ARTICLE IN PRESS

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

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

A semi-empirical approach using gradient boosting and *k*-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting

Jing Huang *, Matthew Perry

CSIRO Oceans & Atmosphere Flagship, GPO Box 3023, Yarralumla, ACT 2601, Australia

ARTICLE INFO

Keywords:
Solar power
Probabilistic forecasting
Gradient boosting
k-nearest neighbors regression
GEFCom2014

ABSTRACT

The aim of this work is to produce probabilistic forecasts of solar power for the Global Energy Forecasting Competition 2014 (GEFCom2014). The task involves predicting the outputs from three solar farms at an hourly resolution using data from the ECMWF numerical weather prediction model.

The annual cycle of solar radiation and power is modelled using a low-pass filter built using a Fourier transformation. The diurnal cycle is handled by fitting separate models for each hour of the day with the positive solar radiation. A model for simulating PV power production, taking the effect of temperature into account, is also included.

The forecasting methods were gradient boosting for the deterministic forecasting of solar power and *k*-nearest neighbors regression for estimating prediction intervals in order to provide probabilistic forecasts. A cross-validation strategy, splitting the data into monthly folds, was employed for comparing the performances of alternative methods and in an attempt to avoid overfitting issues.

1. Introduction

Solar power forecasting is a problem which has increased in both interest and importance as a result of the increasing penetration of solar power into the global energy mix. Forecasting is a key issue for the effective management of solar photovoltaic power installations, for both large plants and distributed small rooftop systems. For example, forecasting can help with issues such as smoothing the power output to the grid, the efficient usage of battery storage, and the effective operation of energy markets.

Solar power forecasting was included as a track in the Global Energy Forecasting Competition 2014 (GEFCom2014). This paper describes the methods used by our team *Gang-gang*, which was one of the leading participants in the competition. The design and organization of the competition are described by Hong et al. (2016).

The core of the approach taken by our team involves the use of the gradient boosting method for deterministic forecasting. Gradient boosting is a machine learning technique for regression problems which was first developed by Friedman (2001). Boosting builds an ensemble of prediction models iteratively in order to optimize the loss function.

In many applications, a knowledge of the expected uncertainty of the predictions is crucial for decision making. The competition required the participants to provide forecasts of the probabilistic distribution of solar power generation in the form of 99 quantiles for each step over the forecast horizon. Our approach estimated the prediction intervals using the '*k*-nearest neighbors' regression method. Note that the *k*-nearest neighbors algorithm was also a

* Corresponding author.
E-mail addresses: jing.duke@gmail.com (J. Huang), matthew.perry@csiro.au (M. Perry).

http://dx.doi.org/10.1016/j.ijforecast.2015.11.002

Please cite this article in press as: Huang, J., & Perry, M., A semi-empirical approach using gradient boosting and *k*-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.11.002

Table 1
Times (hour and day of the year) when the solar farms have positive solar radiations.

Local hour	Day of year (DOY)		
	Farm 1	Farm 2	Farm 3
5	318–364	316–364	316–365
6	1–61, 267–365	1–61, 267–365	1–60, 267–365
7	1–135, 223–365	1–133, 224–365	1–133, 223–365
8–17	1–365	1–365	1–365
18	1–133, 197–365	1–133, 197–365	1–133, 197–365
19	1–83, 278–365	1–83, 278–365	1–83, 278–365
20	1–34, 344–365	1–34, 344–365	1–35, 343–365

winning method in the wind power forecasting track of GEFCom2012 (Mangalova & Agafonov, 2014).

This paper details the general considerations and technical approaches for data pre-processing (Section 2), the modelling of solar radiation and power (Section 3), deterministic and probabilistic forecasting, and performance evaluation (Section 4).

2. Data

2.1. Data pre-processing

There are twelve independent variables from the ECMWF numerical weather prediction (NWP) model output that are available for use in predicting the target variable (solar power). The data are provided for each of three adjacent solar farms, with an hourly time resolution. Initially, twelve months of training data were provided, with this growing to 26 months as the competition progressed.

The first part of the data pre-processing task involved differentiating the four accumulated fields, namely the total surface solar radiation directed downwards at the surface (SSRD), the surface thermal radiation downwards (STRD), the net solar radiation at the top of the atmosphere (TSR), and the total precipitation (TP). The units of SSRD, STRD and TSR were also converted to W/m^2. For the rest of the report, the processed results are referred to whenever these variables are used.

It is more convenient to use local time than UTC time, as the local time matches the expected diurnal process of solar radiation. This piece of information can be obtained by looking at the time series of SSRD and requiring the maximum radiation to occur at around the local noon. Thus, the local time zone is marked as UTC+10. In addition, solar power production relies predominantly on the solar radiation received by the solar photovoltaic (PV) panels, which are near the land surface. As such, it is important to flag the times (day and hour) that have positive solar radiations at the surface; this is accomplished by a visual inspection of the SSRD time series. Solar power is only going to be predicted for the flagged times. For the times when the surface radiation is expected to be zero, the resulting power production will be either zero or very little, due to the errors arising from measurement and the numerical weather prediction (NWP) model. The flagged times for the three solar farms are listed in Table 1.

2.2. Data splitting for cross-validation

Cross-validation was used to compare the performances of alternative methods, by assessing how the results of the model fitted on the training data would generalize to the independent prediction data. To avoid overfitting, it is crucial to split the training data appropriately for cross-validation. For this competition, it is natural to split the data into separate months, as the data are released month by month. Thus, the number of cross-validation folds n is equal to the number of months available in the training data (see Fig. 1). The model is run n times, with each fold being used as the test data in turn, and the remaining months used as the training data. This method of validation gives us confidence that any improvement in the proposed model made based on the training data will lead to an improvement for the test data as well.

3. Solar modelling

3.1. De-trending

The most prominent features of the solar radiation and power time series are their diurnal and annual cycles. As the time resolution of prediction is only hourly, it is feasible to develop an individual model for each hour of the day and each farm. In fact, there are only 16 models to be developed for different hours, as each solar farm has 8 h per day with zero radiation (cf. Table 1). However, the annual cycle needs to be modelled explicitly and removed from the original solar radiation and power time series. The annual cycle is caused mainly by the change in the solar azimuth angle for a given hour.

A low-pass filter is built for modelling the annual cycle of the solar radiation and power time series (i.e., the corresponding clear sky values), using a Fourier transformation, due mainly to its smoothness, flexibility and natural periodicity. For each hour with positive solar radiation (hours 5–20 in local time), the training data are used to determine the optimum frequency threshold for the low-pass filter. The optimum threshold is chosen to be that which produces the lowest mean absolute error (MAE) for the solar power prediction on average over all cross-validated folds. Note that we force the threshold of frequency here to be the same for radiation and power. Although it is conceptually possible to find the optimum threshold of frequency in the two-dimensional space of radiation and power, the computation time required makes this approach unattainable. Fig. 2 shows an example

Please cite this article in press as: Huang, J., & Perry, M., A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.11.002

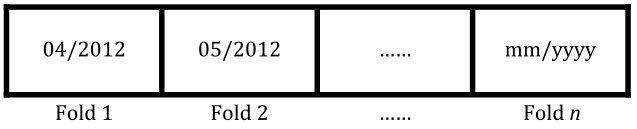

Fig. 1. Illustration of the data splitting strategy for cross-validation.

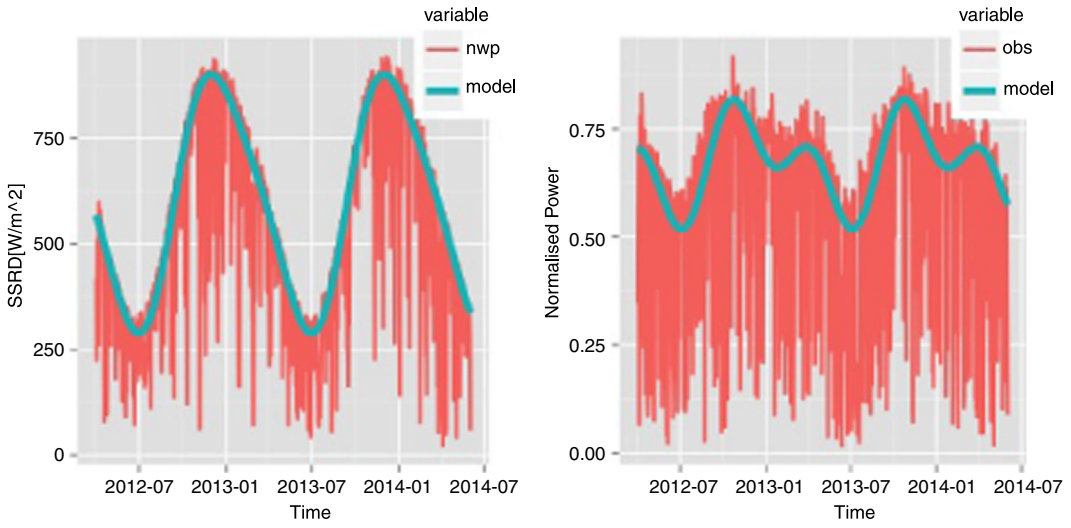

Fig. 2. Comparison of the original time series of radiation (left) and power (right), with their corresponding annual cycle models for Farm 1 and local hour 10 am in Task 15.

for Farm 1 and the local hour 10 am of the original radiation and power time series, with their corresponding clear sky models. The highest frequency kept in the low-pass filter is two cycles per year. Table 2 lists the frequency threshold for each solar farm and each hour used in the final task (Task 15), which is the highest frequency each low-pass filter keeps. The thresholds vary considerably across sites and hours, which may reflect the differences in the configurations of solar panels among the three sites, such as the panel orientations and tilt angles, the site altitudes and the solar panel types.

When the size of the training dataset is small (typically below two years), it is beneficial to filter the original power and radiation time series first using the weighted quantile regression (WQR) model mentioned by Bacher, Madsen, and Nielsen (2009), before entering the low-pass filter. For example, the implementation of the WQR reduces the quantile score by approximately 6E−4 for Task 1. However, the WQR should not be applied once the training dataset is large enough, as this is found to weaken the performances of the resulting forecasts slightly. For Task 15, adding the WQR increases the quantile score by around 2E−4.

3.2. Solar PV simulation

In addition to the clear sky model developed above, a model for simulating solar PV power production is also built for each solar farm (see Lorenz et al. 2011). This model relies on physical information from the solar research community, and accounts for the effects of ambient temperature. As such, it may provide additional value to the model, and help to improve the accuracy of the final

Table 2
The list of frequency thresholds (times per year) by local hour used in Task 15 for each farm.

Local hour	Frequency threshold (cycles/year)		
	Farm 1	Farm 2	Farm 3
5	6	9	12
6	2	2	12
7	9	2	1
8	10	3	8
9	15	9	1
10	2	3	9
11	1	4	15
12	2	2	14
13	2	2	15
14	2	11	15
15	8	4	10
16	2	2	4
17	3	9	4
18	4	8	6
19	4	2	5
20	4	4	11

prediction of solar power. The formulas from Lorenz et al. (2011) are rewritten as follows:

$$\eta_{MPP}\left(I_t, 25\ ^\circ\mathrm{C}\right) = a_1 + a_2 I_t + a_3 \ln\left(I_t\right), \quad (1)$$

$$\eta_{MPP}\left(I_t, T_m\right) = \eta_{MPP}\left(I_t, 25\ ^\circ\mathrm{C}\right)\left(1 + \alpha\left(T_m - 25\ ^\circ\mathrm{C}\right)\right), \quad (2)$$

$$T_m = T_a + \gamma I_t, \quad (3)$$

where I_t is the tilted solar irradiance perpendicular to the solar PV panel, T_m is the temperature of the solar PV panel, T_a is the ambient temperature, $\eta_{MPP}\left(I_t, T_m\right)$ is the estimated solar power production given I_t and T_m under maximum power point (MPP) conditions, and a_1, a_2, a_3 and γ are constant coefficients. Note that SSRD is the down-

Please cite this article in press as: Huang, J., & Perry, M., A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.11.002

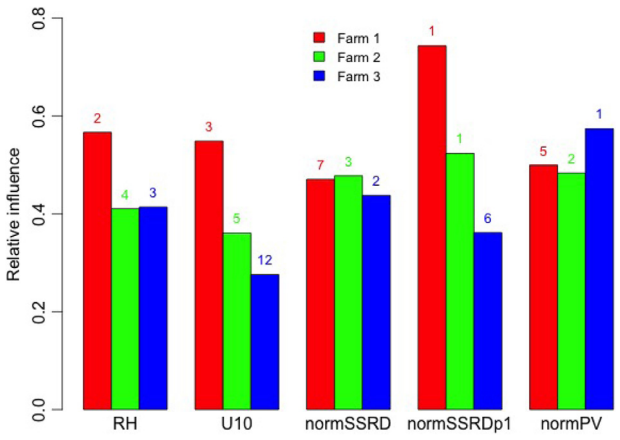

Fig. 3. Relative influences of the five most important independent variables for local hour 12 pm in Task 15, given by the GBM package. The ranking is marked for each farm and each independent variable.

Table 3
List of the dependent and independent variables used in gradient boosting for deterministic forecasting.

	ID	Description
Dependent variable	normpower	Normalized power, defined as the real power divided by the annual cycle model of power
Independent variables	TCIW	Total column cloud ice water content (kg/m^2)
	RH	Relative humidity at 1000 mbar (%)
	U10	10 m U wind component (m/s)
	V10	10 m V wind component (m/s)
	T2	Air temperature 2 m above ground level
	normTSR	Normalized net solar radiation at the top of the atmosphere, defined as TSR divided by the annual cycle model of TSR
	normPV	Normalized PV simulation model (see Section 3.2), defined as the solar PV model value divided by the annual cycle model of power
	normSSRD	Normalized downward surface solar radiation, defined as the pre-processed SSRD divided by its annual cycle model value
	normSSRDp1	normSSRD of the preceding hour
	normSSRDp2	normSSRDp1 of the preceding hour
	normSSRDp3	normSSRDp2 of the preceding hour
	normSSRDf1	normSSRD of the following hour
	normSSRDf2	normSSRDf1 of the following hour
	normSSRDf3	normSSRDf2 of the following hour

ward irradiance, which is different from I_t defined here. However, it is reasonable to assume that SSRD will be proportional to I_t for a given hour and solar farm. T_a is also equivalent to one of the variables provided, the temperature at two meters above ground level (T2). Therefore, a linear regression model is constructed for each hour and each solar farm, with the following formula:

$$power \sim (SSRD + \ln(SSRD))\,(T2 + SSRD)\,. \tag{4}$$

4. Forecasting

4.1. Deterministic forecasting using gradient boosting

The models built in Section 3 are used to develop the variables to be used as dependent and independent variables when applying the gradient boosting machine (GBM) technique to provide a deterministic forecast for solar power. The variables used are listed and described in Table 3.

A regression model is developed for each solar farm and each hour of positive irradiance. In order to account for temporal correlation, it is helpful to include the normalized SSRD for the preceding and following three hours, as well as for the prediction hour itself. Since the three solar farms are adjacent to each other, there are potential correlations between the power production of one solar farm and the independent variables of all three solar farms. As such, the power for each solar farm is forecast using the independent variables of all three solar farms (42 in total). The remaining variables were not used because they were not found to improve the accuracy of the predictions. Using the *relative.influence* function in the *GBM* package of *R*, we are able to gauge the relative influence of the independent variables quantitatively. Fig. 3 demonstrates such results for local hour 12 pm in Task 15. Not surprisingly, the

Please cite this article in press as: Huang, J., & Perry, M., A semi-empirical approach using gradient boosting and *k*-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.11.002

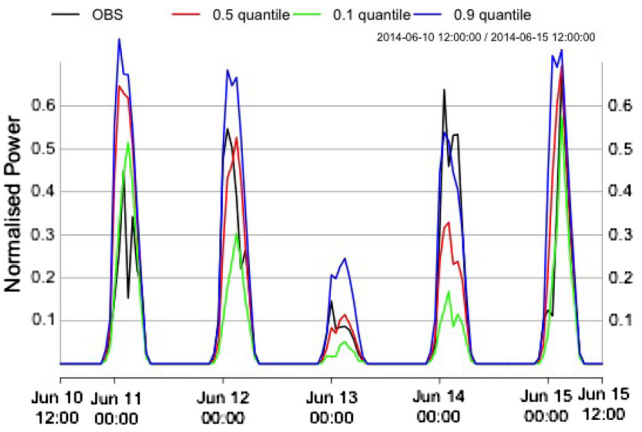

Fig. 4. Comparison of the normalized power time series between observations and three forecast quantiles (0.1, 0.5 and 0.9) for Farm 1 and five days of Task 15.

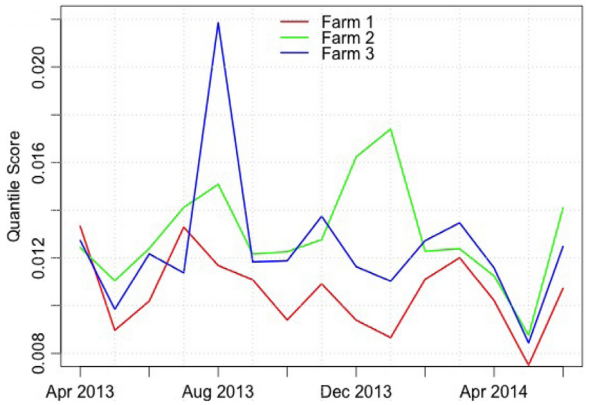

Fig. 5. Time series of the quantile score of our submissions in the trial and interim periods for the three farms.

most important independent variables include *normPV*, *normSSRD*, *normSSRDp*1, *RH* and *U*10.

4.2. Adding prediction intervals using k-nearest neighbors regression

So far, we have described how to build a model to make deterministic forecasts of solar power. However, probabilistic forecasting requires prediction intervals to be estimated and superimposed on the deterministic forecasts. The errors are likely to be related to weather variables; for example, higher errors are expected in cloudy conditions than in clear-sky situations.

We use the *k*-nearest neighbors regression method for this issue. The idea is basically to search for *k* similar scenarios from the historical data in order to form a probabilistic distribution of the forecasting error for each time point to be predicted. Here, the extent of the similarity is quantified by the Euclidean distance in a hyperspace with the dimensions defined by a number of selected variables. *k* is empirically set to 200.

Given that there may not be enough training data to form a stable probabilistic distribution function if the prediction interval is modeled for each hour, we use all of the training data for this purpose, and select *hour* as one of the relevant variables for calculating the Euclidean distance. The corresponding formula used is:

$$error \sim hour + fit + fitp1 + fitf1, \tag{5}$$

where *fit* is the cross-validated prediction of *normpower* using the data splitting strategy described in Section 2, *fitp*1 and *fitf* 1 represent *fit* in the preceding hour and the following hour, respectively, and *error* is the cross-validation error ($\equiv$ *normpower* $-$ *fit*). Note that *fitp*1 and *fitf* 1 may be dropped from the formula when the training dataset is small.

The estimated prediction intervals for solar power are superimposed onto the deterministic forecasts and a file is produced for submission in accordance with the required format. A sample period of five days is shown in Fig. 4 for Farm 1 and Task 15. While the 0.5 quantile forecast deviates from the observation significantly, it is expected that the observation will fall within the band composed of all quantiles.

4.3. Performance evaluation and evolution

The quantile scores of our forecast submissions are calculated for each task and each farm, and plotted in Fig. 5. The score is generally a reflection of the performance of

Please cite this article in press as: Huang, J., & Perry, M., A semi-empirical approach using gradient boosting and *k*-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.11.002

our forecasting methodology. However, it should be noted that the score depends on the time and location/farm as well. For example, the quantile score tends to be noticeably lower for Farm 1 than for the other two farms, which may be due to differences in the configuration settings of the solar PV panels and in the surrounding environment. Nevertheless, it can be observed that the quantile score averaged over the three farms generally decreases over time. This is a result of our ongoing efforts to improve the forecasting model throughout the competition period. The basic ideas described in the paper were implemented at the beginning of the competition, then features such as the use of low-pass filtering for modelling the annual solar trend and the consideration of temporal correlation were added later on.

5. Concluding remarks

The proposed algorithm has been implemented using R. Due to the computational intensity of the proposed algorithm, the program will ideally be run on a parallel platform with Message Passing Interface (MPI) support. A test on a 256-core Intel 2.6 GHz platform for Task 15 shows the following: the pre-processing step takes about 5 s, the modelling step takes about 17 s, the gradient boosting step takes about 640 s, and the PDF estimation step takes about 96 s. As such, the program can be finished within 13 min.

Gradient boosting was used to train the model in converting the NWP output to point forecasts of solar power. The step of de-trending the annual cycle of the time series of solar radiation and power was found to be critical for deterministic forecasting. For this purpose, low-pass filters using the Fourier transformation were built for modelling the annual trends. In addition to the variables from the NWP output, a model of solar PV power was also built for inclusion in the independent variables. A nonparametric approach was adopted for calculating the prediction intervals using a *k*-nearest neighbors regression.

These methods have proved to be successful for producing accurate forecasts of solar power outputs, together with realistic prediction intervals. Our team performed well in the leaderboard of the GEFCom2014 probabilistic solar power forecasting track, achieving an overall quantile score, calculated using the pinball loss function (Hong et al., 2016), of 0.01211. Therefore, these methods could prove useful for the operational forecasting of outputs from solar power plants for timescales from a few hours to a few days ahead, assisting in the effective management and decision making associated with this variable resource.

References

Bacher, P., Madsen, H., & Nielsen, H. A. (2009). Online short-term solar power forecasting. *Solar Energy*, *83*, 1772–1783.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *28*, 337–374.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: state-of-the-art 2015. *International Journal of Forecasting*, this issue.

Lorenz, E., Scheidsteger, T., Hurka, J., Heinemann, D., & Kurz, C. (2011). Regional PV power prediction for improved grid integration. *Progress in Photovoltaics: Research and Applications*, *19*, 757–771.

Mangalova, E., & Agafonov, E. (2014). Wind power forecasting using the *k*-nearest neighbors algorithm. *International Journal of Forecasting*, *30*(2), 402–406.

Jing Huang is currently a research scientist at CSIRO in Canberra, Australia. His research interests include renewable energy forecasting, numerical weather prediction and boundary-layer meteorology. He obtained his Ph.D. in Environmental Engineering at Duke University in 2010. Prior to that, he earned his bachelor degrees in Mechanical Engineering and Computer Science at the University of Science and Technology of China.

Matthew Perry is a meteorological data analyst at CSIRO in Canberra, Australia, working on solar energy applications, including solar forecasting and resource assessment. Matthew gained his MSc. in Environmental Statistics and Systems from the University of Lancaster, UK. He has also worked on climatological and extreme value analysis at the UK Met Office.

Please cite this article in press as: Huang, J., & Perry, M., A semi-empirical approach using gradient boosting and *k*-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.11.002