Applied Energy 225 (2018) 998–1012

Contents lists available at ScienceDirect

Applied Energy

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

Missing value imputation for short to mid-term horizontal solar irradiance data

Haydar Demirhan*, Zoe Renwick

RMIT University, School of Science, Mathematical Sciences, 3000 Melbourne, Australia

HIGHLIGHTS

- Accuracy of imputation methods for missing solar irradiance data are compared.
- Linear and Stineman interpolations are very precise for minutely series.
- Linear and Stineman interpolations and Kalman filters are precise in hourly series.
- Weighted moving average performs well for daily and weekly solar irradiance.

ARTICLE INFO

Keywords:
Horizontal solar irradiance
Imputation
Gap filling
Kalman filter
Missing value
Time series

ABSTRACT

Improving the accuracy of solar irradiance forecasting has become crucial since the use of solar energy power has become more accessible due to increased efficiency and decreased costs associated with its production. Data quality and availability are essential to producing accurate solar irradiance forecasts. In this article, we focus on the estimation of missing values in minutely, hourly, daily, and weekly solar irradiance series using an extensive number of imputation methods. We compare the accuracy of 36 imputation methods for solar irradiance series over a real dataset recorded in Australia under 16 experimental conditions. The experiments are run in a semi-Monte Carlo setting, in which missing values are randomly generated in the solar irradiance series. Our results identify the most reliable and robust approaches for the imputation of solar irradiance for each of the mentioned frequencies. While linear and Stineman interpolations and Kalman filtering with structural model and smoothing are found accurate for minutely and hourly series, weighted moving average gives the highly precise imputations for daily and weekly solar irradiance.

1. Introduction

As a potential renewable and clean energy resource, the interest in the solar energy has been increasing significantly in fields such as agriculture, meteorology, energy, business and ecology. Prediction of solar energy potential is essential in these fields, and the frequency of predictions change depending on the specific needs of each field. For instance, while yearly estimation of the amount of solar irradiance is important to assess the feasibility of a solar farm or a microgrid investment, a forecast horizon of less than an hour is required to connect a solar energy generator to a grid or an energy storage system [1,2]. Governments around the world have been committing to reduce greenhouse gas emissions by amounts ranging from 20% to 40% below the past levels before 2030 [3]. To achieve these reductions in greenhouse gas emissions, it is integral to integrate solar power plants composed of photovoltaic panels to the existing electric power systems. Due to the high variability in the amount of electricity generated by solar systems and its dependency on other meteorological factors, synchronisation of photovoltaic panels with power systems to generate power for the energy market is a challenging research area [4,5]. The assessment of the impact on the transient power system stability with photovoltaic (PV) systems requires irradiation forecasting in the ultra-short terms such as second intervals. Scolari et al. [6] focus on forecasting the AC active power output of a PV system in ultra-short-terms. Torregrossa et al. [7] propose a dynamic interval predictor for ultra-short-term forecasting of solar irradiance. Barbieri et al. [8] provide a comprehensive review on the methods for forecasting PV power within very-short terms over a large-scale grid-connected PV farm. Accurate and reliable short-term forecasting of the amount solar irradiance is very important for the harmonisation of solar power generators with the electric grids to ensure power continuity and ramp rates of the overall power system Barbieri et al. [8].

* Corresponding author.
E-mail addresses: haydar.demirhan@rmit.edu.au (H. Demirhan), zoe.renwick@rmit.edu.au (Z. Renwick).

https://doi.org/10.1016/j.apenergy.2018.05.054
Received 9 March 2018; Received in revised form 25 April 2018; Accepted 13 May 2018
0306-2619/

Forecasting methods applied to solar radiation prediction include physical, probabilistic and statistical, soft computing (machine learning) methods, and their hybrids [2,9,10]. Diagne et al. [9] provide an extensive review of the forecasting methods used for solar radiation up till 2013. Physical models include photovoltaic models, numerical weather prediction (NWP) models, and satellite imagery based models [11–15]. One of the main drawbacks of physical models is their dependence on temporal and spatial resolution, which determine the length of focused time interval and the vertical and horizontal spatial extent, respectively. NWP models require a high spatial resolution to resolve the cloud cover and satellite forecasts require high temporal and spatial resolution to produce accurate forecasts [9]. Grantham et al. [3] propose a probabilistic approach based on nonparametric bootstrapping to give one-hourly forecasts of solar irradiance. Shakya et al. [2] focuses on using Markov switching model to give one-day ahead forecasts of solar irradiance for remote microgrids.

Statistical methods include linear and non-linear models. The most primitive statistical approach is to use the last observation as the one-step ahead forecast. This is called naïve or persistence forecasting. More complicated statistical models include regression models, autoregressive integrated moving average (ARIMA) methodology and its nonlinear versions such as nonlinear autoregressive (NAR) and autoregressive conditional heteroscedastic (ARCH) models [16–20]. Hassan et al. [21,22] propose a modelling approach for PV modules power estimation along with a decomposition model for the solar radiation. Benmouiza and Cheknane [18] focus on a small-scale solar irradiance database to produce hourly forecasts for a period of one day using ARIMA and NAR models, and a hybrid of these approaches. Hussain and Al Alili [23] use ARIMA models to forecast global horizontal solar irradiance on hourly basis. Boland and Soubdhan [19] utilise the ARCH model along with some other modelling tools for forecasting solar irradiance at three sites in Guadeloupe in the Caribbean. Voyant et al. [24] focus on mean absolute log-return from econometrics literature as an estimator of global solar irradiance. This is in accordance with the work of Boland and Soubdhan [19] since ARCH models are applied with log-returns series [25 (p. 277)]. Trapero et al. [26] use a dynamic nonlinear regression approach called dynamic harmonic regression to forecast 1–24 hourly short-term solar irradiance forecasts. Trapero et al. [26] employ state space framework to find forecasts. Also, they provide benchmarks with ARIMA, exponential smoothing, persistence, and seasonal persistence forecasts. Jing et al. [27] propose a coupled autoregressive and dynamical system (CARDS) model to forecast solar irradiance on hourly and less then hourly forecast horizons. They observe that the CARDS model performs well consistently for cloudy and clear days [9]. Grantham et al. [3] evaluate the CARDS approach for short term forecasting of solar irradiance. The CARDS approach works through filtering the seasonal component using a Fourier approach and employing autoregressive model to account for the serial correlation. Although it is possible to get accurate forecasts with the statistical models, they have some assumptions to hold to give reliable and generic results [28].

Soft computing techniques include artificial intelligence methods such as support vector machines (SVM), artificial neural networks (ANN), fuzzy and genetic algorithms (GA). Soft computing techniques have been widely applied for long-term forecasting of solar irradiance [1,29–33]. As for short-term forecasting of solar irradiance, recently Wang et al. [34] used ANFIS, M5Tree, and Angstrom methods along with meteorological variables to forecast daily global solar radiation and compare their performance over 21 stations in China. They mention that because of the differences between model performances, more attention should be devoted to the key influencing factors in modelling solar irradiance [34]. Hassan et al. [21,22] focus on ensemble methods to model solar radiation series. Patrick et al. [35] focus on a semi-parametric time series approach including the spatial information of the locations for short-term forecasting of solar irradiance. Pedro and Coimbra [36] use k-nearest neighbour (KNN) and ANN methodology to generate 15 minutely to 2 hourly forecasts. They also assess the effect of microclimates on solar radiation forecasts. Cao and Lin [37] apply a hybrid of wavelets and ANNs for short term forecasting of solar radiation. There are also other hybrid approaches to forecast solar irradiance for short horizons [38–40]. Recently, use of deep learning methods are also considered for the analysis of solar irradiation series [41]. It is known that deep learning algorithms improve the overall accuracy, but they can be less responsive to changes in data patterns [42]. Also, special attention is required to overcome the overfitting, which causes serious problems in imputation. Specifically, although Deep Networks provide more accurate results than ARMA models, the need for a representative training set and an adapted output dimension persists [42].

All the mentioned approaches, regardless of their classification, depend on availability of solar irradiance data. Regression-like approaches need auxiliary independent variables to feed meteorological or spatial information into the analysis which is a problem with remote locations [2]. In relation to the availability of data, missing values is an important challenge in both short and long-term forecasting of solar irradiance using statistical methods, soft computing techniques, and hybrid approaches. It is very common in solar and meteorological data to have missing observations due to the faulty measurement devices in meteorological stations [43]. The negative impact of missingness is seen on the frequency of time series data. If the series is observed in irregular time intervals, the frequency of the series can no longer be uniform due to the missing observations and it turns out to be an irregularly observed series. In this case, for example, it is not possible to compute the autocorrelation function (ACF) and partial autocorrelation function (PACF), which are essential for ARIMA related models. Thus, we need to address the missing values before applying ARIMA or similar modelling techniques. Alternatively, we can use soft computing methods like KNN that are able to handle time series data with missing observations [36]. However, all the approaches including the soft computing ones suffer from loss of efficiency, complications in handling the data, and the bias due to the missingness [44]. Another approach is to use general statistical methods developed for irregularly observed time series [45]. The main drawback of these techniques is the determination of forecast horizon, which can be unclear for irregular series. Therefore, one of the most reliable ways of dealing with missing values in time series is to impute them using accurate statistical approaches.

Although having missing values in solar and meteorological variables is not rare, the literature on the imputation of missing values is limited not only in solar radiation modelling but also in the field of renewable energy. Within the renewable energy domain, Layanun et al. [46] propose use of a linear interpolation approach based on the mean of solar irradiance values under different weather types. Akcay and Filik [47] consider imputation of missing values for wind-speed forecasting by spectral analysis of long term forecasting. Rivero et al. [48] consider an energy associated tuning method using ANNs for short term forecasting by complete and incomplete time series data. They benchmark their approach using chaotic time series. Mahmudvand et al. [49] use singular spectrum analysis (SSA) to impute missing values in series. They use an iterative approach to impute missing values. Filho and Lima [50] apply SSA technique to impute missing values in precipitation series, which is also a significant exploratory variable in modelling the solar radiation [28]. They observe that the performance of SSA approach for precipitation series is closely related to the length of the series. Brooks et al. [43] use linear interpolation, ARIMA methodology and decompositions to replace missing values in global horizontal solar irradiance series. Moreno-Tejera et al. [51] propose an approach for imputation of missing values based on the frequency distribution of original series. They reflect the population characteristics of solar irradiance series into the imputation of missing values. Oluwaseyi and Song [52] focus mainly on the length of the missingness period in the solar irradiance series and compare the accuracy of SSA, statistically adjusted solar radiation, and a temperature-based approach in the