Energy Conversion and Management 96 (2015) 277–286

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Generation of daily global solar irradiation with support vector machines for regression

F. Antonanzas-Torres *, R. Urraca, J. Antonanzas, J. Fernandez-Ceniceros, F.J. Martinez-de-Pison

EDMANS Group, Department of Mechanical Engineering, University of La Rioja, Logroño, Spain

ARTICLE INFO

Article history:
Received 10 December 2014
Accepted 27 February 2015
Available online 13 March 2015

Keywords:
Solar resource estimation
Global horizontal irradiation
Solar energy
Soft computing

ABSTRACT

Solar global irradiation is barely recorded in isolated rural areas around the world. Traditionally, solar resource estimation has been performed using parametric-empirical models based on the relationship of solar irradiation with other atmospheric and commonly measured variables, such as temperatures, rainfall, and sunshine duration, achieving a relatively high level of certainty. Considerable improvement in soft-computing techniques, which have been applied extensively in many research fields, has lead to improvements in solar global irradiation modeling, although most of these techniques lack spatial generalization.

This new methodology proposes support vector machines for regression with optimized variable selection via genetic algorithms to generate non-locally dependent and accurate models. A case of study in Spain has demonstrated the value of this methodology. It achieved a striking reduction in the mean absolute error (MAE) – 41.4% and 19.9% – as compared to classic parametric models; Bristow & Campbell and Antonanzas-Torres et al., respectively.

1. Introduction

Solar photovoltaic energy has experienced enormous growth in recent years due to its mass scale economy and subsequent cost reduction, which has rendered its price competitive in many electricity markets. The globally installed capacity reaches 138.9 GW [1] and this power is expected to continue increasing. Accurate solar resource assessment is critical to the proper development of solar technologies [2], as it mitigates uncertainties in these investments.

Solar global irradiation has traditionally been estimated from other related and commonly measured variables, with relatively simple parametric models, generally, calibrated onsite, aiming to parameterize the atmospheric transmittance and relate it to the extraterrestrial irradiation. Extraterrestrial irradiation accounts for the stationary component of solar irradiation, which is only dependent on solar geometry, being the atmospheric transmittance the stochastic component. Angstrom first proved the existence of a linear relationship between sunshine duration and extraterrestrial irradiation and daily global irradiation [3]. Many other approaches consisted of the daily range of temperatures [4–8] or the daily range of temperatures and rainfall [8–11]. The daily range of temperatures is associated with cloud cover and cleanliness of the atmosphere. Thus, high daily ranges of temperatures are typical of sunny, predominantly clear-sky days. Another different alternative resulted from the cloud cover measurement [12]. For a more detailed description of parametric models, the authors refer to their previous study [13]. This research concluded that some of the drawbacks of parametric models were the complexity of variable selection for model tuning and high mean absolute errors, ranging between 2.2–3.3 MJ/m^2 day.

Solar irradiation can also be estimated from satellite images and clear sky models. Basically, a satellite image collates the upwelling radiance from the Earth. This radiance varies depending on ground albedo and atmospheric transmittance, from clear sky periods to completely overcast, providing direct information about cloudiness and clear skies, throughout the cloud and clear-sky indexes [14,15]. The cloud index is computed from the reflectivity recorded outside the atmosphere, normalized with the range between the darkest pixel (corresponding to clearest sky conditions) and the brightest value (corresponding to the most overcast conditions). The clear sky index is calculated from the relationship between global horizontal irradiance and the clear sky global horizontal irradiance. These indexes are used to attenuate irradiance obtained via clear sky transmittance models, which are generally based on aerosol and precipitable water vapor content in the atmosphere [16,17].

* Corresponding author.
E-mail address: antonanzas.fernando@gmail.com (F. Antonanzas-Torres).

http://dx.doi.org/10.1016/j.enconman.2015.02.086
0196-8904/© 2015 Elsevier Ltd. All rights reserved.

The fact that satellite derived solar irradiation deviation is associated with the spatial resolution of sensors (image's pixel size) and that this resolution generally falls within the range of kilometers implies a high level of uncertainty [18,19]. Other sources of uncertainty are found in the inherent uncertainty of aerosol and water vapor estimations [14,20], which are normally estimated for very low spatial resolutions in the range of the Multi-angle Imaging SpectroRadiometer – MISR 0.5 × 0.5° [21] and the Moderate Resolution Imaging Spectroradiometer – MODIS 1 × 1° [22]).

In addition, solar irradiation can be estimated with soft-computing techniques using different atmospheric transmittance parameterizations and techniques. Artificial neural networks (ANN) have been widely employed to estimate solar irradiation using different sets of inputs depending on climatic criteria: in Turkey [23], Brazil [24], Saudi Arabia [25] or Spain [26], among others. Bayesian neural networks were also found to be useful when trained with maximum and minimum air temperatures [27]. Other techniques such as fuzzy genetic (FG) and adaptive neuro fuzzy inference systems (ANFIS) have been applied using spatial information (latitude, longitude and elevation) as inputs for models to account for spatial dependence [28]. Eventually, support vector machines for regression (SVR) began to be used to estimate solar irradiation from sunshine duration [29] and air temperatures [30] in China, detecting a remarkable spatial influence induced by elevation and temperature differences between training and testing sites.

The main drawbacks of these soft-computing techniques are the high computational costs, the complexity in variable selection and the low capacity of generalization if over-fitted, rendering them extremely locally dependent. In this study, the authors propose a new methodology to simplify the estimation of solar irradiation with support vector machines for regression with a wrapper-based scheme for input selection to obtain non-locally dependent models. This methodology was proven useful for developing a general (non-locally dependent) model for solar irradiation estimation, which was implemented in a case of study in Spain, under different climates and on diverse terrain. The results are compared with the classic parametric models [5,13].

2. Methodology

This study aims to develop a methodology capable of generating spatially general solar irradiation models, using data from different locations. To this end, SVR was the predictive technique chosen (see Section 2.1). In order to improve on the quality of the predictions, model optimization parameter (MPO) of SVR and feature selection (FS) were performed simultaneously using genetic algorithm (GA), as an evolution-based optimization algorithm, as detailed in Section 2.2.

This methodology was also applied to locally-trained models, i.e. models trained with data from a specific location, in order to quantify differences between local and general prediction models. Furthermore, some classical parametric techniques (Section 2.3) are included in the analysis as a benchmark for comparison with the proposed methodology.

2.1. Support vector regression

Support vector machines (SVM) were originally developed by [31] for classification problems. The popularity of this technique rapidly increased due to its ability to deal with non-linear data whilst maintaining satisfactory generalization ability and avoiding overfitting during the training process. The regression variant of SVM, also known as support vector regression, was later introduced by [32] who proposed the ε-intensive loss function (ε-SVR). In the present methodology, ε-SVR is applied and it is hereafter described.

SVR can be more easily understood by first assuming linear data. Here, the general equation for a linear regression model is as follows:

$$f(x) = \langle w, x \rangle + b \tag{1}$$

where x is the set of input patterns, w the unkown weight vector, $\langle w, x \rangle$ is the dot or inner product between w and x and b a threshold value. Traditional models, such as multiple linear regression, compute the weight vector based on the reduction of quadratic errors. On the contrary, ε-SVR are based on optimizing the absolute error. The initial goal of ε-SVR is to develop a function where all errors lie under a predefined value ε but with the best generalization capacity possible (generally related to model flatness). These two conditions are imposed as follows:

$$\begin{aligned} &\textit{minimize} \quad \frac{1}{2}||w||^2 \\ &\textit{subject to} \quad \begin{cases} y_i - (\langle w, x_i \rangle + b) \leqslant \varepsilon \\ (\langle w, x_i \rangle + b) - y_i \leqslant \varepsilon \end{cases} \end{aligned} \tag{2}$$

A flat model is obtained by minimizing the norm of the weight vector $||w||$. Moreover, the constraints of Eq. (2) guarantee that every error is lower than ε. Nevertheless, this formulation assumes that a solution for the optimization problem exits, which is not always true. In order to overcome this problem the condition imposed in Eq. (2) is relaxed and samples with errors higher than ε are admitted:

$$\begin{aligned} &\textit{minimize} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ &\textit{subject to} \quad \begin{cases} y_i - (\langle w, x_i \rangle + b) \leqslant \varepsilon + \xi_i \\ (\langle w, x_i \rangle + b) - y_i \leqslant \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geqslant 0 \end{cases} \end{aligned} \tag{3}$$

where ξ_i and ξ_i^* are the slack variables. A second term is included to measure the amount of loss via the slack variables. Here is where the foundation under the ε-intensive loss function $\xi_{|\varepsilon|}$ lies:

$$\xi_{|\varepsilon|} = \begin{cases} 0 & \text{if } |y_i - \hat{y}_i| < \varepsilon \\ |y_i - \hat{y}_i| - \varepsilon & \text{otherwise} \end{cases} \tag{4}$$

where y_i and $\hat{y}_i$ are the measured and predicted outcome, while ε is a parameter defined by the user. Points inside the ε-intensive region have null slack variables ($\xi_i = 0$ and $\xi_i^* = 0$), while points out of this region have either ($\xi_i > 0$ and $\xi_i^* = 0$) or ($\xi_i = 0$ and $\xi_i^* > 0$), as slack variables are constrained to be non-negative. Therefore, SVR tuning is influenced solely by points out of the ε-intensive region, also known as support vectors.

The trade-off between the two terms of Eq. (4) is controlled by the regularization parameter C, also referred to as cost. For low C values, the first term dominates the equation. A flat general model is then obtained but at the expense of under-training the model. On the contrary, for high C values, the second term dominates. The training error is then reduced, but at the same time, a risk of overfitting appears.

Standard dual optimization through Lagrange multipliers is used to solve the optimization problem of Eq. (3). Once the Lagrangian is computed, several transformations are conducted until the following expression is then obtained:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x_i, x \rangle + b \tag{5}$$

where α_i and α_i^* are Lagrange multipliers. A unique solution to this optimization problem can be obtained via quadratic programming (QP) techniques.