Applied Energy 181 (2016) 367–374

Contents lists available at ScienceDirect

Applied Energy

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

Prediction of diffuse solar irradiance using machine learning and multivariable regression

Siwei Lou [a,*], Danny H.W. Li [a], Joseph C. Lam [a], Wilco W.H. Chan [b]

[a] Building Energy Research Group, Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China
[b] School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

HIGHLIGHTS

- 54.9% of the annual global irradiance is composed by its diffuse part in HK.
- Hourly diffuse irradiance was predicted by accessible variables.
- The importance of variable in prediction was assessed by machine learning.
- Simple prediction equations were developed with the knowledge of variable importance.

ARTICLE INFO

Article history:
Received 13 May 2016
Received in revised form 28 July 2016
Accepted 16 August 2016

Keywords:
Solar energy
Diffuse irradiance
Boosted regression tree
Logistic regression

ABSTRACT

The paper studies the horizontal global, direct-beam and sky-diffuse solar irradiance data measured in Hong Kong from 2008 to 2013. A machine learning algorithm was employed to predict the horizontal sky-diffuse irradiance and conduct sensitivity analysis for the meteorological variables. Apart from the clearness index (horizontal global/extra atmospheric solar irradiance), we found that predictors including solar altitude, air temperature, cloud cover and visibility are also important in predicting the diffuse component. The mean absolute error (*MAE*) of the logistic regression using the aforementioned predictors was less than 21.5 W/m^2 and 30 W/m^2 for Hong Kong and Denver, USA, respectively. With the systematic recording of the five variables for more than 35 years, the proposed model would be appropriate to estimate of long-term diffuse solar radiation, study climate change and develope typical meteorological year in Hong Kong and places with similar climates.

1. Introduction

Solar irradiance data are crucial to the active solar energy facilities [1] and passive energy-efficient building designs [2]. The global irradiance on a horizontal plane consists of two differently featured components, namely the direct and diffuse irradiance. The two components are essential to estimate the solar irradiance in arbitrary surface directions [3,4], obstructed environments [5] and interior spaces [6] for building energy simulation, and may affect the photovoltaic system analysis [7,8]. The irradiance varies with latitude, seasons and time due to the different solar positions under unpredictable weather conditions [9]. Long-term data measurement is the most effective and accurate way of setting up databases [10,11]. However, the measurements of diffuse and direct irradiance are less straightforward than global component. In order to measure the diffuse irradiance, a shadow ring is usually adopted to obstruct the direct component but part of diffuse irradiance is blocked at the same time. Thus correction is required [12]. Sun trackers can be utilized to trace the solar position consistently but such apparatus can be costly. Therefore, the basic horizontal solar irradiance data are not always readily available in many parts of the world [13]. In Hong Kong, only hourly global solar irradiance on a horizontal plane (E_{HG} W/m^2) was systematically recorded by the local meteorological station called Hong Kong Observatory (HKO) for a time period of more than 35 years while the hourly diffuse and direct irradiance were not measured until Aug. 2008.

In absence of hourly horizontal diffuse irradiance (E_{HD} W/m^2), there are a number of prediction models of different complexities from simple empirical models [14–17], semi-empirical models [18–21] to rigorous radiative transfer models [22]. The radiative transfer models and semi-empirical models have the advantage of high accuracy. However, they tend to rely on more input variables and are usually more complex than empirical models. It seems that they are not readily applicable for the study of long

* Corresponding author.
E-mail address: swlou2-c@my.cityu.edu.hk (S. Lou).

http://dx.doi.org/10.1016/j.apenergy.2016.08.093
0306-2619/

term weather conditions since the required variables such as aerosol and ozone amount may be unavailable. In terms of the empirical models, the most common approach is to correlate the diffuse fraction (K) given in Eq. (1) with other readily available variables in simple regression equations.

$$K = E_{HD}/E_{HG} \tag{1}$$

The diffuse fraction depicts the level of diffuse component (E_{HD}) with respect to the global solar irradiance (E_{HG}) of the ground level. The most frequently used variable for correlation is clearness index (K_t) as Eq. (2), which displays the percentage of solar irradiance neither being absorbed nor reflected by atmosphere.

$$K_t = E_{HG}/E_{HE} \tag{2}$$

where E_{HE} is the hourly extra atmospheric solar irradiance on a horizontal plane (W/m^2). Other predictors depict weather conditions that may also contribute to the correlations. In 2001, Boland et al. [23] proposed a model based on logistic equation for estimating K from K_t. The model was modified in 2007 [24] and 2010 [25] to employ other variables and improve the prediction ability [26]. There are, however, a number of available meteorological variables which have potential to be well correlated with K. In order to improve the performance while keep the simplicity of the equation, it is expected to quantify the relevance between K and the available meteorological variables.

An alternative approach is to use the technique of machine learning (ML) that extracts knowledge from the available database and estimates the relationship between K and other readily accessible variables [27–30]. A limitation of many ML algorithms is that the models being established are restricted in interpretability [31], which have limited physical implications to the problem of interest. However, the model developed by the algorithm of Boosted Regression Tree (BRT) can be employed to identify the contribution of each input variable in the prediction of output [32]. The method was used for the analysis of many ecological [33–35] and environmental [36,37] problems that were featured with large dataset, great number of relevant variables and complex interrelationship between input and output variables. This paper analyses the solar irradiance data recorded by the HKO from Aug. 2008 to Dec. 2013. The importance of meteorological variables in predicting K were analyzed by the algorithm of BRT and logistic regressions [38] were developed to predict K using the important variables. The findings and building design implications are discussed.

2. Data collection and solar irradiance in Hong Kong

Hong Kong (HK) is located in south coast of China (22.3°N, 114.3°E) with a subtropical climate. The summer (May to Sep.) is hot and humid while the winter (Dec. to Feb.) is dry and cool. In Aug. 2008, HKO established a station to measure hourly E_{HG}, E_{HD} and direct normal irradiance (E_{NB} W/m^2) in an outlying island. The data collection starts just before sunrise and ends after sunset every day. The hourly measurements in local civil time between Aug. 2008 and Dec. 2013 were applied in the study.

Fig. 1 presents the monthly averaged daily direct and diffuse solar radiation energy on a horizontal plane (W h/m^2). It can be seen that the daily diffuse radiation varies from 1469 W h/m^2 in December to 2553 W h/m^2 in June while daily direct radiation ranges between 778 W h/m^2 in March and 2991 W h/m^2 in July. The average global, direct and diffuse radiation energy are 3773, 2070 and 1703 W h/m^2, respectively. The maximum daily global solar radiation observed in July is mainly due to the high solar altitude and long day-length in summer. The below average daily solar radiation from Nov. to Apr. are resulted from the low solar altitude in winter and unstable weather conditions in spring. It is noted that the months receive low global solar radiation containing a high percentage of its diffuse component. Generally, the variation in diffuse solar radiation is comparatively less than that in direct component. The annual averages of horizontal diffuse and direct are around 54.9% and 45.1%, respectively, which implies the importance of diffuse solar radiation in HK. The findings are in good agreement with our previous work [39,40].

Hourly data in solar irradiance would be more appropriate for examining cooling load. Fig. 2 plots the monthly-average-hourly E_{HG} and E_{HD} in February and July representing respectively the lowest and largest recorded monthly-average-daily E_{HG}. The peak solar irradiance appears at solar noon for both E_{HG} and E_{HD}. The E_{HG} of just over 700 W/m^2 and approximately 400 W/m^2 are observed at noon in July and Feb., respectively. The E_{HD} values are quite close for these two months.

To eliminate spurious data and erroneous measurements, stringent quality-controls tests based on the CIE guidance were adopted [41]. Table 1 summarizes the criteria and number of data for each test. Over 2400 groups of data (observations) near sunrise and sunset with a solar altitude of less than 4° or E_{HG} of less than 20 W/m^2 were excluded. Totally, 326 observations were rejected under Level 1 and 2 tests. After the tests, 21,515 hourly solar irradiance sets (i.e. E_{HG}, E_{HD} and E_{NB}) were retained for the subsequent analysis.

Table 2 lists the predictors whose importance to K prediction were analyzed by the machine learning algorithm of Boosted Regression Tree (BRT). Sine function of solar altitude (μ) is used because the variable implicates the contribution of normal irradiance to the horizontal plane in a magnitude from 0 to 1. ΔK_t is the stability index developed by Perez et al. [42], which differentiates the stable (e.g. hazy) and unstable (e.g. partly cloudy) sky conditions. The unit of cloud amount (*cld*) was converted from Okta to percentile where 0 Okta and 8 Okta correspond to 0% and 100%, respectively. All the variables were recorded in civil local time. A further check of the variables in Table 2 removes 45 observations due to either sky obstructed from view (for *cld*) or the missing record of wind speed. Ultimately, 21,470 datasets were adopted.

Fig. 3 shows the correlation between K and K_t. It can be found that K remains nearly constant (between 0.9 and unity) up to about K_t = 0.25. It indicates that E_{HD} is the predominant component of E_{HG} when there is little solar irradiance received on the ground surfaces. When K_t is increased, a trend close to the linear decreasing of K can be identified, though the data get very scattered. The linear correlation implies the strong interrelationship between K_t and K while the scattering infers that K relies on more predictors other than K_t. There are few observations with K_t greater than 0.8. Similar trends were also observed by Boland et al. and Ridley et al., who studied the hourly solar irradiance in Adelaide, Australia and Bracknell, UK [25].

3. Boosted Regression Tree (BRT)

The Boosted Regression Tree (BRT) algorithm was used to generate an ensemble of Regression Trees (RT) that correlates hourly diffuse fraction (K) with the relevant variables (predictors) based on the 21,740 observations. As given in Fig. 4, the model is composed by a number of sequentially introduced RTs [43]. The first RT is developed to predict the value of K while the subsequent trees in step n are evolved targeting at the residuals (R_{n-1}) of the previously established RT ensemble in step $n-1$. As such, the ensemble of RTs is trained to focus on the variation of K that are not yet modelled by the existing tree ensemble in every step.

Each RT [44] categorizes observations with a number of predictors (Pre_1, Pre_2, ...) into various groups by a sequence of binary partitions (i.e. splits). Fig. 5 illustrates a split that separates a single group of N_i observations into two child groups according to the