Energy Conversion and Management 100 (2015) 380–390

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Solar irradiation mapping with exogenous data from support vector regression machines estimations

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

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

A R T I C L E I N F O

Article history:
Received 27 March 2015
Accepted 8 May 2015
Available online 22 May 2015

Keywords:
Solar resource estimation
Global horizontal irradiation
Solar mapping
Ordinary kriging
Support vector regression

A B S T R A C T

Exactly how to estimate solar resources in areas without pyranometers is of great concern for solar energy planners and developers. This study addresses the mapping of daily global irradiation by combining geostatistical interpolation techniques with support vector regression machines. The support vector regression machines training process incorporated commonly measured meteorological variables (temperatures, rainfall, humidity and wind speed) to estimate solar irradiation and was performed with data of 35 pyranometers over continental Spain. Genetic algorithms were used to simultaneously perform feature selection and model parameter optimization in the calibration process. The model was then used to estimate solar irradiation in a massive set of exogenous stations, 365 sites without irradiation sensors, so as to overcome the lack of pyranometers. Then, different spatial techniques for interpolation, fed with both measured and estimated irradiation values, were evaluated and compared, which led to the conclusion that ordinary kriging demonstrated the best performance. Training and interpolation mean absolute errors were as low as 1.81 MJ/m^2 day and 1.74 MJ/m^2 day, respectively. Errors improved significantly as compared to interpolation without exogenous stations and others referred in the bibliography for the same region.

This study presents an innovative methodology for estimating solar irradiation, which is especially promising since it may be implemented broadly across other regions and countries under similar circumstances.

1. Introduction

Nowadays, solar irradiation research is in high demand thanks to its numerous applications-such as solar energy planning, agriculture and climate monitoring, among others. Solar irradiation is rarely measured in meteorological stations and if it is, the historical registers generally do not date back far enough to be of use for the afore-mentioned applications. As a result, during recent years different techniques have been proposed for solar irradiation estimation such as parametric models, in which atmospheric transmittance is correlated with other commonly measured meteorological variables. Angstrom [1] pioneered establishing the relationship between solar irradiation and sunshine duration. Later, Hargreaves [2] introduced the use of temperatures as an explanatory variable, which was further developed by Bristow and Campbell [3] considering maximum and minimum temperatures. Rainfall, together with temperatures, was also considered to model solar irradiation [4]. The combination of temperatures and rainfall yielded the best results in the evaluation of different models assessed in Australia [5]. A review of 24 of these parametric models concluded that the daily range of temperatures together with the rainfall achieved the best performing models in Spain [6].

Additionally, solar irradiance has also been estimated from satellite images and clear sky models. Different algorithms have been proposed to obtain cloud content and cleanness of atmosphere from upwelling brightness recorded in on-board satellite sensors [7]. Clear sky models physically describe downwelling solar irradiation and parametrize the atmosphere with aerosol content, via the Linke turbidity factor or the aerosol optical depth at different wavelengths, water vapor content and other gases content [8]. Another approach to estimate solar irradiation is found in soft-computing techniques, which usually consider commonly measured variables related to irradiation as inputs. Artificial neural networks (ANN) and fuzzy logic models using different sets of input variables have been extensively applied in solar irradiation modeling. Recently, support vector machines (SVM) were introduced for solar irradiation estimation using sunshine duration [9]. Chen et al. [10] analyzed the accuracy of estimating solar irradiation with a SVM model trained in one site and estimating at a

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

http://dx.doi.org/10.1016/j.enconman.2015.05.028
0196-8904/

different one, showing that estimations were mainly affected by the distance, temperature and altitude differences. Antonanzas-Torres et al. [11] developed an automatic procedure for training models and selecting the most relevant meteorological variables to be used in SVM. A comparative assessment between ANN, SVM and extreme learning machines was performed for the estimation of daily solar irradiation in Spain [12]. In this study, SVM outperformed the other techniques.

Nevertheless, evaluating the potential of solar energy technologies and establishing energy policies (i.e. via feed-in tariffs) not only requires estimation in specific sites, but preferably the development of solar irradiation maps as well. Most of these maps are based on satellite-derived estimates due to the wide spatial and temporal coverage of satellites. However, in many cases their spatial resolution is in the range of kilometers (due to satellite image resolution), which is inadvisable in those regions under micro-climatic conditions or with complex terrain [13]. Furthermore, these models require accurate estimations of aerosols, water vapor and other gases content which are not available in high resolution for extensive areas of the planet (only available in the range of 0.5–1°). Subsequently, these solar maps need to be correctly validated with on-ground ancillary measurements with pyranometers [14]. Different approaches have been proposed to develop solar irradiation maps using satellite-derived irradiation estimates and digital elevation models [15]. Geostatistical techniques have also proven useful in irradiation mapping when many pyranometers are available, such as the 3-D inverse distance model [16] and universal kriging using pyranometers and a low spatial resolution layer of satellite-derived solar irradiation [17]. Moreno et al. [18] generated maps of air temperatures and rainfall with ordinary kriging from a set of stations. These estimates were then utilized as inputs of an ANN to spatially estimate solar irradiation.

This article proposes an innovative method based on a support vector regression model trained and validated in a reduced set of meteorological stations with pyranometers using double optimization with genetic algorithms for variable selection and fine tuning. Afterwards, this model was used to estimate solar irradiation in a massive set of meteorological stations without pyranometers, but rather with records of other variables related to solar irradiation. In the third step, several spatial methods for interpolation were applied to map solar irradiation: *inverse distance weighting* (*IDW*), *ordinary kriging* (*OK*) and *universal kriging* (*UK*), to evaluate the accuracy of each approach.

This method proved to be useful in Spain using 400 meteorological stations (35 of them with pyranometer measurements). Ordinary kriging with support vector regression showed a striking mean absolute error of 1.74 MJ/m^2 day and an annual average error as low as 2.69%. This method promises high applicability in regions with a limited set of pyranometers, but with a dense network of basic meteorological stations, which is indeed the situation in many countries.

2. Method

The method aims to generate a map of solar irradiation by first increasing the density of solar irradiation values in the area of study with a general predictive model. Then, both the values measured by pyranometers and by the predictive model estimations are used to develop a continuous map utilizing different classical geostatistical techniques. In short, the method consists of three main steps.

- Train a general predictive model in stations where on-ground measurements of global solar irradiation are available. In this case, SVR, a well-known machine learning technique, is selected. In addition, Genetic Algorithm (GA) (Section 2.2) is used to perform model optimization parameter (MPO) and feature selection (FS), which should improve the quality of predictions.
- Estimate global solar irradiation in locations where records from different meteorological variables excluding solar irradiation are available.
- Develop a continuous map of global solar irradiation by using geostatistical interpolation techniques. The techniques tested are *IDW* (Section 2.5), *OK* (Section 2.3) and *UK* (Section 2.4). Latitude and elevation are the variables tested as external drift for the *UK*.

2.1. Support vector regression

Support vector regression (SVR) machines are a supervised modeling technique based on the concepts originally introduced by Vapnik [19]. Later, it was further developed for classification purposes [20]. The regression variant was firstly applied with the release of the ε-intensive loss function (ε-SVR) [21]. In this study, this ε-SVR was applied.

SVR is characterized by its strong generalization capability, its robustness against extreme data and its capacity to deal with non-linear problems thanks to the kernel trick. Nonetheless, the linear version of SVR is described first to facilitate understanding of the basic theory. The general equation in a multiple linear regression (MLR) problem is:

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

where x is the set of inputs, w the weight vector, $\langle w, x \rangle$ is the dot product between w and x, and b the bias. Traditionally, w is computed by means of least squares method, i.e. minimizing the sum of the squares of residuals. This method boosts the influence of large errors producing non-robust models that are highly sensitive to outliers. SVR copes with these issues by employing an alternative loss function comprised by two terms:

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

- The first term accounts for the accuracy of the model. The absolute error is used to mitigate the influence of outliers. Additionally, only errors beyond a threshold value ε are taken into account (errors within the *tube* or *funnel*). Small errors (absolute error lower than ε) are considered null, while the absolute error of poorly predicted samples is linearly considered via the so-called slack variables ξ_i and ξ_i.
- The second term is the penalty that controls the complexity or *flatness* of the model.

Cost parameter C handles the trade-off between errors in the predictions (first term) and complexity (second term).

Eq. (2) is solved via standard dual optimization through Lagrange Multipliers. After carrying out several mathematical transformations, the following expression is obtained:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x_i, x \rangle + b \tag{3}$$

where α_i and α_i^* are Lagrange multipliers. According to this equation, training samples are required for new predictions. However, only