Sustainable Cities and Society 24 (2016) 33–41

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

Generation of accurate weather files using a hybrid machine learning methodology for design and analysis of sustainable and resilient buildings

Debaditya Chakraborty*, Hazem Elzarka, Raj Bhatnagar

College of Engineering and Applied Science, University of Cincinnati, P.O. Box 210071, Cincinnati, OH 45221-0071 USA

ARTICLE INFO

Article history:
Received 30 January 2016
Received in revised form 13 April 2016
Accepted 14 April 2016
Available online 16 April 2016

Keywords:
Weather prediction
Solar irradiance
Energy efficiency
Sustainability
Resilience
Support vector machine regression
Probability estimation

ABSTRACT

Accurate weather information is required to assess the performance of sustainable buildings. Currently such performance is assessed using Typical Meteorological Year (TMY) weather files. TMY3 is the latest TMY file introduced by Department of Energy (DOE) and it represents the historical 30-year average of weather data more closely than any other available datasets. However, TMY3 data reflects the past and do not accurately forecast weather information thus leading to erroneous estimation of performance and economic feasibility. Therefore, more accurate methods are necessary to generate weather files so as to design buildings for sustainability and resilience. This article proposes an effective hybrid modeling methodology based on support vector machine regression and probability estimation to predict long term future weather variables. The methodology has been validated by using a dataset containing historical solar irradiance data for 54 years (1961–2014) for Cincinnati, Ohio. By implementing the proposed technique multiple models were trained with different segments of the dataset, which were used to predict the hourly solar irradiance for 8, 9, 10 and 11 years respectively. Subsequently, the predicted data and the current TMY3 data was compared with the actual historical data. The average increase in accuracy of the predicted over TMY3 direct normal and diffuse horizontal components of solar irradiance was found to be 38.43% and 11.8%. These findings suggest that the long term accuracy of the proposed modeling technique is greater than the current state-of-the-art.

1. Introduction

A major environmental issue facing the world today is climate change, which is being caused by the increased emission of greenhouse gases and artificial aerosols. It is now widely recognized that the gases emitted from burning fossil fuels are causing climate change, which is capable of disrupting the ecosystem, human comfort and health (Roaf, Crichton, & Nicol, 2009). Buildings as part of the infrastructure must be capable of adapting to the changing climate so as to provide comfortable environments for human welfare. Nonetheless, presently design and analysis of building systems is performed by using weather data that reflects various annual climate averages based on historical weather data. Since buildings are typically designed for a long lifespan (20 years or more), building professionals must incorporate strategies that address future regional climate change. Weather data is used for a variety of design and operational decisions of buildings, which include but are not limited to the selection of systems such as, heating ventilation and air conditioning (HVAC), building materials, building orientation and configuration. Currently, the state-of-the-art building simulation tools used during the design process utilize historical climate data, with the intrinsic belief that the climate will remain moderately stable in the future (Larsen, 2011). Although these simulation tools accurately represent the physical models of buildings, they fail to account for the variation of weather parameters with time, thus making their results vulnerable to future changes in weather (Guan, 2009). Therefore, building professionals need tools, which utilize algorithms to accurately project weather information based on historical trends and present conditions.

The objective of this paper is to provide its readers with an effective method to more accurately predict long term future weather parameters using support vector machine regression and probability estimation. The method has been validated by developing models to predict an important weather parameter – solar

* Corresponding author.
E-mail address: chakrada@mail.uc.edu (D. Chakraborty).

http://dx.doi.org/10.1016/j.scs.2016.04.009
2210-6707/

irradiance. This paper first discusses the importance of solar irradiance prediction, then it briefly describes the support vector machine regression methodology. The paper then discusses the developed model in detail and ends by analyzing the predicted solar irradiance data obtained from the models and comparing it with data obtained from the TMY3 weather file for Cincinnati, Ohio.

2. Importance of solar irradiance prediction

The earth receives a vast amount of energy from the sun in the form of solar radiation. Therefore, a variety of energy efficient strategies can be implemented in buildings to harness the sun's energy, including:

- Solar electric (photovoltaic) to convert sunlight directly into electricity.
- Solar heat (thermal) to heat domestic hot water.
- Passive solar to light and heat buildings.

Solar resource data must be accurately forecasted so as to better optimize these systems and also to evaluate the return on investment, so that people can make informed decisions (NREL, 2016). However, present methods such as numerical prediction cannot forecast long term weather variables due to the complex nature of its partial differential equations, which is impossible to solve accurately using analytical methods (Richardson, 2007). Another practical approach is morphing present day weather with General Circulation Model (GCM) data in order to generate long-term weather files. GCM is based on sound physical processes in the atmosphere, ocean, cryosphere and land surface. However, a source of uncertainty in the GCM model data is that it depicts the climate over a coarse resolution. However, standard building energy simulation programs ideally requires weather data generated from climate model at a fine spatial and temporal resolution (Jentsch, James, Bourikas, & Bahaj, 2013). Therefore, a downscaling method called 'morphing' was proposed in 2005 to convert coarse resolution of GCM to fine spatial and temporal resolutions required for building energy simulation (Belcher, Hacker, & Powell, 2005). Morphing is based on shifting and stretching the climatic variables in a baseline 'present day' weather file such as TMY to produce new weather time series (Chan, 2011). But, previous research has shown that the morphing technique can cause under or over estimation of climate change impacts, resulting in unrealistically high or low values for morphed weather parameters (Jentsch, Bahaj, & James, 2008). In order to prevent unrealistic drift in the predicted results, TMY, WYEC, International Weather Year for Energy Calculation (IWEC), Test Reference Years (TRY) data sets are widely used in both academia and industry to design and evaluate energy performance of buildings. It has been shown that TMY3 characterizes long term solar irradiance data more closely to the actual irradiance data than any other weather dataset (Crawley & Huang, 1997). The TMY3 data set is composed of 12 typical meteorological months (January through December) that are essentially historical weather data concatenated to form one typical year weather data set. TMY3 data sets are widely used by building designers and others for modeling renewable energy conversion systems and long term building energy performance evaluation. However, TMY3 data do not provide a range of results that capture yearly variations due to changing weather, which is important for design analysis, building energy management, and for performing life cycle assessments of energy efficiency investments (Hong, Chang, & Lin, 2013). Therefore, to overcome the shortcomings of TMY3 an alternative methodology is proposed in this article.

3. Support vector machine regression (SVR)

Support vector machine (SVM) is a very useful technique for data classification. SVM in its present form was developed by Boser, Guyon, and Vapnik (1992) and has gained popularity due to many attractive features, and promising empirical performance. Although SVMs were initially developed for classification problems, it can also be used to solve regression problems by introducing an alternative ϵ insensitive loss function proposed by Smola and Schölkopf (2004). The main advantage of SVR over other regression methods is that it has a very strong theoretical background, which is motivated by statistical learning theory (Bennett & Campbell, 2000). The training of SVM involves optimization of a convex constrained cost function, which produces global and unique solutions as proved by Fletcher (1980). SVR was the choice for supervised learning algorithm for this research because of its ability to accurately forecast time series data when the underlying system processes are typically nonlinear, non-stationary and not defined a-priori (Sapankevych & Sankar, 2009). In addition, it has been shown previously that SVR performs better than neural network and genetic algorithm when the underlying relationship between the input features and their corresponding target values are complex (Dong, Cao, & Lee, 2005; Li, Meng, Cai, Yoshino, & Mochida, 2009; Olatomiwa et al., 2015).

3.1. Theory of SVR

Given a set of data points, $\{(x_1, y_1), \ldots, (x_l, y_l)\}$, such that $x_i \in R^n$ are feature vectors and $y_i \in R^1$ are the target values, where l is the total number of training samples, the goal is to find a linear function $f(x) = \langle w, x \rangle + b$ so that the predicted target values $\widehat{y_i}$ are within ϵ distance from the actual target values y_i (Fig. 1) and also as flat as possible (Smola & Schölkopf, 2004). This method is called ϵ-SVR and

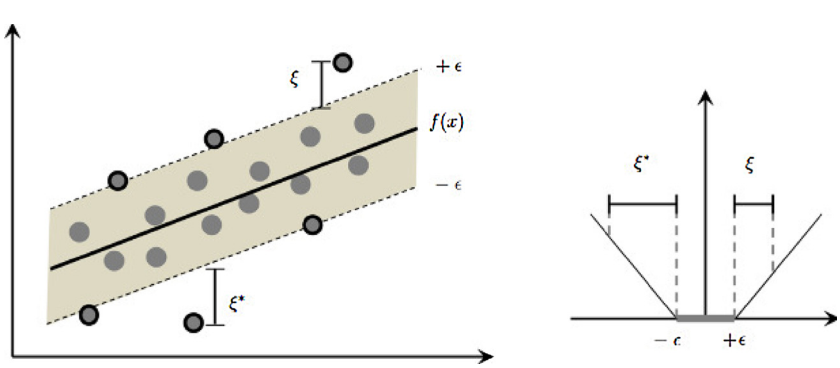

Fig. 1. SVR tube with radius ϵ fitted with data (Yu et al., 2013).