Proceedings of the 2018 13th World
Congress on Intelligent Control and Automation
July 4-8, 2018, Changsha, China

# A Simulation Method of Solar Irradiance Data Based on Feature Clustering and Markov Transition Probability Matrix

Xingbo Fu, Feng Gao, Jiang Wu, Xiaohong Guan, Xuan Li, Pengyuan Liu, Pai Li

*Abstract*—**Solar irradiance is one of the significant influential factors of solar photovoltaic power generation and it is necessary to model and simulate abundant solar irradiance data. In this paper, we propose a simulation approach of solar irradiance data based on feature clustering and Markov transition probability matrix. We introduce the features of solar irradiance data, k-means algorithm and Markov transition probability matrix of solar irradiance conditions, which make up simulation algorithm of solar irradiance. According to this method, a simulation example of National Renewable Energy Laboratory (NREL) one-minute data is presented and the paper gives analysis and evaluation of the results. Finally, there are the conclusion and some possible extensions.**

**Key Words: Solar irradiance; feature clustering; Markov transition probability matrix**

## I. INTRODUCTION

Renewable energy sources, such as solar energy and hydropower, are playing a significant role in energy utilize, which obvious influence the effectiveness and stability of power generation. In photovoltaic generation systems, the output power is fluctuant and transformative because of some meteorological conditions, especially solar irradiance [1]. In order to do research on output power of photovoltaic system, we need abundant solar irradiance data. When we face the fact that the data are deficient, modeling and simulation of solar irradiance is a must.

Some researchers have paid attention to modeling and simulation of solar irradiance. C. W. Richardson experienced stochastic simulation of daily precipitation, temperature and solar radiation data and he gave a multivariate model with means and standard deviations of the variables conditioned on the wet or dry status of the day as determined by the precipitation model [2]. Indira Karakoti et al. predicts monthly mean daily diffuse radiation for India by daily sunshine and relative humidity using the partial regression analysis [3].

These researchers provide useful attempts to simulate solar irradiance data. However, some data, such as precipitation and relative humidity, are not available. Therefore, in this paper we concentrate on the properties of solar irradiance itself. The NREL collects one-minute solar irradiance data [4]. Using the raw data as the input for analysis may not be an efficient choice. Therefore, we utilize several features to take the place of the raw data.

Sanorita Dey et al. concentrates on smartphone accelerometers and when they study the time series of sensor data they take some features of data as unique fingerprints and classify these devices successfully [5]. From this inspiration, we use several statistic index, such as means and standard deviation, as the input rather than the raw data. Weidong Zhang et al. propose the evaluation indexes of fluctuation including average fluctuation magnitude (AFM), reverse fluctuation count (RFC) and moving fluctuation intensity (MFI) [6]. These indexes are also used in our model in this paper.

Mellit et al. found that the transition probability that the daily irradiance leads to a choice of a first order Markov-type analysis [7], based on Markov transition matrices technique [8]. Therefore, Markov chain in this paper can consist of solar irradiance states of each day and Markov transition probability matrix is used to simulate data.

The remainder of the paper is organized as follows. We introduce the simulation method of solar irradiance in Section II. Section III is a simulation example of this method and we show some analyses and evaluations in Section IV. Section V gives the conclusions and some possible extensions.

## II. SIMULATION METHOD OF SOLAR IRRADIANCE

### A. Features of Solar Irradiance Data

NREL's Solar Radiation Research Laboratory (SRRL) monitors and records irradiance data and meteorological data for several years and the data include one-minute solar irradiance data [4]. The Baseline Measurement System(BMS) are located at NREL's SRRL site (Latitude: 39.742° North, Longitude: 105.18° West, Elevation: 1828.8 meters AMSL). Solar irradiance data consist of global irradiance (Global CMP22), direct irradiance (Direct CHP1-1) and diffuse irradiance (Diffuse 8-48) data. Figure 1 shows these three irradiance data on June 20, 2015. In this paper, we focus on the global irradiance (Global CMP22).

One-minute direct irradiance data have 1,440 values and 1,440 time stamps each day. As for a time series, if taking the raw data set as the input, the dimension of the vector is enormous. Hence, we extract several features to decide whether a time series is similar to another rather than by using the raw direct irradiance data.

There are many researches concentrating on extracting features of time series. Sanorita Dey's research centers on accelerometer characteristics of smartphones and he select 8

The research presented in this paper is supported in part by the National Key Basic Research Program of China (2016YFB0901900), the National Natural Science Foundation of China(61773308) and Open Fund of State Key Laboratory of Operation and Control of Renewable Energy & Storage System.

Xingbo Fu and Feng Gao are with State Key Laboratory for Manufacturing System Engineering, Xi'an Jiaotong University, Xi'an 710049. E-mail: fgao@sei.xjtu.edu.cn.

Jiang Wu, Xiaohong Guan, Xuan Li and Pengyuan Liu are with Systems Engineering Institute and Ministry of Educational Key Lab for Intelligent Networks and Networks Security, Xi'an Jiaotong University, Xi'an 710049.

Pai Li is with China Electric Power Research Institute, Beijing, 100192.

978-1-5386-7346-1/18/$31.00 © 2018 IEEE

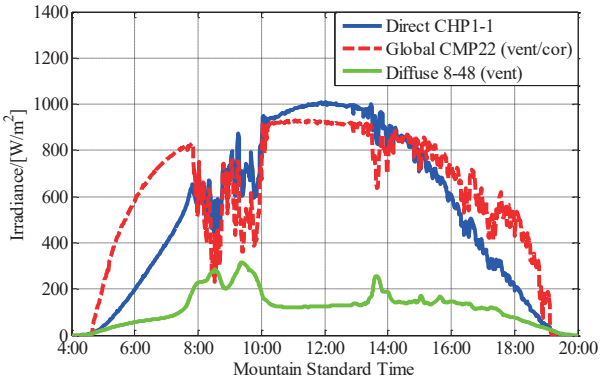


Figure 1 Load curve of a steel plant

time domain features and 10 frequency domain features to describe the data [5]. Weidong Zhang proposes evaluation indexes of fluctuation and characterizes irradiance fluctuations by Average Fluctuation Magnitude (AFM), Reverse Fluctuation Count (RFC) and Moving Fluctuation Intensity (MFI) [6].

We select 5 features (see Table I) to describe the irradiance data.

Table I
LIST OF 5 FEATURES

| Feature Name | Description |
|---|---|
| Mean | $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x(i)$ |
| Standard Deviation | $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x(i)-\bar{x})^2}$ |
| Skewness | $\gamma = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x(i)-\bar{x}}{\sigma}\right)^3$ |
| Kurtosis | $\beta = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x(i)-\bar{x}}{\sigma}\right)^4 - 3$ |
| Moving Fluctuation Intensity | MFI = AFM ×RFC |

Average Fluctuation Magnitude (AFM) is defined as

$$\mathrm{AFM} = \frac{1}{N}\sum_{i=1}^{N} abs[x(i+1) - x(i)]. \quad (1)$$

Reverse Fluctuation Count (RFC) is defined as count value that when fluctuation trend is reversed, then count value add 1. Fluctuation trend is reversed when

$$[x(i+1) - x(i)][x(i) - x(i-1)] < 0. \quad (2)$$

In this way, we transfer 1,440-dimension raw data to 5-dimension vector as the input.

## B. K-Means Clustering Algorithm

Clustering classifies data as several clusters in which properties of data are as similar as possible and k-means algorithm is one of common clustering algorithms [9]. K-means algorithm finds the optimal solution when the sum of the distance between each point and the classification center that each point belongs to.

When we classify data set $x$ with m values into $k$ clusters $c$, K-means algorithm is as follows:

Step 1: Set k random points of data set x as the initial cluster center $\mathrm{cc}(j), j = 1,2,\cdots\cdots,k$;

Step 2: Compute the Euclidean distance between each point and each cluster center $\mathrm{d}(i,j), i = 1,2,\cdots\cdots,m, j = 1,2,\cdots\cdots,k$;

Step 3: As for each point $x(i) \in x$, if

$$\mathrm{d}(i,j) = min\{\mathrm{d}(i,j), j = 1,2,\cdots\cdots,k\}$$

then the point $x(i)$ will be classified into the $j$ th cluster $(x(i) \in c(j))$;

Step 4: Set the average value of all the points in each cluster as the new cluster center $\mathrm{cc}(j), j = 1,2,\cdots\cdots,k$;

Step 5: Go back to Step 2, until stop.

In this paper, each day is classified into several clusters by k-means algorithm according to the 5 features of each day.

As we know, photovoltaic power generation is significantly influenced by solar irradiance. In this paper, we regard photovoltaic power generation suitability of each day as the state of Markov chain according to solar irradiance [8]. When we classify all the days into $k$ clusters by K-means algorithm, each cluster corresponds to one of states of photovoltaic power generation suitability and there are k states in all. And the next part introduces correlations between each two states.

## C. Markov Chain and Transition Probability Matrix

Markov chain is a kind of Markov random process for discrete time and discrete state. $\mathrm{E} = \{e_1, e_2, \cdots\}$ is the discrete state space of a discrete time series $\mathrm{X}(n), n = 1,2,\cdots$. $\mathrm{X}(t) = e$ means that the random process stays in the state $e$ at the time $t$. According to non-aftereffect property of Markov process [10], the future state $\mathrm{X}(t+1)$ only depends on certain variables and changes of current state $\mathrm{X}(t)$:

$$\mathrm{P}\{\mathrm{X}(t+1) = j|\mathrm{X}(0) = i_0, \mathrm{X}(1) = i_1, \cdots, \mathrm{X}(t-1) = i_{t-1}, \mathrm{X}(t) = i\} = \mathrm{P}\{\mathrm{X}(t+1) = j|\mathrm{X}(t) = i\}. \quad (3)$$

We define $\mathrm{P}(i,j)$ as the transition probability from $\mathrm{X}(t) = i$ to $\mathrm{X}(t+1) = j$. When the data set has $n$ days and the event that one day's state is $i$ with the next day's being $j$ occurs for $m$ times, then

$$\mathrm{P}(i,j) = \mathrm{P}\{\mathrm{X}(t+1) = j|\mathrm{X}(t) = i\} = \frac{m}{n}. \quad (4)$$

Transition probability makes up Markov transition probability matrix

$$\mathrm{P} = \begin{pmatrix} \mathrm{P}(1,1) & \mathrm{P}(1,2) & \cdots & \mathrm{P}(1,k) \\ \mathrm{P}(2,1) & \mathrm{P}(2,2) & \cdots & \mathrm{P}(2,k) \\ \vdots & \vdots & & \vdots \\ \mathrm{P}(k,1) & \mathrm{P}(k,2) & \cdots & \mathrm{P}(k,k) \end{pmatrix}. \quad (5)$$

Evidently, as for any state $i$, the transition probability in Markov transition probability matrix satisfies

$$\sum_{j=1}^{k} \mathrm{P}(i,\mathrm{j}) = 1. \quad (6)$$

## D. Simulation Algorithm

Now we can continually simulate data based on transition probability matrix and the raw solar irradiance data. The algorithm is as follows:

Step 1: Choose an initial state $i_1$ randomly, set simulation days $n_e$, generated data set $S_d = \{i_1\}$, the counter $n = 1$;

Step 2: Find out the next day's state $i_2$, which satisfies

$$P(i_1, i_2 - 1) < r < \sum_{i=1}^{i_2} P(i_1, i). \quad (7)$$

r is a random number between 0 and 1;

Step 3: $S_d = S_d \cup \{i_2\}$, $i_1 = i_2$ and $n$=$n$+1;

Step 4: Go to Step 5 when $n > n_e$, or go back to Step 2;

Step 5: As for each element of $S_d$, select the raw one-minute data of a day whose state is the element as simulated data set;

Step 6: Connect the data set of each day together and get the simulated data.

## III. SIMULATION EXAMPLE

### A. Features of Solar Irradiance Data

The data of this example come from Baseline Measurement System(BMS) located at NREL's SRRL site (Latitude: 39.742° North, Longitude: 105.18° West, Elevation: 1828.8 meters AMSL). In this experiment, we utilize the one-minute global irradiance data (Global CMP22) from January 1, 2006 to December 31, 2016.

The irradiance data are divided into four season because of the remarkable deviations among photoperiods of seasons. Meanwhile, we notice that solar irradiance from 21:00pm to 3:00am is too small to satisfy the power generation threshold of photovoltaic plant so the data from 21:00pm to 3:00am are deleted. Hence, the scale of processed data is shown in Table II.

Then we calculate the 5 features (introduced in Part A of Section II) of each day using processed data. Avoiding the impact of units, we normalize these 5 features of each day and the 5 normalized features are the input of clustering.

TABLE II
SCALE OF PROCESSED DATA

| | |
|---|---|
| Years | 2006~2016 |
| Days | Spring:982<br>Summer:1012<br>Autumn:1012<br>Winter:1012 |
| Samples of each day | 1080 values<br>from 3:00am to 21:00pm |

### B. K-Means Clustering

We classify all the days from the same season into 4 clusters according to the 5 features by k-means algorithm. To show the clustering result intuitively, we calculate the average one-minute irradiance of all days of the same cluster shown in Figure 2.

More solar radiations, fewer fluctuations and smaller skewness are three of conditions for output power of Photovoltaic system, so we distinguish these curves according to these three factors. Each curve of summer and winter in Figure 2 is a level of photovoltaic power generation suitability that Class I is the best condition while Class IV the worst. However, curves of spring and autumn in Figure 2 are not as obvious as the first two since the blue curve and black curve intertwine with each other. We regard the blue curves as Class II and the black curves as Class III because the blue curves have fewer fluctuations and smaller skewness.

TABLE III show the quantities of each cluster. The first step of simulation algorithm depends on it.

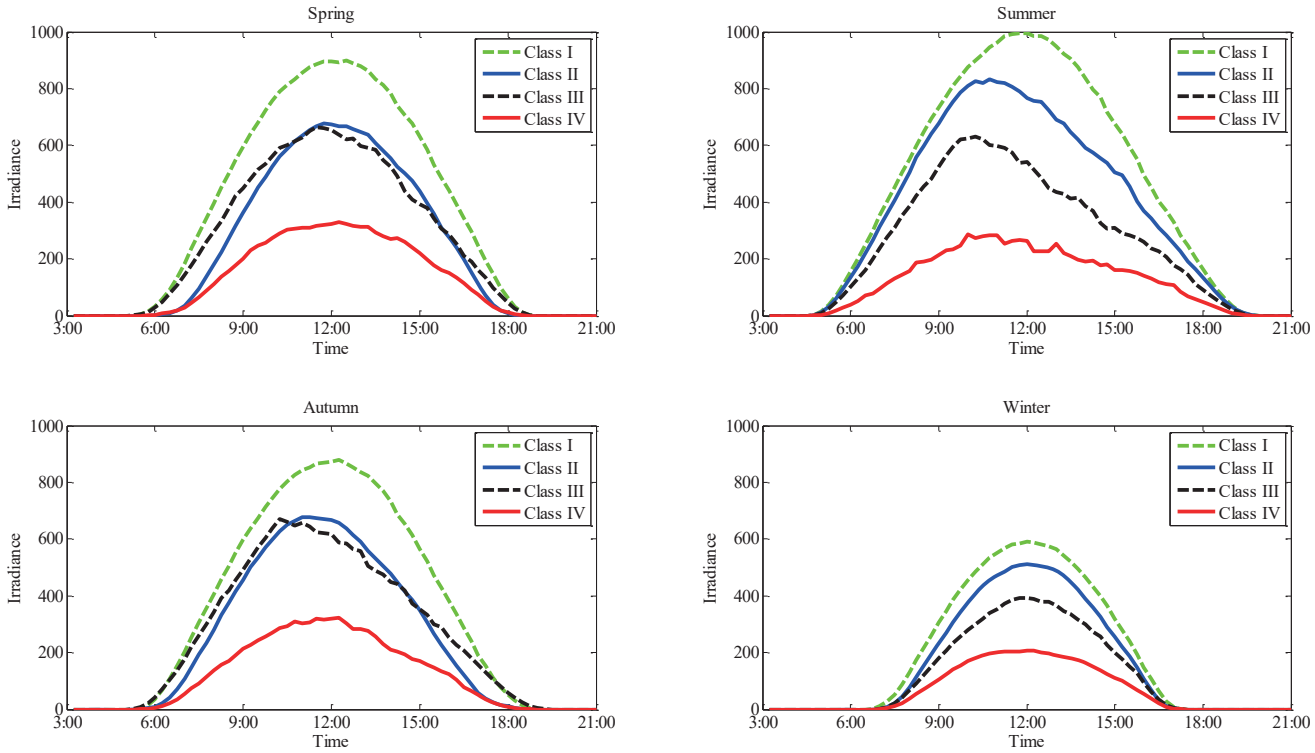


Figure 2 Average one-minute irradiance of all days of the same cluster

TABLE III
QUANTITIES OF EACH CLUSTER

| | Spring | Summer | Autumn | Winter |
|---|---|---|---|---|
| Class I | 274 | 383 | 326 | 239 |
| Class II | 312 | 364 | 315 | 367 |
| Class III | 184 | 193 | 213 | 237 |
| Class IV | 212 | 72 | 158 | 169 |
| Total | 982 | 1012 | 1012 | 1012 |

### C. Markov Transition Probability Matrix

According to Equation 4, we calculate each $P(i,j)$. And Markov Transition Probability Matrices of each season is as follows:

$$P_{sp} = \begin{pmatrix} 0.516 & 0.136 & 0.234 & 0.114 \\ 0.129 & 0.521 & 0.109 & 0.241 \\ 0.279 & 0.180 & 0.311 & 0.230 \\ 0.195 & 0.376 & 0.133 & 0.295 \end{pmatrix}. \quad (8)$$

$$P_{su} = \begin{pmatrix} 0.522 & 0.323 & 0.126 & 0.029 \\ 0.334 & 0.420 & 0.204 & 0.041 \\ 0.230 & 0.377 & 0.298 & 0.094 \\ 0.225 & 0.211 & 0.183 & 0.380 \end{pmatrix}. \quad (9)$$

$$P_{au} = \begin{pmatrix} 0.527 & 0.212 & 0.212 & 0.049 \\ 0.166 & 0.478 & 0.127 & 0.229 \\ 0.378 & 0.151 & 0.392 & 0.080 \\ 0.140 & 0.408 & 0.121 & 0.331 \end{pmatrix}. \quad (10)$$

$$P_{wi} = \begin{pmatrix} 0.581 & 0.182 & 0.136 & 0.102 \\ 0.113 & 0.452 & 0.245 & 0.190 \\ 0.145 & 0.286 & 0.286 & 0.184 \\ 0.144 & 0.395 & 0.275 & 0.186 \end{pmatrix}. \quad (11)$$

### D. Simulation

We simulate solar irradiance data using the simulation algorithm (introduced in Part D of Section II) according to Markov Transition Probability Matrices (8) ~ (11) and the raw solar irradiance data. The simulated data are one-minute solar irradiance data for three years.

## IV. ANALYSIS AND EVALUATION

The analysis and evaluation will focus on these three aspects: (1) the number of clusters, (2) the performance of clustering and (3) comparison between the indexes of simulated data and the raw data.

### A. The number of clusters

Silhouette coefficient is a index to evaluate the performance [11]. For object $i$, $a_i$ is defined as the average dissimilarity of $i$ to all other objects that is in the same cluster as $i$ and $b_i$ is defined as the average dissimilarity of $i$ to all other objects that is in the different cluster from $i$. Then the silhouette coefficient of object $i$ is

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (12)$$

From the definition above, it is obvious that

$$-1 \le s_i \le 1. \quad (13)$$

When $s_i$ is close to 1, the clustering result has a good performance. When $s_i$ is close to -1, the clustering result has a terrible performance.

The average silhouette coefficient is

$$S = \frac{1}{N}\sum_{i=1}^{N} s_i. \quad (14)$$

Figure 3 show the average silhouette coefficients in different numbers of clusters. When the data are classified into 4 clusters, the silhouette coefficient is 0.808, which is acceptable.

Some researchers did classify days into 4 states [12]. That is in accordance with meteorology.

### B. The performance of clustering

In this part, we reduce the 5 features of data to 2 dimensions in order to evaluate the performance of clustering. The principal components analysis (PCA) is a powerful method of dimensional reduction for highly correlated data [13]. In this paper, the PCA is chosen for data dimensionality reduction by PCA function in Matlab.

Table IV shows the accuracy of dimensional reduction at different dimensions for each season. As it shows, the average accuracy is 90.7% at 2 dimensions. After dimensional reduction to 2 dimensions, we draw the point graph to show the clustering result in Figure 4.

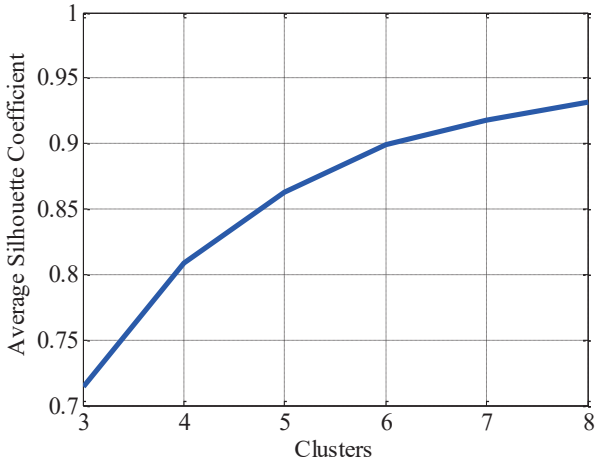

Figure 3 Average silhouette coefficient at different number of clusters

TABLE IV
ACCURACY OF DIMENSIONAL REDUCTION

| Dimensions | Accuracy | | | | |
|---|---|---|---|---|---|
| | Spring | Summer | Autumn | Winter | Average |
| 1 | 67.5% | 70.6% | 67.5% | 75.1% | 70.2% |
| 2 | 89.1% | 90.8% | 91.1% | 91.6% | 90.7% |
| 3 | 98.4% | 97.9% | 98.4% | 99.3% | 98.5% |
| 4 | 99.9% | 99.9% | 99.9% | 99.9% | 99.9% |
| 5 | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% |

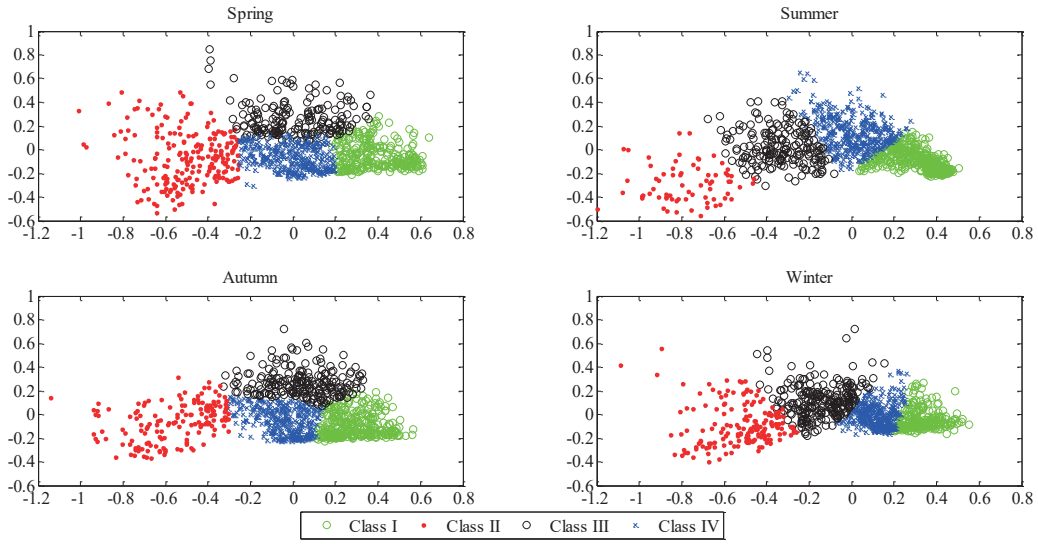


Figure 4 Point graphs of all the seasons

### C. Comparison

We calculate the indexes of the 3-year simulated data and the raw data to make comparison in this part. The result is shown as Table V.

TABLE V
COMPARISON BETWEEN SIMULATED DATA AND RAW DATA

| | Mean | Std-Dev | Skewness | Kurtosis | MFI |
|---|---|---|---|---|---|
| 2006 | 263.1 | 275.0 | 0.736 | 2.37 | 2835748.75 |
| 2007 | 255.2 | 324.8 | 0.660 | 3.10 | 3411530.84 |
| 2008 | 267.0 | 275.9 | 0.726 | 2.37 | 2980057.68 |
| 2009 | 256.2 | 277.7 | 0.674 | 3.85 | 3523860.97 |
| 2010 | 264.2 | 287.6 | 0.681 | 3.40 | 3700079.16 |
| 2011 | 265.4 | 275.4 | 0.721 | 2.38 | 3111849.03 |
| 2021 | 264.9 | 290.3 | 0.526 | 8.63 | 3074894.21 |
| 2013 | 258.9 | 304.0 | 0.625 | 5.63 | 4017362.12 |
| 2014 | 253.6 | 267.1 | 0.804 | 2.62 | 4083371.37 |
| 2015 | 248.2 | 306.9 | 0.368 | 16.09 | 4825031.90 |
| 2016 | 265.1 | 277.1 | 0.740 | 2.38 | 3872833.70 |
| Year 1 | 256.4 | 288.8 | 0.634 | 6.68 | 4199649.77 |
| Year 2 | 259.2 | 286.1 | 0.641 | 5.74 | 3823214.69 |
| Year 3 | 259.5 | 288.5 | 0.650 | 5.24 | 3718644.62 |

## V. CONCLUSION

The simulation method of solar irradiance based on feature clustering and Markov chain and transition probability matrix has been proposed.

Some researchers simulate solar irradiance according to other meteorological data such as precipitation and relative humidity, which may not be available. In this paper, we classify the days into 4 states by 5 features of the raw one-minute solar irradiance data each day of NREL using k-means algorithm. Then the transition probability from the state today to that the next day makes up Markov transition probability matrix. Based on Markov transition probability matrix and the raw data, we simulate one-minute solar irradiance data for 3 years.

We evaluate the simulation method from 3 aspects—the number of clusters, the performance of clustering and comparison between the indexes of simulated data and the raw data.

In this paper, the solar irradiance data of each year are regarded as having the same distribution. But in the real world solar irradiance may have a tendency (increase or decrease) with years, which can be a factor that influences clustering and simulation. This is another possible future work direction.

## REFERENCES

[1] Kaplani E, Kaplanis S. A stochastic simulation model for reliable PV system sizing providing for solar radiation fluctuations[J]. Applied Energy, 2012, 97(3):970-981.

[2] Richardson C W. Stochastic simulation of daily precipitation, temperature, and solar radiation[J]. Water Resources Research, 1981, 17(1):182–190.

[3] Karakoti I, Das P K, Singh S K. Predicting monthly mean daily diffuse radiation for India[J]. Applied Energy, 2012, 91(1):412-425.

[4] http://midcdmz.nrel.gov/srrl_bms/

[5] Dey S, Roy N, Xu W, et al. AccelPrint: Imperfections of Accelerometers Make Smartphones Trackable[C]// Network and Distributed System Security Symposium. 2014.

[6] Zhang W, Liu Z. Simulation and analysis of the power output fluctuation of photovoltaic modules based on NREL one-minute irradiance data[C]// International Conference on Materials for Renewable Energy and Environment. IEEE, 2014:21-25.

[7] Mellit A, Benghanem M, Arab A H, et al. A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach[J]. Solar Energy, 2005, 79(5):469-482.

[8] Aguiar R J, Collares-Pereira M, Conde J P. Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices[J]. Solar Energy, 1988, 40(3):269-279.

[9] Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C]// Proc. of, Berkeley Symposium on Mathematical Statistics and Probability. 1966:281-297.

[10] Häggström O. Finite Markov Chains and Algorithmic Applications[M]. Cambridge University Press, 2002.

[11] Rousseeuw P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational & Applied Mathematics, 1987, 20(20):53-65.

[12] Ming D, Ningzhou X U. A Method to Forecast Short-Term Output Power of Photovoltaic Generation System Based on Markov Chain[J].

Power System Technology, 2011, 35(1):152-157.
[13] Zhang T, Yang B. Big Data Dimension Reduction Using PCA[C]// IEEE International Conference on Smart Cloud. IEEE, 2016:152-157.
[14] Li S, Ma H, Li W. Typical solar radiation year construction using k-means clustering and discrete-time Markov chain[J]. Applied Energy, 2017, 205:720-731.