

# Short-Term Photovoltaic Output Forecasting with Weakly Related Meteorological Data

C.H. Lin, Y. Xiao, J.G. Chen, X.K. Wen

Guizhou Electric Power Test & Research Institute  
China

C. Du

North China Electric Power University  
China

**Abstract**—Photovoltaic (PV) output is influenced by many meteorological factors. The significant degree of meteorological data influences the accuracy of forecasting result. This paper proposed a short-term PV output forecasting method while the weather data and PV output data were weak correlated. By analyzing meteorological historical data and PV output historical data, the main factors effecting PV generation were found out by Pearson correlation coefficient. Based on relevant factors, fuzzy clustering analysis method was used to select similar days, and then support vector regression (SVR) forecasting model was built. SVR model has excellent learning ability for small sample. To determine model parameters, a two-step method was proposed. First, using the global search method to determine the value of parameter  $\epsilon$  and the appropriate range of kernel parameter  $p$  and regularization parameter  $C$ , then using self-adaptive differential evolution algorithm to find the optimal  $p$  and  $C$ , in order to improve the forecast accuracy when parameter  $\epsilon$  was selected in large scale. Examples show that the method proposed in this paper has good forecasting ability when the weather data and PV output data are weak correlated.

**Keywords**—distributed photovoltaic; PV output forecasting; support vector machine; parameter selected

## I. INTRODUCTION

Clean energy represented by solar energy attracts more and more widespread attention. After large-scale PV plants connect to the grid, multiple effects on safe, economic and reliable operation will be emerged [1]. So it is necessary to forecast the short-term PV output [2], in order to control it actively. Output of roof-mounted PV power systems is always forecasted by statistical method. Statistical method relies on historical record to build forecasting models, predicting output directly.

Reference [3] proposes a method to forecast roof-mounted PV system output by artificial neural network (ANN). Reference [4] describes an algorithm based on weather patterns. Firstly, divide historical samples into subclasses by empirical mode decomposition, and then use support vector machine (SVM) to forecast PV output in the different subclasses. Reference [5] forecasts short-term small-scale PV plant output by the use of season time series model and SVM method.

The literatures show that statistical forecasting more depends on meteorological data where the PV plant is located. For traditional statistical forecasting, only when the number of samples is sufficiently large, the algorithm performance can be theoretically guaranteed. But the meteorological measurement

instruments are still not perfect in China, so they cannot provide precise enough information to do forecasting. Even if the PV plants use the same meteorological information measured by one device nearby can also cause problems such as the correlation to the output data is weak, the noise in the sample is large, so the number of effective samples is reduced. The applicability of traditional statistical method is lessening.

To overcome the disadvantage that less correlation between meteorological data and distributed PV output data, we use support vector regression (SVR) model to forecast output. First, fuzzy clustering analysis (FCA) was used to extract similar days. Subsequently, support vector regression model which has good prediction ability for small sample was established. Then a two-step method which combines global grid search (GS) algorithm and self-adaptive differential evolution (SADE) algorithm was proposed to select optimal parameters, forming a short-term PV output forecasting method called FCA-GS/SADE-SVR method. Finally, a practical example was tested to verify the effectiveness of the proposed method.

## II. STRONG CORRELATION SAMPLE CONSTRUCTION

### A. Characteristics of Short-Term PV Output

PV output is affected by many factors [6-7] like the intensity of solar radiation, solar incidence angle, angle of the solar cell module installed, temperature, wind speed, cloud amount, dust amount, shadow, etc. For a given PV system, relevance of PV output and external factors are inherent in the historical data [8]. Therefore, through the study of historical data, the ability to predict the future information can be obtained.

### B. Data Pre-Processing

The roof-mounted PV power plant in this article locates at longitude  $106^{\circ}07' \sim 107^{\circ}17'$ , between latitude  $26^{\circ}11' \sim 27^{\circ}22'$ . As used herein, the PV output data are taken from the roof-mounted PV power plant actual record from February 1, 2013 to March 31, 2013, while the sample time is 10min. Wind speed and temperature data are from the local numerical weather forecasting, time resolution is 1h. Because solar radiation intensity measuring devices are not yet universal, so we use HOMER software to simulate radiation intensity in hours. Thus, the test data have a gap between real situation, and correlation between output data is weak.

Normalized all data to eliminate the impact of different dimensions. The formula is as follows:

$$x = \frac{x' - x'_{\min}}{x'_{\max} - x'} \quad (1)$$

Where:  $x'$  are the real value of PV output, light intensity, temperature, and wind speed;  $x'_{\max}$  are the maximum PV output, light, temperature and wind speed in all samples;  $x'_{\min}$  are the minimum output, light, temperature and wind speed;  $x$  are the normalized value.

We use the most widely used Pearson correlation coefficient method to analyze the relationship between the PV output and light intensity, temperature, wind speed.

Set two variables  $X$  and  $Y$ , and each group of samples is expressed as  $(X_i, Y_i) \ (i = 1, 2, \dots, n)$ , the Pearson linear correlation coefficient formula is:

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

$$\text{Where, } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

According to the historical data within 2 months, the correlation coefficients of each meteorological factors associated with PV output are calculated, shown in Table 1.

TABLE 1. CORRELATION COEFFICIENTS BETWEEN PV OUTPUT DATA AND METEOROLOGICAL DATA.

Meteorological factor	Solar irradiance	Temperature	Wind speed
Correlation coefficient	0.55	0.64	0.13

The two variables will have higher degree of linear relationship if their correlation coefficient closer to 1. When  $|\hat{\rho}| \geq 0.8$ , it is regarded as highly relevant. Table 1 shows the verification of weak correlation among samples.

Choose relatively strong correlated factors: light intensity and temperature, as the main factors affecting the photovoltaic power generation, building strong correlated small-sample.

### C. Similar Days Selected By Fuzzy Clustering Algorithm

By appropriate screening and classification, it can improve the similarity between samples and improve accuracy of forecasting. Fuzzy clustering algorithm is used to select similar days in this paper, according to factors relatively strong affecting the PV output.

Set  $X = \{x_1, x_2, \dots, x_k\}$  as the historical sample, and set  $x_i \ (i = 1, 2, \dots, k)$  as a single sample.  $x_i \ (i = 1, 2, \dots, k)$  is a vector including all weather data of one day.  $k$  is the number of historical days. There are  $m$  factors, and  $x_i$  can be expressed as

a vector  $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$ . Because different meteorological factors impact PV output in different level, so we give appropriate weighting factor for each factor. In this paper, taking Euclidean distance method to calculate:

$$d_i = \sqrt{\sum_{j=1}^m \lambda_j (x_{ij} - x_j)^2} \quad (3)$$

Where:  $m$  is the number of influence factors;  $\lambda_j$  is the correlation coefficient calculated by equation (2);  $x_{ij}$  is  $i$ -th day's  $j$ -th meteorological factor;  $x_j$  is the  $j$ -th meteorological factor of the day to be predicted.

When  $d_i$  is smaller, the association is stronger. Select the three largest correlation days as the similar days. The meteorological and PV output data of the similar days constitute the training and testing samples, acting as SVM model input as well.

## III. SUPPORT VECTOR MACHINE AND PARAMETER OPTIMIZATION

### A. Support Vector Machine

SVMs are statistics learning tools introduced by Vapnik in 1995, these are usually used in classification and regression problems. SVR algorithm is mainly used v-SVR,  $\epsilon$ -SVR and LS-SVR [9-10], etc.  $\epsilon$ -SVR requires less parameters and has good generalization performance [10], which make it the prediction model used in this article.  $\epsilon$ -SVR is as follows:

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 + C \frac{1}{l} \sum_{i=1}^l (\xi_i + \xi_i^*) \\ \text{s.t.} \quad & y_i - w\phi(x_i) - b \leq \epsilon + \xi_i, \xi_i \geq 0 \\ & w\phi(x_i) + b - y_i \leq \epsilon + \xi_i^*, \xi_i^* \geq 0 \end{aligned} \quad (4)$$

The model can be written as:

$$f(x) = \sum_{i=1}^N y_i (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (5)$$

Where  $w$  is a vector of weights, and  $b$  a constant. The slack variables  $\xi_i, \xi_i^*$  are introduced to compensate the possible presence of excessive noise or outliers. So  $\xi_i$  and  $\xi_i^*$  are the positive and negative error respectively. The positive constant  $C$  is a hyper-parameter adjusting the compromise between the amounts authorized error and the flatness of the function  $f$ .

The function  $k$  is called kernel function; sigmoidal and radial basis function (RBF) defined as follows:

$$K(x_i, x) = \exp\left(-\|x_i - x\|^2 / 2p^2\right) \quad (6)$$

In summary, parameters  $C, \epsilon$  and  $p$  are the key parameters that affect SVM prediction performance, and therefore it needs to select the optimum parameters.

### B. Parameters Optimization

The current methods of determining the parameters are: experience/experiment select method [11-13], genetic algorithm [14], grid search and

other optimization algorithms. Grid search can traverse all possible combinations of parameters, having advantages to solve small-sample forecasting problems. Therefore, we used a two-stage method to determine  $\epsilon$ -SVR parameters: firstly, using a grid search to determine  $\epsilon$  and possible range of C and p; secondly, using SADE for C and p optimization.

The data from February 1 to March 31 were the training samples, the data from February 1 to February 8 were the test samples. Set  $(\epsilon, p, C)$  as optimization variable to design grid search test. Set  $\epsilon$  within the range of 0.001 to 0.01, changing in steps of 0.001. Set p in the range of 0.1 to 2, changing in steps of 0.1. Set C within the range of 0.1 to 10, changing in steps of 0.5. A three-dimensional grid was formed. RMSE (Root Mean Squared Error) was calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2}$$

Where, n is the number of all parameter groups,  $y_i$  is the predicted values,  $y'_i$  is the true values.

Project RMSE curve onto  $\epsilon$  plane. The results that RMSE changing with  $\epsilon$  are shown in Figure 1.

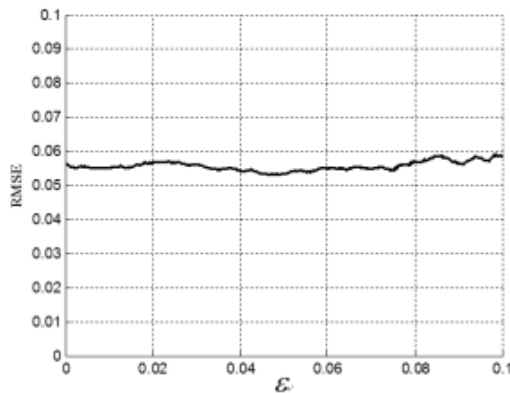


FIGURE I. RMSE VALUES CHANGE WITH  $\epsilon$ .

Figure 1 shows that, when the parameter  $\epsilon$  changes, RMSE are in between from 0.05 to 0.06. RMSE changes stably, meaning the impact of  $\epsilon$  on the performance of the model is not significant. Known from the literature [14], when p and C are fixed on a suitable value range, SVR performance is not sensitive to  $\epsilon$ .

$\epsilon$  controls the sparseness of support vectors. The larger  $\epsilon$  is, the fewer support vectors are. When  $\epsilon$  exceeds a certain figure, less learning phenomenon will exist, increasing the prediction error. So we make  $\epsilon = 0.01$  in test 2.

Set  $(p, C)$  as optimization variable to design grid search test again. Set  $\epsilon$  to 0.01. Set p in the range of 0.1 to 2, in steps of 0.1. Set C in the range of 0.1 to 10, in steps of 0.5. RMSE is calculated in the same way, and results are shown in Figure 2.

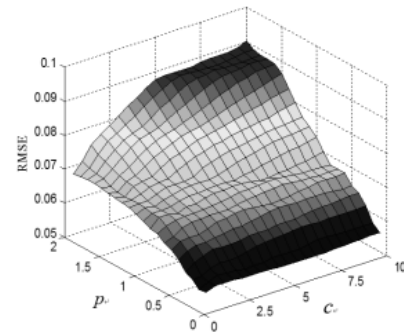


FIGURE II. RMSE VARY WITH P AND C.

So we determined the appropriate range of C and p through the grid search, and used self-adapting differential evolution (SADE) [15] algorithm to select optimum figure.

Using SADE algorithm to generate initial population randomly, the number of individuals is 20. Then calculate the fitness of the initial population of individuals. New offspring individuals were generated in solution space by mutation and crossover strategy based on differential evolution algorithm [16], evaluated by fitness. Offspring and parent populations were selected by greedy algorithm, based on individual fitness. After a certain number of iterations, we got the best C and p, making the structural risk minimum.

#### IV. CASE STUDY AND RESULTS

The output data of the roof-mounted PV power plant and the light, temperature data in the region from February 1, 2013 to March 31 were acted as test data. Two methods were applied in the case. One is the FCA-GS/SADE-SVR model proposed above, another is a typical traditional statistical method—back propagation (BP) neural network forecasting method. The comparative results are shown in Figure 3~4.

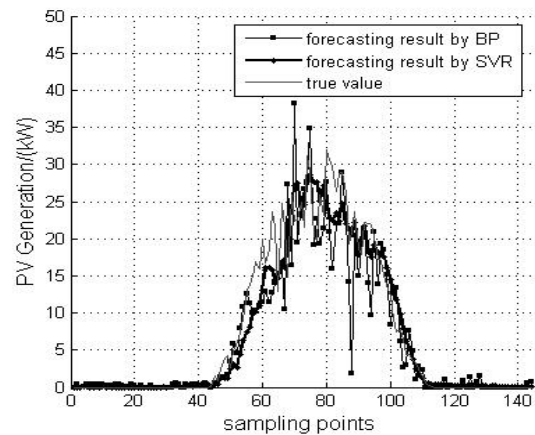


FIGURE III. COMPARISON OF PREDICTION RESULT USING BP AND FCA-GS/SADE-SVR.

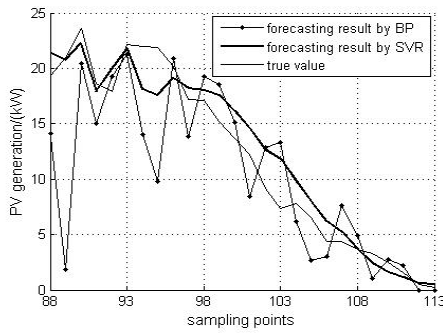


FIGURE IV. PARTIAL COMPARING FIGURE OF PREDICTION.

Multiple comparison tests were proceeded by the two methods, getting the average RMSE and illustrating as follows:

TABLE II. AVERAGE RMSE OF TEST RESULT.

	For ecas t day 1	For ecas t day 2	For ecas t day 3	For ecas t day 4	For ecas t day 5	For ecas t day 6	For ecas t day 7	For ecas t day 8
BP	8.9	7.29	7.37	7.41	4.88	3.72	9.01	5.64
FCA- GS/SA DE- SVR	7.28	4.86	6.58	2.84	4.55	2.98	9.7	4.62

Known from Table 2, Figure 3 and Figure 4, the method proposed in this paper can describe PV output short-term characteristics more accurately except the seventh day. The average RMSE of 8 days calculated by BP prediction is 6.789%. While using SVR method, the figure is 5.551%, and improving 1.238% by contrast.

## V. CONCLUSIONS

This paper analyzes the actual output data of distributed PV and local weather data, selecting light intensity and temperature as the relatively strong correlation factors. Based on two related factors, fuzzy clustering theory is applied to select similar days. A two-stage method is proposed to determine  $\epsilon$ -SVR parameters. Case shows that the FCA-GS / SADE-SVR method can describe characteristics of PV output in short-term more accurately than the BP neural network forecasting method.

## ACKNOWLEDGEMENTS

This work was supported by the National Key Technology R&D Program of China (No. 2013BAA02B02).

## REFERENCES

- [1] Tan Y, Meegahapola L, Muttaqi K M. A review of technical challenges in planning and operation of remote area power supply systems[J]. Renewable and Sustainable Energy Reviews, 38: 876-889, 2014.
- [2] Brouwer A S, van den Broek M, Seebregts A, et al. Impacts of large-scale Intermittent Renewable Energy Sources on electricity systems, and how these can be modeled[J]. Renewable and Sustainable Energy Reviews, 33: 443-466, 2014.
- [3] MaoMeiqin, GongWenjian, ZhangLiuchen, etc. Short-term photovoltaic generation forecasting based on EEMD-SVM combined

- method[J]. Proceedings of The CSEE, 33(34): 17-24, 2013.
- [4] MaoMeiqin, GongWenjian, ZhangLiuchen, etc. Short-term photovoltaic generation forecasting based on EEMD-SVM combined method[J]. Proceedings of The CSEE, 33(34): 17-24, 2013.
- [5] Bouzerdoum M, Mellit A, MassiPavan A. A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant[J]. Solar Energy, 98: 226-235, 2013.
- [6] Ren Hong, Ye Lin. Operation characteristics of PV system under the influence of environmental factors[J]. Transactions of China Electrotechnical Society, (10): 158-165, 2010.
- [7] Zhang Xueli, Liu Qihui, Ma Huimeng, Li Bei[J]. Power System and Clean Energy, 28(5): 76-81, 2012.
- [8] Zhang Jiawei, Zhang Zijia. Short-term photovoltaic system power forecasting based on PSO-BP neural network[J]. Renewable Energy Resources, 8: 006, 2012.
- [9] Vapnik V N. The Nature of Statistical Learning Theory. Berlin: Springer-Verlag, 1995.
- [10] Vapnik V N, Golowich S, Smola A. Support vector machine for function approximation, regression estimation and signal processing//Michael C, et al, ed. Advances in Neural Information Processing Systems 9, Cambridge, MA: Massachusetts Institute of Technology Press, 1997.
- [11] Lan Hua, Liao Zhimin, Zhao Yang. ARMA model of the solar power station based on output prediction[J]. Electrical Measurement & Instrumentation, (2): 31-35, 2011.
- [12] Safie F M. Probabilistic Modeling of Solar Power Systems[C]. Atlanta, GA: Reliability and Maintainability Symposium, 1989.
- [13] Muselli M, Poggi P, Nottton G, et al. First order Markov chain model for generating synthetic "typical days" series of global irradiation in order to design photovoltaic stand-alone systems[J]. Energy Conversion and Management, 42(6): 675-687, 2001.
- [14] Ding Ming, Zhou Ning. A method to forecast short-term output power of photovoltaic generation system based on Markov Chain[J]. Power System Technology, 35(1): 152-157, 2011.
- [15] Brest J, Greiner S, Bošković B, et al. Self-adapting control parameters in differential evolution: a comparative study on numerical benchmark problems[J]. IEEE Transactions on Evolutionary Computation, 10(6): 646-657, 2006.
- [16] Advances in differential evolution[M]. Springer Verlag, 2008.