Solar Energy 135 (2016) 682–691

Contents lists available at ScienceDirect

Solar Energy

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

Modeling and forecasting hourly global solar radiation using clustering and classification techniques

Pedro F. Jiménez-Pérez, Llanos Mora-López *

Departamento de Lenguajes y Ciencias de la Computación, ETSI Informática, Universidad de Málaga, Campus de Teatinos, 29071 Málaga, Spain

ARTICLE INFO

Article history:
Received 8 October 2015
Received in revised form 7 June 2016
Accepted 16 June 2016

Keywords:
k-means
Clearness index
Forecasting and modeling hourly global solar radiation
Data mining

ABSTRACT

A new system to forecast and model hourly global solar radiation is proposed. The system works in two different phases and applies different data mining techniques in each phase. A clustering algorithm to identify the type of days is proposed. The use of decision trees, artificial neural networks and support vector machines to estimate the parameters that characterize each type of day is also advanced. Two procedures are put forward to analyze data and estimate models. The proposed procedures have been validated using data recorded in Malaga, Spain. Two different input data sets were used and the corresponding errors for each case are presented. The results show that it is possible to predict next-day hourly values of solar radiation values with an *rMAE* of 15.2% for one of the input data sets; while the *rMAE* is 16.7% for the other input parameter set.

1. Introduction

The significant increase in recent years in the number of grid-connected photovoltaic systems raises one important issue, to establish how much energy these systems will produce. In the case of large facilities, produced energy forecasting is necessary to ensure correct integration into the electrical system (Widén et al., 2009) and it can help to achieve higher percentages of self-consumption (for instantaneous net metering) and to ensure electricity consumption is managed efficiently for residential and small commercial installations.

On the one hand, in the case of larger facilities, some electricity markets require energy producers to predict their hourly production for the following day. Moreover, if that prediction is wrong, some regulations such as the Spanish set a financial penalty which depends on the prediction error (Red Elétrica de España, 2016). On the other hand, intelligent power management of residential installations based on knowledge of daily consumption profiles and the prediction of the energy produced by the photovoltaic system can help to make these systems more profitable (Luthander et al., 2015).

Moreover, as regards performance evaluations of small photovoltaic facilities, where there is an insufficient number of monitoring stations, it would be also desirable to know how much energy they should have produced taking into account the solar irradiance they have received to closely manage them. The production of photovoltaic energy facilities mainly depends on the availability of solar irradiance that is variable and intermittent. Solar global irradiance is not a deterministic variable due to the meteorological conditions. Cloud presence is the most important of these factors in terms of the attenuation of solar irradiance. The problem is that the cloud attenuation process is highly stochastic in nature and it is difficult to predict how it will affect solar irradiance. The attenuation process is nonlinear, complex, dynamic and widely scattered due to the influence of the physical phenomena involved and to their variability in space and time.

Initial approaches to estimate solar irradiance would be the numerical weather prediction models, using current weather conditions as input for mathematical models of the atmosphere to predict the weather (Lorenz et al., 2008; Remund et al., 2008). The Mesoscale models, based on equations using the physics and dynamics of the atmosphere (Heinemann et al., 2006; Bacher et al., 2009) are another option. Alternatively, models based on satellite images or visible photographs of the sky that incorporate information of the current atmospheric state have been also used to forecast time horizons from a few minutes to hours (Hammer et al., 2001; Mayer et al., 2008).

Several statistical and probabilistic models have also been developed for solar irradiance forecasting tasks. In particular, ARMA (autoregressive moving average) and ARIMA (autoregressive integrated moving average) models have been widely used; for instance Aguiar et al. (1988); Aguiar and Collares-Pereira (1992)

* Corresponding author.
E-mail addresses: pjimenez@lcc.uma.es (P.F. Jiménez-Pérez), llanos@lcc.uma.es (L. Mora-López).

http://dx.doi.org/10.1016/j.solener.2016.06.039
0038-092X/

Nomenclature

α	solar elevation	I_{sc}	solar constant
δ	declination	$k^*_{h,d}$	daily-detrended hourly clearness index
γ	parameter for the kernel function in SVM	k_d	daily clearness index
$\hat{A}$	estimated value of A	k_h	hourly clearness index
ω_0	constant that determines the activation threshold	k_t	clearness index for period t
ω_h	hour angle	MAE	mean absolute error
ω_i	weight of input i	MSE	mean square error
ω_{sr}	sunrise angle	P	pressure
ϕ	latitude	$rMAE$	relative mean absolute error
d	day	$RMSE$	root mean square error
E_0	eccentricity factor	s	forecast skill over 24 h persistence forecast
G_t	average values of measured global horizontal irradiation for period t	T	temperature
$G_{0,d}$	daily extraterrestrial global radiation on a horizontal surface	x	value of an input or of a neuron of previous layer in ANN
$G_{0,h}$	hourly extraterrestrial global radiation on a horizontal surface	ANN	artificial neural network
$G_{0,t}$	extraterrestrial global horizontal radiation on period t	DT	decision tree
H	humidity	SVM	support vector machine
h	hour	SVM-C	support vector machine for classification
		SVM-R	support vector machine for regression

and Mora-López and de Cardona (1998) propose different methods for modeling hourly and daily series of solar irradiance (using the clearness index).

All the above subjects can be considered as modeling and simulation problems and can be addressed by data mining techniques. Data mining techniques usually involve analysis of large numerical data sets to determine statistical characteristics that can be used to infer future behavior. One of the advantages of these techniques is that they have great generalization capability along with their ability to manage different types of data. Several previous models have been developed to forecast and model solar global radiation, for instance Perez et al. (2007), Mora-López et al. (2005), Viorel (2008), Guarnieri et al. (2008), Heinemann et al. (2005), Mellit and Pavan (2010), Ramirez Santigosa et al. (2003), Koca et al. (2011), Ozgoren et al. (2012) and Jiménez-Pérez and Mora-López (2014). Clustering and classification techniques have been also used in the field of solar energy, such as in Moreno Sáez et al. (2013) where clustering techniques were used to model solar spectral irradiance for use in the evaluation of the performance of photovoltaic solar modules.

Regarding the prediction of hourly solar radiation results obtained by several authors, one conclusion is that the errors and accuracy of the models vary but the errors are significant in all cases, see for instance Perez et al. (2007), Mora-López et al. (2005), Viorel (2008), Mellit and Pavan (2010), Reikard (2009), Koca et al. (2011) and Voyant et al. (2013). Mellit and Pavan (2010), The MAPE ranges from 30% to 40% in Reikard (2009) and the *RMSE* range from 23% to 28% in Voyant et al. (2013). Hence, improving models to predict hourly solar radiation is a key issue.

Taking into account all these previous results, the aim of this paper is to explore the use of different technologies to improve accuracy in the prediction of short-term hourly global solar radiation. This paper proposes a new procedure to forecast next-day hourly solar global radiation. In a previous paper (Jiménez-Pérez and Mora-López, 2014), we proposed a model to estimate hourly profiles of radiation for a day. This previous model used the daily clearness index as input and was built using the cumulative probability distribution function of hourly clearness index for each day. Several types of days were identified. The new proposed procedure extends the previous results and allows us to estimate not only the values of hourly global solar radiation for a day but also the value of the clearness index for the next day. In addition, fewer types of different days are now proposed, thanks to the use of a new variable instead of using cumulative probability distribution functions. This procedure is based on using several data mining techniques and has been checked for recorded data in Malaga, Spain. Finally, as an extension of this proposal, a procedure to simulate the hourly global solar radiation for a day using different meteorological parameters (usually public data recorded by the weather services) is proposed. These simulated data can be used for the evaluation of the performance of photovoltaic facilities that do not have monitoring systems.

The rest of the paper is organized as follows. The materials and methods used are described in the second section. The third section sets out the data used and model inputs. The proposed models are presented in the fourth section. The model and forecasts evaluations are detailed in the fifth section. Finally, the last section summarizes the conclusions of the work.

2. Materials and methods

The different methods and models to be used for modeling and forecasting hourly global solar radiation are presented in this section. Specifically, four data mining approaches are explained as they will be used in the study.

2.1. Clustering techniques

Clustering allows a data set to be partitioned into groups in such a way that one sample is more similar to samples of its cluster than to samples in other clusters according to some objective function that defines similarity or dissimilarity among samples (Han et al., 2006). *k*-means clustering is one of the most used partitional techniques. Clustering techniques are commonly used in different domains such as text mining, statistical learning and pattern recognition (Jain et al., 1999; Duda et al., 2001; Hastie et al., 2001). Clustering is based on analyzing one or more attributes (variables) to identify a cluster of correlating results. The distance from the samples to the centroid of its cluster is used to measure the similarity between samples in a cluster. Squared Euclidean distance is used, as it is defined in Jain et al. (1999).