables were determined using **least square method (LSM) based on regression model**. Paudel et al. [10] used outside air temperature of prediction day to select "**few representative data**" to predict heating load for an office building. Authors [16,17] used electrical energy load and climatic data of prediction day to select small representative data to build an **AI model**. For these authors as well, weight factors of selected variables were determined using **LSM based on regression model**. Heating degree-day (HDD) and cooling degree-day (CDD) measured in $^{\circ}$C/day were used by Roldan-Blay et al. [18] to select small representative data using estimated HDD and CDD of prediction day to predict electrical load. Various authors [19–28] used clustering/classification methods to select few representative data from particular clusters/classes based on estimated daily average energy load of prediction day and energy load of previous day from prediction day for model training. The detailed variables used to select small representative day data for model training are summarized in Table 1.

Although previously studied methods have advantages due to small representative data selection, there are still some limitations. First, the methods that focus on selection of "**few representative data**" do not consider past day climatic conditions due to large time constant of building (for example, more than 100 h in LEBs [29]) which is an essential factor for LEBs. Second, these methods do not consider the solar gain impact. Finally, the methods that are based on daily average energy load of prediction day or previous day to select representative data are not suitable for LEBs. If the learning mechanism of prediction model is not only for a day ahead but also for a longer period in advance, then the prediction models will rely on previously predicted daily average energy load values, and errors will be accumulated. Thus, it is not pragmatic for real applications. To bridge the aforementioned research gap, this paper aims to develop a prediction model from hours to couple of days (or even for longer periods depending upon the forecast range of climatic conditions) using few representative data for model training that include the influence of many past day climatic conditions applied to LEBs. The other objectives are to compare the developed methodology with existing methods, i.e., "**relevant data**" approach to select few representative data using **LSM based on regression** and "**all data**" approach. Consequently, this paper selected support vector machine (SVM) as machine learning **AI model** since it provides global optimal solutions and higher generalization performance compared to neural networks because of its non-linear problems solving by empirical risk minimization [30].

2. Methodology

The initial step of our methodology is the collection of building energy consumption, climatic variables, occupancy profile, and building operating conditions data (e.g., set-point temperature,