Applied Energy 180 (2016) 392–401

Contents lists available at ScienceDirect

Applied Energy

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

Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines

Yanting Li [a], Yong He [a], Yan Su [b], Lianjie Shu [c,*]

[a] Department of Industrial Engineering and Logistics Management, Shanghai Jiao Tong University, ShangHai, China
[b] Department of Electromechanical Engineering, University of Macau, Taipa, Macau
[c] Faculty of Business Administration, University of Macau, Taipa, Macau

HIGHLIGHTS

- Suggests a nonparametric model based on MARS for output power prediction.
- Compare the MARS model with a wide variety of prediction models.
- Show that the MARS model is able to provide an overall good performance in both the training and testing stages.

ARTICLE INFO

Article history:
Received 14 December 2015
Received in revised form 12 July 2016
Accepted 15 July 2016

Keywords:
Solar power output
Weather information
NWP model
MARS

ABSTRACT

Both linear and nonlinear models have been proposed for forecasting the power output of photovoltaic systems. Linear models are simple to implement but less flexible. Due to the stochastic nature of the power output of PV systems, nonlinear models tend to provide better forecast than linear models. Motivated by this, this paper suggests a fairly simple nonlinear regression model known as multivariate adaptive regression splines (MARS), as an alternative to forecasting of solar power output. The MARS model is a data-driven modeling approach without any assumption about the relationship between the power output and predictors. It maintains simplicity of the classical multiple linear regression (MLR) model while possessing the capability of handling nonlinearity. It is simpler in format than other nonlinear models such as ANN, k-nearest neighbors (KNN), classification and regression tree (CART), and support vector machine (SVM). The MARS model was applied on the daily output of a grid-connected 2.1 kW PV system to provide the 1-day-ahead mean daily forecast of the power output. The comparisons with a wide variety of forecast models show that the MARS model is able to provide reliable forecast performance.

1. Introduction

Fossil fuels, such as coal, oil and gas, are becoming more and more deficient. Nowadays photovoltaic (PV) technology has been rapidly developed and explored, due to the fact that the solar energy is abundant, environmental-friendly and renewable [1]. With the development and advancement in PV technology, in addition to the wide establishment in remote areas, PV systems are also becoming popular in grid-connected applications.

There are many factors that could affect the power output of a PV system, including the environmental factors such as solar irradiance, temperature, insolation, wind speed, and some physical factors like the installation angle and location. Due to the variability of solar irradiance and other environmental factors, the power output of a PV system dynamically changes with time. The variability of power output not only adversely affects the stability of the electrical system being connected but also adds more risk to the profit of PV system owners [2,3]. For this reason, there is an increasing need for more accurate prediction of power output.

Various models have been proposed for forecasting the power output of a PV system. Depending on whether the model is intrinsically linear, one may classify these models into linear and nonlinear models. The multiple linear regression (MLR) and time series models are typical examples of linear models. The most commonly used MLR establishes a linear relationship between power output and climatic variables. The power output generated from a PV system varies with time, which can be viewed as a time series. Thus, it is also natural to apply the traditional time series modeling techniques to forecast the power output. The widely-used time series

* Corresponding author.
E-mail address: ljshu@umac.mo (L. Shu).

http://dx.doi.org/10.1016/j.apenergy.2016.07.052
0306-2619/

Nomenclature

t	time in day	b	the penalty coefficient
Y_t	average daily power in day t	$RMSE$	root mean square error
$\hat{Y}_t$	estimate of Y_t	MAD	mean absolute deviation
$\hat{f}(x)$	estimated functional relationship	$MAPE$	mean absolute percent error
α_0	intercept of the MARS model	m	number of external covariates
α_i	regression coefficient on the ith basis function	$\theta(B)$	MA polynomial
$BF_i(x)$	the ith basis function in MARS	$\phi(B)$	AR polynomial
$h(x)$	hinge function	p	AR order
c	knot	d	difference order
$GCV(M)$	generalized cross validation statistic	q	MA order
$C(M)$	penalty function of model complexity	k	number of neighbors in KNN
N	total number of observations in the training set		
M	number of non-constant basis functions		

models include the auto-regressive (AR) models, moving average (MA) models, and their generalizations such as the auto-regressive moving average (ARMA) and the auto-regressive integrated moving average (ARIMA) models also known as Box-Jenkins models [4]. Prema and Uma Rao [5] used a time series model for short-term solar power prediction. Pedro and Coimbra [6] used ARIMA model to forecast solar power output without exogenous inputs. To allow for exogenous inputs into the time series model, the ARMAX model can be used, which has been proved to be a powerful tool in time series forecasting [7]. Bacher et al. [8] showed that the ARX model with numerical weather conditions (NWPs) as inputs performs much better than the AR model in forecasting short-term (2-h ahead) power output. Li et al. [9] applied the ARMAX model to forecast the power output of a grid connected PV system.

In addition to linear models, there are numerous models that are inherently nonlinear in nature, which do not require the pre-specification of the exact form of nonlinearity prior to modeling. A sample of nonlinear models used for solar power output forecast includes artificial neural network (ANN), support vector machines (SVM), and K-nearest neighbors (KNN). The tree-based models like classification and regression trees (CART) are also nonlinear. A comparison between the performance of linear and nonlinear models for power output was conducted in [10].

Among the nonlinear models, the ANN model is one of the most widely used approaches. For example, [11,12] applied the ANN and evolutionary ANN for output prediction of PV systems. Chen et al. [13] used ANN together with weather classification to build an online 24-h solar power forecasting model. Almonacid et al. [14] employed the dynamic ANN for short-term forecast of the power output of a PV generator. Amrouche and Pivert [15] combined the spatial model with ANN technique for the daily global solar radiation forecasting. Long et al. [16] compared the prediction performance of ANN models with other data-driven models. Moreover, some researchers combined the ANN models with other models as a novel hybrid model in order to further improve forecast accuracy. See, for example, [17–20].

As a class of powerful and highly flexible modeling techniques, the SVM model has made great achievements in wind power prediction [21], and it has been recently proposed for solar power output forecasting. For example, [22] employed support vector regression (SVR) with numerical cloudiness to forecast power output of a PV system in Japan. Yang et al. [23] developed a weather-based hybrid method for 1-day ahead hourly forecasting of PV system power output. The proposed approach used a self-organizing map (SOM) for classification and SVR for model fitting. Similarly, [24] combined the SVR method with weather classification for power output forecasting.

The use of other nonlinear models based on regression tree and KNN for forecasting the solar power output has also been studied. For example, [25] described the quantile regression forests model to forecast power output. Zhang et al. [26] adopted the KNN method to mine the historical data to find days that were similar to the target forecasting day according to certain similarity measures. Additionally, [27] proposed the application of an analog ensemble (AnEn) method to generate probabilistic solar power forecasts.

The linear model is relatively simple but is mainly limited to model linear trends. On the other hand, the nonlinear model is more flexible but also more complex. The ANN models often involve the design of network architecture and the selection of learning algorithm. However, this heavily relies on past experience and is subject to trial and error processes since the optimal configuration is not known a priori. The SVM model is able to overcome these obstacles while it requires much computational cost, especially when there is huge amount of data. Note that in SVM, every variable has to be multiplied by the corresponding element of every support vector. That can be a slow process if there are many variables and many support vectors. The CART and KNN models are easy to interpret and implement. However, they are more widely used for classification than regression. Therefore, their prediction performance may not be as good as other nonlinear forecasting models.

The objective of this paper is to investigate the use of a fairly simple nonparametric regression model known as multivariate adaptive regression splines (MARS), as an alternative to the forecast of solar power output. The MARS model makes no assumption about the underlying functional relationship between the dependent and independent variables. Instead, it constructs this relation from a set of coefficients and basis functions that are entirely driven by the regression data. Therefore, it can be viewed as an extension of linear models that automatically models nonlinearities and interactions between variables. In a sense, this method is based on the "divide and conquer" strategy, which partitions the predictor into two groups and models the relationship between the response and the predictor in each group. This nature makes the MARS algorithm very computationally efficient, which is suitable for handling fairly large data sets and high input dimensions.

The MARS model was first introduced by [28] as a nonparametric statistical method to fit the relationship between dependent variables and independent variables. Due to the advantages of being nonparametric and computationally efficient, the MARS model has gained increasing applications in areas such as meteorology, transportation and biology. For example, [29] applied MARS model to predict the main hazardous air pollutants in the atmosphere. Zhang et al. [30] assessed the soil liquefaction based on