Daily Clearness Index Profiles Cluster Analysis for Photovoltaic System

Chun Sing Lai, *Student Member, IEEE,* Youwei Jia, *Member, IEEE,*
Malcolm D. McCulloch, *Member, IEEE,* and Zhao Xu, *Senior Member, IEEE*

Abstract—Due to various weather perturbation effects, the stochastic nature of real-life solar irradiance has been a major issue for solar photovoltaic (PV) system planning and performance evaluation. This paper aims to discover clearness index (CI) patterns and to construct centroids for the daily CI profiles. This will be useful in being able to provide a standardized methodology for PV system design and analysis. Four years of solar irradiance data collected from Johannesburg (26.21S, 28.05E), South Africa are used for the case study. The variation in CI could be significant in different seasons. In this paper, cluster analysis with Gaussian Mixture Models, K-Means with Euclidean distance, K-Means with Manhattan distance, Fuzzy C-Means with Euclidean distance and Fuzzy C-Means with Dynamic Time Warping (FCM DTW) are performed for the four seasons. A case study based on sizing a stand-alone solar PV and storage system with anaerobic digestion biogas power plants is used to examine the usefulness of the clustering results. It concludes that FCM DTW and GMM can determine the correct PV farm rated capacity with an acceptable energy storage capacity, with 36 and 46 rather than 1457 solar irradiance profiles respectively.

Index Terms—Clearness Index, Photovoltaic system, Dynamic time warping, Fuzzy C-Means

I. Introduction

SOLAR Photovoltaic distributed generation (PV-DG) systems are being integrated worldwide into distribution systems at a rapid rate [1]. Due to the intermittent nature of PV sources which are generally densely connected in low-voltage distribution network. Voltage and power fluctuations on the grid must be considered. To study the fluctuations, statistical evaluation and localized spectral analysis of the fluctuation power index should be further investigated [2]. As a result of the analytical monitoring costs, there are limited number of studies on PV systems operation in remote areas. To reduce the costs, clustering results are needed for analyzing the performance and sizing of PV systems.

Given the statistical distribution of the solar irradiance, a large quantity of data can be characterized with only very few parameters [3-9]. An example for the practical application of solar irradiance statistical modelling is provided in [3], for a case study in Tahifet, Algeria. It is learnt that the installed PV system produces excess energy in October and energy storage is required in June and December.

Solar irradiance is characterized by short fluctuations mainly introduced by passing clouds. The analysis of these fluctuations with regard to solar energy applications should focus on the instantaneous clearness index (CI) [2, 6, 10, 11]. CI can effectively characterize the attenuating impact of the atmosphere on solar insolation by specifying the proportion of extraterrestrial solar irradiance that reaches the surface of the earth. Performance analysis of the PV systems studied with classification scheme of CI profiles provides useful insights [10, 12]. The ability of generalization of this technique allows the proposed method to be applied to other system configurations for evaluation purposes, such as sizing energy storage system [13]. In particular, it is shown that cloud-induced fluctuations in CI can be treated by statistical analysis.

Manuscript received November 16, 2016; revised February 16, 2017; accepted March 9, 2017.

C. S. Lai is with the Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou, Guangdong 510006 China and also with the Energy and Power Group, Department of Engineering Science, University of Oxford, OX1 3PJ, United Kingdom (e-mail: chun.lai@eng.ox.ac.uk).

Y. Jia is with the Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong, China and also with the Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou, Guangdong 510006 China (e-mail: corey.jia@connect.polyu.hk).

M. D. McCulloch is with the Energy and Power Group, Department of Engineering Science, University of Oxford, OX1 3PJ, United Kingdom (e-mail: malcolm.mcculloch@eng.ox.ac.uk).

Z. Xu is with the Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: eezhaoxu@poly.edu.hk).

This paper provides the grouping of daily CI profiles and to construct centroids with cluster analysis. Section II provides the literature review on statistical analysis of PV and renewable energy sources. Section III presents the clear-sky solar model and real-life solar data collected for CI calculation purposes. The research problem and preliminary understanding will also be provided. Section IV gives the clustering algorithms and distance metrics used for clustering of daily CI profiles. Section V will present the clustering results for the four seasons with the five clustering techniques. To evaluate the usefulness of the clustering results for PV system planning, a case study based on sizing a stand-alone solar PV and storage system with anaerobic digestion biogas power plants is given in Section VI. Section VII provides the conclusion and future work of the research.

II. Literature Review

Fractal analysis of daily solar irradiance measured with a time step of 10 minutes at Golden and Boulder located in Colorado is provided in [14], with the aim to perform classification of daily solar irradiance. These results lead to three classes, namely clear-sky, partially covered sky and overcast sky. The daily distributions of CI were classified by estimating a finite mixture of Dirichlet distribution in [5]. The results display four distinct classes of distributions corresponding to different types of days. However, in the two studies, the CI in different seasons or months has not been studied or given.

The use of models with CIs for any solar system applications, such as solar hydrogen production is appropriate and simple. This is due to the CI only needs the global solar irradiance data [15]. The knowledge of the statistical behavior of short-term variability of solar irradiance will provide a more accurate evaluation of the uncertainty in the long-term annual energy production of solar power plants [16]. CI can be used to train the Markov transitions matrix, in order to approximate the daily irradiance value with Markov model [17]. Irradiance sequences can be generated via this method. Reference [18] uses CI to separate forecasting complexity into the prediction of solar geometry and the prediction of cloudiness and aerosol. The quadratic and cubic equations which are based on global solar irradiance data have the highest accuracy in predicting the diffuse fraction as a function of CI [19-21].

Wavelet analysis is applied to the daily CI profiles in [22], and which is decomposed into components to evaluate the endurance and magnitude of various fluctuations of the solar irradiance. The classification of typical meteorological days from global irradiance data is given in [23]. The classification was performed with aggregation Ward's method. It is learnt that the recorded days are clustered in 3, 4 or 5 groups for monthly time step and 3 groups are classified for annual time-step. The authors relied on discriminant analyses to evaluate the number of clusters and this was achieved by visual inspection.

PV generations are commonly presented by Beta distribution [9]. This assumption has been widely used for the system planning purposes. However, in reality, the underlying distribution may vary widely due to the hemisphere and climate of the location [24]. Reference [8] determined the parameters of the appropriate distribution that provide the best fit for CI. The global solar irradiance is thereafter predicted from CI using inverse transformation of the cumulative distribution function. The proposed method is effective in predicting the monthly average global solar irradiance.

Pattern recognition and cluster analysis have been applied to other renewable sources. A statistical approach was proposed for improvement of short-term wind electric power forecast based on pattern recognition technique [25]. The predictions on wind speed and direction to identify patterns of the wind behavior at the location considered to obtain a stochastic distribution of the daily wind speed were studied in [26]. A statistical hybrid wind power forecast technique was proposed in [27], where weather events are clustered with respect to the most important weather forecast parameters.

III. Clearness Index Development

A. Data Acquisition of Real-life Solar Irradiance

The CI is developed with the solar irradiance data collected from the Skye Instruments SKS 1110 Pyranometer sensor [28, 29]. The cosine-corrected head, a sensor consists of a semiconductor diode, and a light filter system for the wavelength range 350nm-1100nm were used to construct the pyranometer. Cosine-corrected head is required to avoid measurement errors when the sensor is not directly below the sun. The pyranometer can be used for energy balance studies, as the head is perfectly sealed and can be placed indefinitely in outdoor conditions. World Radiometric Reference [30] is used for the calibration of sensor under open sky conditions.

The pyranometer sensor was placed on a perfectly flat surface in order for the top light collecting surface to be exactly horizontal. Four years of solar irradiance data, from 2009 to 2012 were obtained in Johannesburg for this research. Johannesburg has a latitude of 26.21°S, longitude of 28.05°E and with an altitude of 1753m. The data sampling rate is at 1 sample/30min.

B. Clear-Sky Solar Irradiance Model

Under perfect atmospheric condition, the earth will absorb the solar irradiance which is equal to the solar constant minus the amount absorbed by the atmosphere of the earth. The solar constant is at a value of 1367 Wm^{-2}. The global solar irradiance on a horizontal surface has two main components, namely the direct beam component and the diffuse sky irradiance.

The other factor in the attenuation of the atmosphere is a function of the concentrations of the various elements in the atmosphere [31]. Their impacts can be assessed by comparing the actual observed optical depth with the theoretical optical depth of a perfectly clean dry scattering Rayleigh atmosphere. The ratio of the two optical depths is known as the Air mass 2 Linke turbidity factor, T_{LK}. The clear-sky beam irradiance normal to the beam I_{model} at the surface is calculated as mentioned in references [32, 33].

$$I_{model} = I_o \varepsilon \exp\left(-0.8662 T_{LK} m \delta_r(m)\right) \sin\gamma_s \quad (1)$$

$$\varepsilon = 1 + 0.0334 \cos\left(j' - 2.80^o\right) \quad (2)$$

$$j' = \frac{J * 360}{365.25} \quad (3)$$

$$m = (p/p_o)/\{\sin\gamma_s + 0.50572(\gamma_s + 6.07995)^{-1.6364}\} \quad (4)$$

I_o is the solar constant, ε is the correction factor to mean solar distance, m is the optical air mass corrected for station height, γ_s is the solar altitude angle in degrees and δ_r is the Rayleigh optical depth, J is the Julian day and j' is the Julian day angle. p/p_o is the pressure correction for station height and is calculated with Equation (5) given below:

$$\frac{p}{p_o} = \exp\left(-\frac{z}{H_R}\right) \quad (5)$$

z is the site elevation above sea level in meter and H_R is a constant at 8400 meters. δ_r is calculated as follows [34].

$$\frac{1}{\delta_r(m)} = 6.6296 + 1.7513m - 0.1202m^2 + 0.0065m^3 - 0.00013m^4 \quad \text{if } m < 20 \quad (6)$$

$$\frac{1}{\delta_r(m)} = 10.4 + 0.718m \quad \text{if } m >= 20 \quad (7)$$

The solar altitude angle is calculated as a function of time of day with Equation (8) [32].

$$\gamma_s = \sin^{-1}(\sin\phi\sin\delta + \cos\phi\cos\delta\cos\omega) \quad (8)$$

$$\omega = 15(t - 12) \quad (9)$$

ϕ , δ and ω are the latitude of location, solar declination angle and solar hour angle respectively. All are in degrees. t is the instantaneous time of the day in hour with values between 0 and 23.

C. Real-life Solar Irradiance Data Analysis

To examine the nature of the real-life irradiance data, the clear-sky model is used to provide comparisons. T_{LK} has been set to 5 to model the diffuse irradiance. A comparison of solar insolation data from different sources are summarized in Fig. 1. Further comparisons are made with the NASA data obtained in [35]. The maximum amount of insolation received is in December and the minimum amount is in June. The insolation is generally higher in Summer (Dec, Jan, Feb) season as compared to other seasons such as Spring (Sept, Oct, Nov), Autumn (March, Apr, May) and Winter (June, Jul, Aug).

NASA provides the solar insolation for clear-sky condition. The solar model and NASA data will have a higher monthly averaged insolation incident as compared to the real-life data. It can be seen that the three sources give a similar trend and this gives a good indication that the data is statistically accurate.

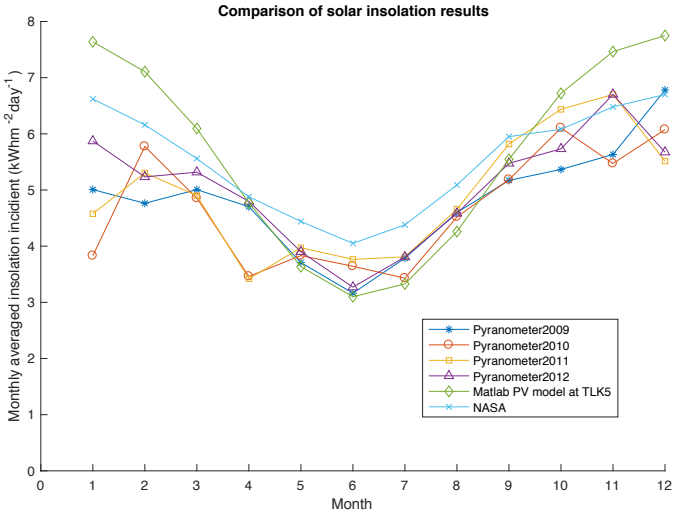


Fig. 1. Comparison of solar insolation data.

D. Clearness Index

CI at instantaneous time t is expressed as a ratio between 0 and 1, where 1 signifies there is no loss in irradiance, i.e. all the insolation is of direct beam irradiance, and 0 means there is no irradiance due to complete cloud cover. It is worth mentioning that CI can be undefined when no irradiance is available, such as before sunrise and after sunset. These conditions are not considered in this work as they are not applicable for the study. CI is calculated with Equation (10) below.

$$\mathrm{CI}(t) = \frac{I_{pyranomter}(t)}{I_{model}(t)} \quad (10)$$

$\mathrm{I_{pyranomter}}$ is the real-life solar irradiance and $\mathrm{I_{model}}$ is the clear-sky solar irradiance from the solar model. To calculate the solar model irradiance for CI, the TLK is set to 1 to remove the effect due to the clear-sky solar irradiance atmospheric absorption and scattering. These phenomena can be reflected in the CI, as it takes into account of the total irradiance reduction from the clear-sky irradiance. Fig. 2 presents the clear-sky and real-life solar irradiance for a typical day in January.

CI for four different seasons between 2009-2012 is shown in Fig. 3. Each color represents a CI profile for a day. 20 profiles were plotted for each season due to the space limitation. It can be seen that in winter there are significantly more clear days, i.e. higher CI. In contrast, CIs in summer are mostly below 0.3. CI also displays the nature of uncertainty and the daily fluctuation.

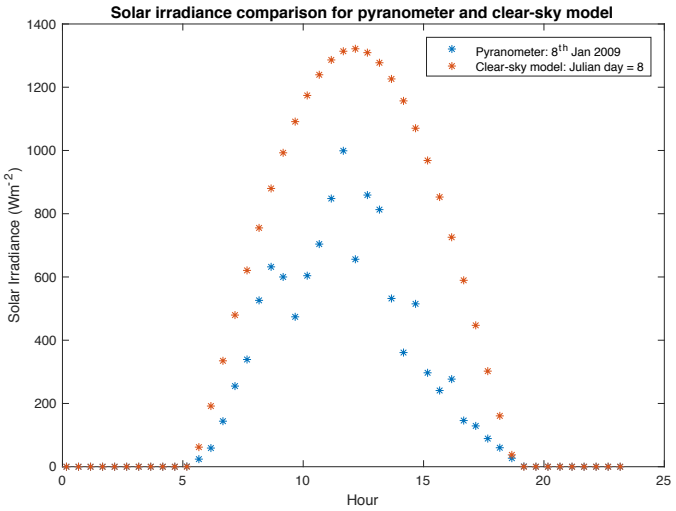


Fig. 2. Solar irradiance for clear-sky model and real-life (pyranometer) data.

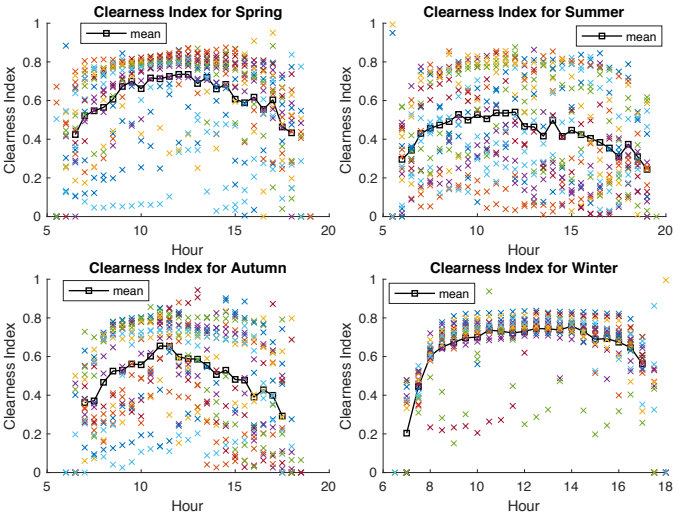


Fig. 3. Clearness index for the four seasons.

IV. Clustering Methods

A. Distribution-Based Clustering

In distribution-based clustering, clusters can be defined as objects belonging most likely to the same distribution. A Gaussian Mixture Model (GMM) is a weighted sum of m components, i.e. the number of clusters. The Gaussian mixture densities for vectors x is given in Equation (11) [36] below:

$$g_{x;w,\theta}(x;w,\theta)=\sum_{i=1}^{m} w_i g_{x;\theta_i}(x;\theta_i) \tag{11}$$

w is the mixture weight with the constraints $w_i>0$ and $\sum_{i=1}^{m} w_i=1$. $g_{x;w_i,\theta_i}(x;w_i,\theta_i)$ is known as the component Gaussian densities. The parameter θ contains the component weights w_i, mean vectors, μ_i and the covariance matrices Σ_i. This is expressed with Equation (12) [36] below:

$$\theta=\{w_1,w_2,\dots,w_m,\mu_1,\mu_2,\dots,\mu_m,\Sigma_1,\Sigma_2,\dots,\Sigma_m\} \tag{12}$$

For h_n to be the number of elements in the vector x_n, the log likelihood function is given in Equation (13) [36] below:

$$L(\theta)=\sum_{n=1}^{h_n}\ln\sum_{i=1}^{m} w_i g_{x;\theta_i}(x_n;\theta_i) \tag{13}$$

The Expectation-maximization (EM) algorithm aims to calculate the maximum likelihood estimation of the marginal likelihood in an iterative process. The process consists of two stages, the expectation step and maximization step.

1. Expectation step: Calculate the expected value of the log likelihood function under the current estimate of the parameters $\theta^{(t)}$ at t iteration [36, 37].

$$B\left(\theta;\theta^{(t)}\right)=E_{X,\theta^{(t)}}[L(\theta)] \tag{14}$$

2. Maximization step: Find the parameter that maximizes the following quantity [36, 37]:

$$\theta^{(t+1)}=\underset{\theta}{\operatorname{argmax}}\, B\left(\theta;\theta^{(t)}\right) \tag{15}$$

B. Partition-Based Clustering

1) K-Means

K-Means clustering aims to classify the objects into the clusters with the nearest mean. It is an iterative algorithm and begins with choosing K initial cluster centers. The distances of all observations to each centroid are computed. The object is assigned to the cluster with the closest centroid. The new centroid locations are determined by calculating the average of the objects in each cluster. Given data with n vectors of equal lengths, $X=\{x_1,x_2,\dots,x_n\}$, K-Means determines cluster centers for k clusters of vectors with equal lengths $V=\{v_1,v_2,\dots,v_k\}$, by minimizing the objective function as given in Equation (16) [38] below:

$$\min_{V}\sum_{i=1}^{k}\sum_{j=1}^{n}\mathrm{D}^2\left(x_j,v_i\right) \tag{16}$$

where D is the distance function, such as Euclidean distance (ED), Manhattan distance (MD) and Dynamic Time Warping (DTW) etc.

2) Fuzzy C-Means

The Fuzzy C-Means (FCM) algorithm is an extended version of the K-Means algorithm by including the fuzzy-partition matrix. Each object can belong to more than one cluster. The iterative process is similar to K-Means. The objective function is given in Equation (17) [38] below:

$$\min_{V}\sum_{i=1}^{k}\sum_{j=1}^{n}(w_{ij})^m\mathrm{D}^2(x_j,v_i) \tag{17}$$

The fuzzifier m determines the level of cluster fuzziness, where $1\le m\le\infty$. A large m results in smaller membership values. The centroid for FCM is calculated with Equation (18) [38].

$$v_i=\frac{\sum_{j=1}^{n}(w_{ij})^m x_j}{\sum_{j=1}^{n}(w_{ij})^m},\ i=1,\dots,k. \tag{18}$$

The fuzzy-partition matrix is given in Equation (19) [38].

$$w_{ij} = \frac{1}{\sum_{h=1}^{k} \left(\frac{\mathrm{D}(x_j, v_i)}{\mathrm{D}(x_j, v_h)}\right)^{\frac{2}{m-1}}} \tag{19}$$

3) Distance Metrics

One of the crucial element in partition-based cluster analysis is the function used to measure the similarity between time series. The distance metric can have a profound effect to the clustering of times series and to their respective clusters.

a) Euclidean distance and Manhattan distance

ED and MD are the most frequently used distance measures for data mining. Although both have been deployed in many time series application fields, the metric has many pitfalls for time series analysis. They can only be used for time series of equal length and are prone to noise and outliers, which are common in real-life temporal sequences especially when noise and uncertainty exist in the data source [39]. Another major issue with ED is that the metric is based on the comparison between data points at the same time interval. Time series regularly suffer transformations in the time axis although the series are in a similar shape, i.e. due to the perturbation to the solar irradiance and in the context of CI, there will be time discrepancies for sunrise and sunset in the clear-sky solar model and real-life solar irradiance data. Let x_i and v_j each be a d-dimensional vector, the ED and MD between the two vectors are presented in Equations (20) [38] and (21) [40] respectively.

$$\mathrm{ED} = \sqrt{\sum_{f=1}^{\mathrm{d}} (x_{if} - v_{jf})^2} \tag{20}$$

$$\mathrm{MD} = \sum_{f=1}^{\mathrm{d}} \left| x_{if} - v_{jf} \right| \tag{21}$$

b) Dynamic Time Warping

Dynamic Time Warping (DTW) has many features that may overcome the drawbacks of ED and MD. In essence, the objective of DTW is to find the optimal alignment between the two series by searching for the minimal path in a distance matrix that defines a mapping between them, whist satisfying the moving restrictions during the searching process, i.e. only vertical, horizontal and diagonal moves are allowed. This results in stretching and compressing of time series. The mapping for every pair of points in the series can be determined by distance metrics such as ED and Manhattan etc. In this paper, the distance metric used for FCM DTW is ED. The outputs of DTW are the cost matrix that denotes the cost values, i.e. the DTW distance between the two coordinates and the warping path. The main weakness of DTW is the computational complexity. The algorithm for calculating the DTW for two time series is given in [41].

The main challenge in applying DTW distance to partition-based clustering techniques is to calculate the average of a set of time series. To overcome this issue, DTW barycenter averaging (DBA) is used for DTW averaging. Unlike the traditional centroid calculation method where the mean is directly determined, the aim of DBA is to minimize the sum of squared DTW distances between the centroid sequence and the set of sequences to be clustered. This is essentially achieved by performing two iterative procedures. The first stage is to perform DTW to the sequence to be clustered and the centroid to be refined. The associations between them are kept and will be stored in the vector or known as the association table. The second stage is to update each coordinate of the centroid with the barycenter of coordinates associated with it from the association table, by calculating the mean. The standard deviation of the centroid can also be calculated with the association table. The algorithm for DBA can be found in [42].

As previously explained, the cluster centers cannot be calculated with the traditional method in Equation (18) when considering DTW as the distance function in Equation (17). The cluster centers are calculated with DBA instead. To initialize the cluster centers in each cluster, the time series are assigned to the cluster having the maximum membership degree. The cluster centers will be refined in the iterative FCM optimization process. The partition matrix is calculated with Equation (19), where D stands for the DTW distance.

C. Comparison of Computational Complexities

Let *N, I* and *d* be the number of profiles, number of iterations and dimension of profiles respectively. The K-Means computational complexity (CC) for ED or MD is approximated as O(*NKdI*) [43]. For FCM with ED, the CC is approximated as O(NK^2dI). The FCM suffers a higher computation costs compared to K-Means, this is due to the need for updating the fuzzy-partition matrix in each iteration [44].

For GMM clustering, the CC is mainly associated with the EM algorithm. This is approximated as O($INK(1+d^2)+K$) [37], which is higher than the partition-based clustering methods for high dimensional data [45, 46]. The covariance and mean matrix grow in the size of Kd^2 and Kd respectively.

For FCM DTW, the differences in CC per iteration with respect to the standard FCM algorithm are the distance and centroid calculations. The CC of DTW is quadratic, d^2, unlike the linear computation costs for ED and MD [42]. Therefore, the total complexity for distance calculation is O(Nd^2). The centroids are calculated with the DBA method and has a computation cost of O(INd^2) [42]. The resulting FCM DTW computation cost is therefore O$\{Nd^2I(K^2Nd^2)\}$, which simplifies to O($IN^2K^2d^4$). FCM DTW has the highest CC compared to other clustering methods and future research should look for more efficient methods in using DTW for FCM clustering.

V. Cluster Analysis and Discussions

A. Clustering of Daily Clearness Index Profiles

Cluster analysis aims to determine the smallest number of cluster for the daily CI profiles, while minimizing the intra-cluster distance. To achieve this, it is required to minimize the total distance between each clustered profile with respect to their centroids. Five clustering methods described previously are studied and compared. These are FCM with Euclidean

distance (FCM ED), FCM with Dynamic Time Warping (FCM DTW), K-Means with Euclidean distance (K-Means ED), K-Means with Manhattan distance (K-Means MD), and GMM. The cluster number to be evaluated is from 2 to 15. The total intra-cluster distance is calculated with Equation (22) and is used as an indicator of merit.

$$\mathrm{D_{Total}} = \sum_{i=2}^{k} \sum_{j=1}^{n} \mathrm{D}^2(x_j, v_i) \tag{22}$$

Due to the clustering algorithms contain random variables, 50 repeated tests were made and the minimum results are kept. The total intra-cluster distance with respect to different number of clusters for the four seasons is provided in Fig. 4. Compared with different clustering methods, there is a huge difference in total intra-cluster distance for Winter and Summer case. This can be explained by the level of uncertainty associated with the season. There is significantly more fluctuation in the Summer season, hence, the total intra-cluster distance will be increased. It is learnt that FCM DTW has significantly smaller total intra-cluster distance for all cases. FCM ED performs marginally better than K-Means ED in most cases, given the fact that it provides better performance in analyzing uncertainty, i.e. clusters with various sizes and shapes due to the fuzzifier used for calculation, where the profiles may belong to more than one cluster. It is worth mentioning that the intra-cluster distance for K-Means MD should not be compared with other intra-cluster distances, as ED and MD are different metrics.

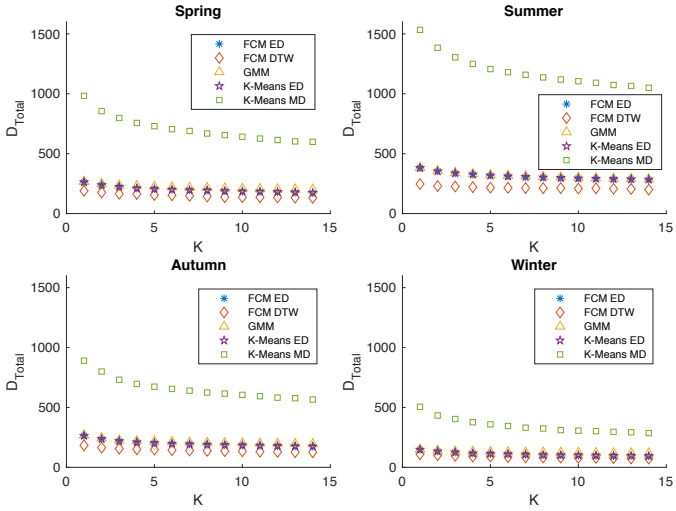


Fig. 4. Total Intra-cluster distance for different number of clusters.

ED is used to compute the intra-clusters distance for GMM. It can be seen that the total intra-cluster distance for GMM is marginally higher than K-Means methods. K-Means aims to minimize the intra-cluster distance with ED, while GMM aims at maximizing the maximum-likelihood via EM algorithm, which does not consider minimizing the distance.

B. Fuzzy Decision Making

One of the major challenges in using cluster analysis is to determine the optimal number of clusters. Traditionally, this is achieved by using the criteria such as Silhouette index, Dunn's index and Calinski-Harabasz index. In essence, these indices aim to determine some form of relationship between the within cluster cohesion and the cluster separation in order to evaluate the clusters validity. The details of these indices and the distance suitability are provided in [40]. It is learnt that the criteria are deemed ineffective for dataset with significant amount of noise or uncertainty [40]. Also, the optimal number of clusters is problem dependent and the mentioned criteria rarely provide the same results. A technique in determining the optimal number of clusters is provided in this research, by first evaluating the number of clusters that provides the best trade-off for minimizing the total intra-cluster distance. Consequently, if the resultant centroids have similar characteristics, i.e. mean and standard deviation, the similar clusters will be grouped together with the new centroid calculated. Given no prior knowledge for the selection of candidate partitions in Pareto set, an un-weighted fuzzy logic decision making strategy [47] is employed to yield the best trade-off solution.

In this paper, it is assumed that the preferences of minimizing the number of clusters and total intra-cluster distance are unbiased. Fuzzy logic decision making is formulated as follows. m is equal to 15 and stands for the number of non-dominated solutions and n is the number of objective functions. In this case, n is equal to 2 due to the objective is to minimize the number of centroids and the total intra-cluster distance. The fuzzy membership is defined below:

$$\mu_i(j) = \frac{f_i - f_i^{\min}}{f_i^{\max} - f_i^{\min}}, i = 1,2 \tag{23}$$

f_i stands for the solutions in the i^{th} objective function. The normalized membership for each solution is expressed as below:

$$\mu(j) = \frac{\sum_{i=1}^{n} \mu_i(j)}{\sum_{j=2}^{m} \sum_{i=1}^{n} \mu_i(j)} \tag{24}$$

The most satisfactory solution in this case is selected with the minimum fuzzy membership value [47]. The normalized fuzzy memberships for the four seasons, calculated with the intra-cluster distances given in Fig. 4 are presented in Fig. 5. The clustering methods show a similar trend for the normalized fuzzy membership with respect to cluster number. This explains that the optimal number of clusters are similar for the different clustering techniques.

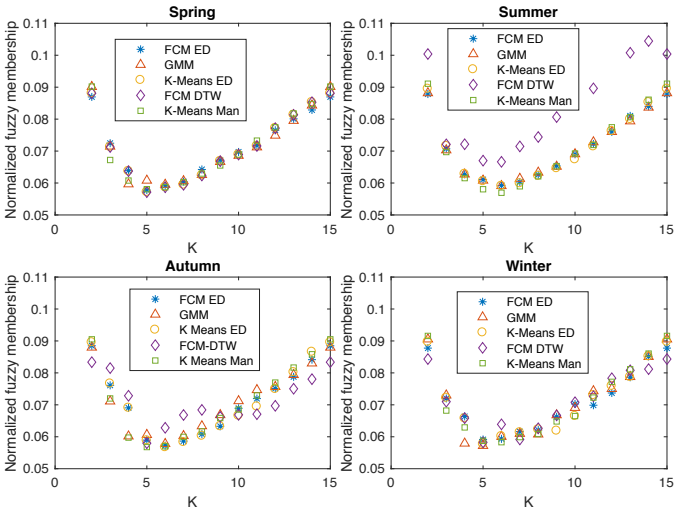


Fig. 5. Fuzzy decision making for both case studies.

Fig. 6 shows the centroids for the Winter case with different clustering techniques. The centroids for FCM ED and K-Means ED in Winter case show a similar shape. Recall the results in Fig. 4, the total distance for the two approaches are very similar. The number of representation for clear days are different. It is one centroid for GMM, two centroids for FCM ED and K-Means ED, and three centroids for K-Means MD and FCM DTW. A method is required to minimize the number of centroids with similar shapes and magnitude to reduce redundancy.

According to Fig. 7, the centroids in the Summer case show that the clustering performance of FCM ED, K-Means ED, K-Means MD and GMM are similar. It is worth noting that the clustering problem for Summer case is significantly more challenging than the Winter case. This is due to the fact that CI in Summer has more fluctuation.

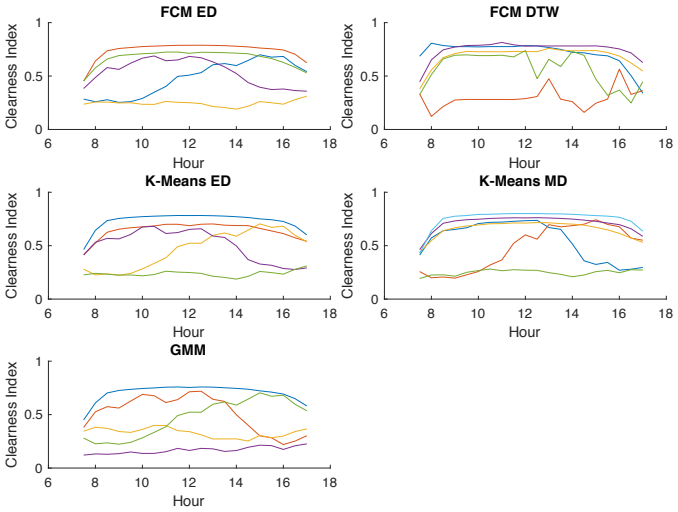


Fig. 6. Centroids for different clustering techniques in Winter case.

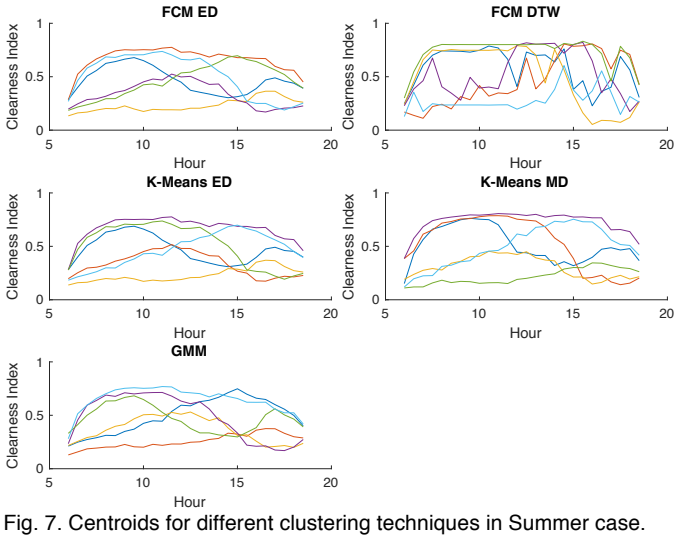


Fig. 7. Centroids for different clustering techniques in Summer case.

C. Reduction of clusters with centroid evaluation

The K-Means related clustering techniques consider solely the intra-cluster compactness, i.e. the distances between the objects and the centroids [48]. The inter-cluster separation, i.e. the distances between the centroids are not well considered during the clustering process. Redundant or similar clusters and centroids can be eliminated by evaluating the centroid's standard deviation and mean. In the first step, a mean similarity matrix and standard deviation similarity matrix, both with size K by K are constructed to determine if the clusters are similar. If the element in the matrix falls below a predefined threshold, in this case 0.1 for both variables, then the element will be set to 1. This signifies that the two clusters have similar characteristic. The corresponding elements in the two matrices with binary numbers will be multiplied together to give the similarity matrix, *M*. Once the similar clusters are grouped together, the new centroid is calculated by calculating the average of the centroids by considering the weights with the number of profiles in the original cluster. The procedure is presented in Table I.

TABLE I
ALGORITHM FOR REDUCTION OF REDUNDANT CENTROIDS

Input: $C = \{c_1, c_2, ..., c_k\}$: the set of all centroids

card: the cardinality of each cluster

Output: C' : the reduced set of centroids

1. Calculate the similarity matrix *M*
2. $s = \{1, 2, ..., k\}$
3. Find the all-zero rows of *M*

   $s' \leftarrow$ the indices of all-zero rows

   $C' = \{c_i \mid i \in s'\}$
4. $s = s - s'$
5. **while** *s* is non-empty

$$s' = \{s_i \mid i = 1 \text{ or } M_{s_1,i} = 1\}$$

$$C\_new = \frac{\sum_{i \in s'} c_i \cdot card_i}{\sum_{i \in s'} card_i}$$

$$C' = C' \bigcup C\_new$$

$$s = s - s'$$

end

Fig. 8 shows that the centroids for FCM DTW Winter case can be reduced from the original 5 centroids as shown in Fig. 6 to 3 centroids. It also presents the daily CI profiles with their respective centroids for FCM DTW in Winter case. The percentage day covered for clear days in Cluster 3 is 78%, which is the highest compared to other seasons. The black, blue and red lines in Fig. 8 and Fig. 9 give the centroids, the centroids with plus one and minus one standard deviation respectively, calculated with the association table in DBA.

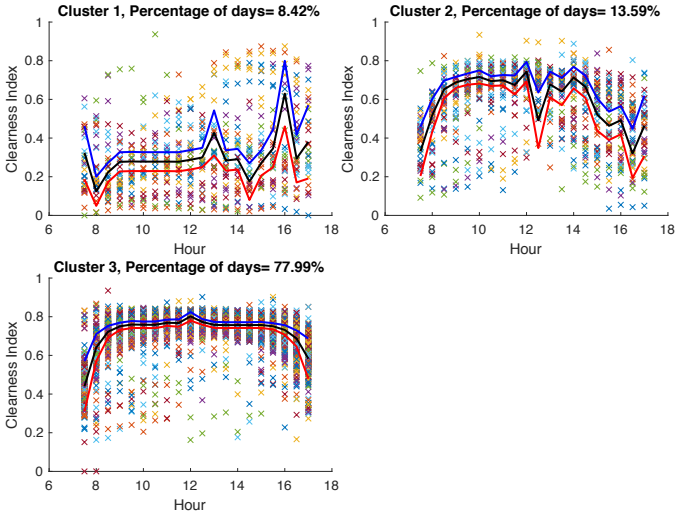


Fig. 8. CI Profiles in respective centroids for FCM DTW Winter case.

Fig. 9 presents the daily CI profiles with respect to their centroids for FCM DTW Summer case. The clear days are presented in Cluster 5, which takes into 19.10% of the days of the season. Cluster 3 shows that the perturbation takes place during the late afternoon. Cluster 6 presents the CI profile where the CI is generally low for the whole day. The clustering results display an interesting pattern and could be understood and quantified. The optimal number of clusters for the four seasons with the five clustering techniques is provided in Table II.

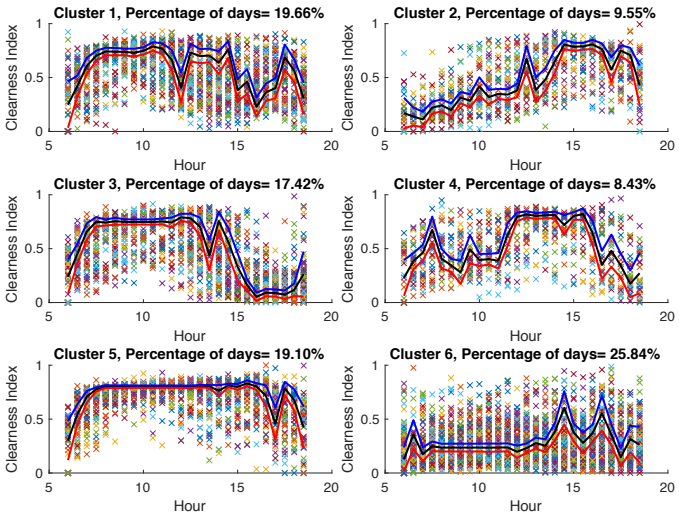


Fig. 9. CI Profiles in respective centroids for FCM DTW Summer case.

TABLE II
OPTIMAL NUMBER OF CLUSTERS

	Spring	Summer	Autumn	Winter
FCM ED	5	6	6	4
FCM DTW	4	6	5	3
GMM	6	6	6	5
K-Means ED	5	6	6	4
K-Means MD	5	6	5	4

VI. CASE STUDY: SIZING OF STAND-ALONE PV AND STORAGE SYSTEM WITH ANAEROBIC DIGESTION BIOGAS POWER PLANTS

In contrast with using the actual real-life daily solar irradiance profiles for system sizing in [49], this paper uses the daily solar irradiance profiles constructed from the cluster centroids with CIs. The daily clear-sky solar irradiance profile for Autumn, Spring, Winter and Summer are calculated with the equinoxes (20th March and 23rd Sept.), the Winter solstice (21st June) and the Summer solstice (21st Dec.) respectively. The equation for the calculation of constructed solar irradiance is given in Equation (25) below:

$$I_{construct}(t) = \mathrm{CI}(t) * I_{model}(t) \quad (25)$$

The constructed solar irradiance profiles for FCM DTW Winter and Summer cases are presented in Fig. 10 and Fig. 11 respectively. To consider the dispersion of the clustered data, the plus one and minus one standard deviations of the centroids are included for sizing purposes.

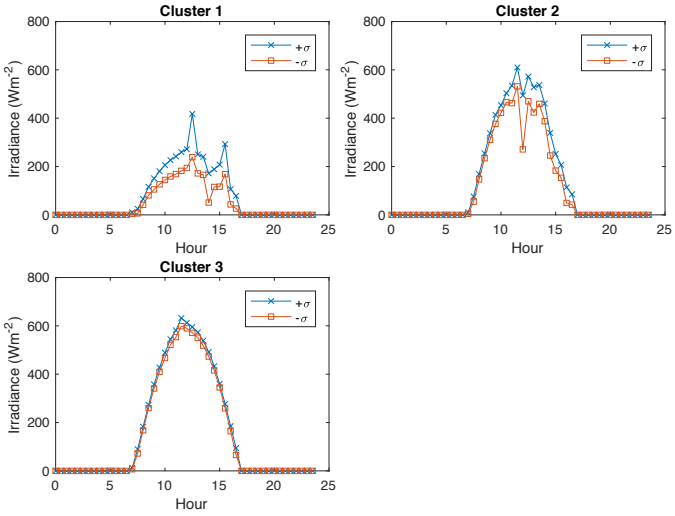


Fig. 10. Constructed irradiance from centroids in FCM DTW Winter case.

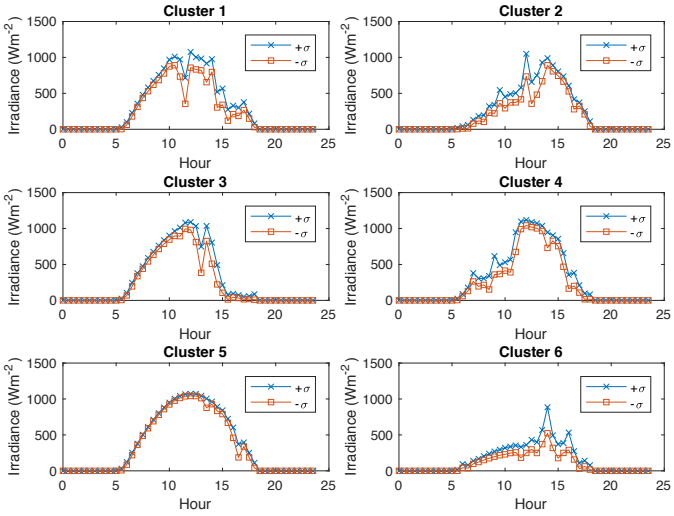


Fig. 11. Constructed irradiance from centroids in FCM DTW Summer case.

A. Sizing of Solar Panels

The required PV solar panel areas to meet the energy deficit of the solar PV hybrid energy system are determined with Particle Swarm Optimization with Interior Point Method [49]. Fig. 12 presents the panel area results with the irradiance profiles developed from five different clustering techniques.

The population of results in a form of boxplot for PV panel sizing for the four seasons are represented in Fig. 12 and the PV farm power capacities are given in Fig. 13. The calculation of PV capacity from panel area can be referred to [49]. In Fig. 13, it can be realized that FCM DTW and GMM need a required PV capacity of 5 MW, whereas FCM ED and K-Means ED need a 4 MW PV capacity, and finally K-Means MD needs a PV capacity of 3 MW. The energy balance of generation and demand are highly related to the shape and arbitrariness of the solar irradiance profile. These are better captured by GMM and FCM DTW clustering methods, which are reflected in the PV panel sizing results in Fig. 12 and the centroids in Fig. 6 and Fig. 7.

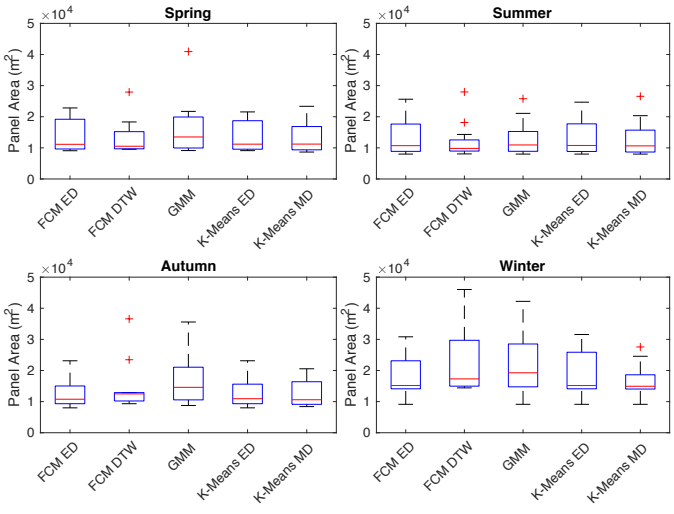


Fig. 12. Optimization results for PV panel sizing.

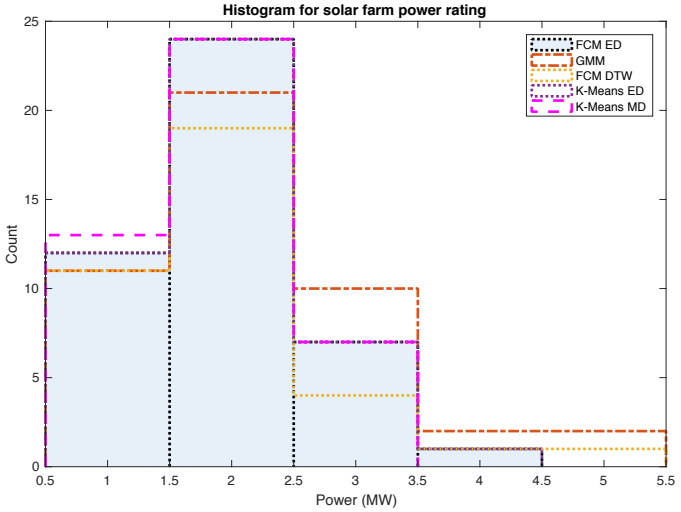


Fig. 13. PV farm rated capacity with different clustering techniques.

B. Sizing of Storage

Turning to sizing of storage, the aim is to determine the maximum energy deficit of the system with the centroid profiles. For a 5 MW solar farm, the maximum energy deficit with GMM results is 4.14 MWh and occurs in Summer. The maximum energy deficit with FCM DTW results is 3.49 MWh and also occurs in Summer, at Cluster 6 in Fig. 11 with the minus one standard deviation. To understand the implications of the energy deficit results, Fig. 14 shows the energy deficit computed with four years of real-life solar irradiance data.

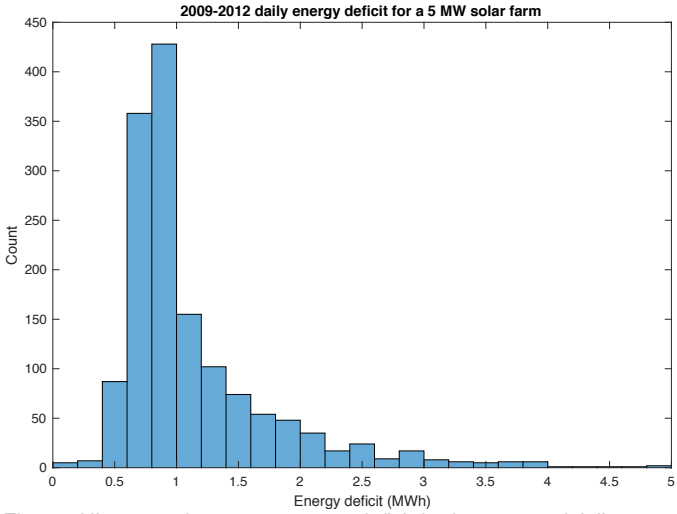


Fig. 14. Histogram for system energy deficit for four years of daily case study.

The real-life solar irradiance profile study shows that the energy deficit can reach 5 MWh. It is also worth mentioning that the total number of clusters for FCM DTW is less than GMM. The total number of sizing cases, i.e. the total count is 1457. The number of additional cases where 4.14 MWh meets the energy demand in contrast to 3.49 MWh is 16. At 3.49 MWh, 1436 cases of energy deficit are covered. The difference in energy storage capacity between FCM DTW and GMM is (4.14-3.49)/3.49 = 18.62%. By increasing the storage capacity from 3.49 MWh to 4.14 MWh, the additional cases of energy deficit covered is 16/1436 = 1.11%. This concludes that the

system can meet an additional 1.11% cases of energy deficit with an additional of 0.65 MWh storage capacity. This may not be an economical solution and the issue with energy deficit can be overcome by optimal scheduling or demand side management, which will be a future work.

VII. Conclusion and Future Work

This paper presents feature extraction for daily clearness index profiles with five different cluster analysis techniques. An optimal sizing case study for a PV system with energy storage and anaerobic digestion biogas power plants is used to compare the clustering results for PV system planning. As different to the 1457 daily irradiance profiles used in [49] for the system sizing, the data set can be represented with 36 and 46 profiles, with Fuzzy C-Means (FCM) dynamic time warping (DTW) and Gaussian mixture model clustering respectively. It is worth mentioning that the optimal number of clusters is problem dependent and may vary depending on the application.

For future work, it is possible to include an extra temporal constraint in DTW, by limiting the number of vertical or horizontal steps that the path can take consecutively. This adjustment avoids the matching of points that are very far from each other in time and, in addition, it reduces the computation cost. The fuzzifier parameter for FCM can be further explored. The centroids can be used for other planning and operation purposes for PV systems, such as optimal placement of phasor measure unit and evaluation of scheduling algorithms.

Acknowledgement

Chun Sing Lai would like to express his appreciation to Guangdong University of Technology, Guangzhou, China in giving financial support under the grant from Department of Financial and Education of Guangdong Province 2016[202]: Key Discipline Construction Programme and The Hong Kong Polytechnic University, Hong Kong, China for a three-month exchange under its postgraduate scholarship scheme in 2016.

References

[1] F. Katiraei and J. R. Aguero, "Solar PV integration challenges," *IEEE Power and Energy Magazine,* vol. 9, pp. 62-71, 2011.

[2] A. Woyte, R. Belmans, and J. Nijs, "Fluctuations in instantaneous clearness index: Analysis and statistics," *Solar Energy,* vol. 81, pp. 195-206, 2007.

[3] A. Maafi and S. Harrouni, "Preliminary results of the fractal classification of daily solar irradiances," *Solar Energy,* vol. 75, pp. 53-61, 2003.

[4] H.-T. Yang, C.-M. Huang, Y.-C. Huang, and Y.-S. Pai, "A weather-based hybrid method for 1-day ahead hourly forecasting of pv power output," *IEEE Transactions on Sustainable Energy,* vol. 5, pp. 917-926, 2014.

[5] T. Soubdhan, R. Emilion, and R. Calif, "Classification of daily solar radiation distributions using a mixture of Dirichlet distributions," *Solar Energy,* vol. 83, pp. 1056-1063, 2009.

[6] S. Kheradmanda, O. Nematollahi, and A. R. Ayoobia, "Clearness index predicting using an integrated artificial neural network (ANN) approach," *Renewable and Sustainable Energy Reviews,* vol. 58, pp. 1357-1365, 2016.

[7] A. Mellit, S. Kalogirou, S. Shaari, H. Salhi, and A. H. Arab, "Methodology for predicting sequences of mean monthly clearness index and daily solar radiation data in remote areas: application for sizing a stand-alone PV system," *Renewable Energy,* vol. 33, pp. 1570-1590, 2008.

[8] T. Ayodele and A. Ogunjuyigbe, "Prediction of monthly average global solar radiation based on statistical distribution of clearness index," *Energy,* vol. 90, pp. 1733-1742, 2015.

[9] Z. Ren, W. Yan, X. Zhao, W. Li, and J. Yu, "Chronological probability model of photovoltaic generation," *IEEE Transactions on Power Systems,* vol. 29, pp. 1077-1088, 2014.

[10] A. Woyte, V. Van Thong, R. Belmans, and J. Nijs, "Voltage fluctuations on distribution level introduced by photovoltaic systems," *IEEE Transactions on Energy Conversion,* vol. 21, pp. 202-209, 2006.

[11] J. Liu, W. Fang, X. Zhang, and C. Yang, "An improved photovoltaic power forecasting model with the assistance of aerosol index data," *IEEE Transactions on Sustainable Energy,* vol. 6, pp. 434-442, 2015.

[12] R. Kumar and L. Umanand, "Estimation of global radiation using clearness index model for sizing photovoltaic system," *Renewable Energy,* vol. 30, pp. 2221-2233, 2005.

[13] Y. Ghiassi-Farrokhfal, S. Keshav, C. Rosenberg, and F. Ciucu, "Solar power shaping: An analytical approach," *IEEE Transactions on Sustainable Energy,* vol. 6, pp. 162-170, 2015.

[14] S. Harrouni, A. Guessoum, and A. Maafi, "Classification of daily solar irradiation by fractional analysis of 10-min-means of solar irradiance," *Theoretical and Applied Climatology,* vol. 80, pp. 27-36, 2005.

[15] H. Khorasanizadeh, K. Mohammadi, and N. Goudarzi, "Prediction of horizontal diffuse solar radiation using clearness index based empirical models; A case study," *International Journal of Hydrogen Energy,* vol. 41, pp. 21888-21898, 2016.

[16] C. M. Fernández-Peruchena and A. Bernardos, "A comparison of one-minute probability density distributions of global horizontal solar irradiance conditioned to the optical air mass and hourly averages in different climate zones," *Solar Energy,* vol. 112, pp. 425-436, 2015.

[17] L. Wang, O. Kisi, M. Zounemat-Kermani, G. A. Salazar, Z. Zhu, and W. Gong, "Solar radiation prediction using different techniques: model evaluation and comparison," *Renewable and Sustainable Energy Reviews,* vol. 61, pp. 384-397, 2016.

[18] A. Sanfilippo, L. Martin-Pomares, N. Mohandes, D. Perez-Astudillo, and D. Bachour, "An adaptive multi-modeling approach to solar nowcasting," *Solar Energy,* vol. 125, pp. 77-85, 2016.

[19] L. Wang, W. Gong, M. Luo, W. Wang, B. Hu, and M. Zhang, "Comparison of different UV models for cloud effect study," *Energy,* vol. 80, pp. 695-705, 2015.

[20] K. Bakirci, "Models for the estimation of diffuse solar radiation for typical cities in Turkey," *Energy,* vol. 82, pp. 827-838, 2015.

[21] T. E. Boukelia, M.-S. Mecibah, and I. E. Meriche, "General models for estimation of the monthly mean daily diffuse solar radiation (Case study: Algeria)," *Energy Conversion and Management,* vol. 81, pp. 211-219, 2014.

[22] A. Peled and J. Appelbaum, "Evaluation of solar radiation properties by statistical tools and wavelet analysis," *Renewable Energy,* vol. 59, pp. 30-38, 2013.

[23] M. Muselli, P. Poggi, G. Notton, and A. Louche, "Classification of typical meteorological days from global irradiation records and comparison between two Mediterranean coastal sites in Corsica Island," *Energy Conversion and Management,* vol. 41, pp. 1043-1063, 2000.

[24] C. Tiba, A. N. Siqueira, and N. Fraidenraich, "Cumulative distribution curves of daily clearness index in a southern tropical climate," *Renewable Energy,* vol. 32, pp. 2161-2172, 2007.

[25] S. Buhan and Y. Özkazanç, "Wind Pattern Recognition and Reference Wind Mast Data Correlations With NWP for Improved Wind-Electric Power Forecasts," *IEEE Transactions on Industrial Informatics,* vol. 12, pp. 991-1004, 2016.

[26] C. S. Ioakimidis, L. J. Oliveira, and K. N. Genikomsakis, "Wind power forecasting in a residential location as part of the energy box management decision tool," *IEEE Transactions on Industrial Informatics,* vol. 10, pp. 2103-2111, 2014.

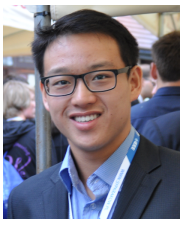
[27] M. B. Ozkan and P. Karagoz, "A novel wind power forecast model: statistical hybrid wind power forecast technique (SHWIP)," *IEEE Transactions on Industrial Informatics,* vol. 11, pp. 375-387, 2015.

[28] Skye Instruments Ltd, "SKS 1110 Pyranometer," [Online]. Available: http://www.skyeinstruments.info/index_htm_files/Pyranometer.pdf. (Visited on 12th Feb. 2017).

[29] Skye Instruments Ltd, "Solar Radiation System for Photo Voltaics," [Online]. Available: http://www.skyeinstruments.info/index_htm_files/Solar Radiation System for Photovoltaics.pdf. (Visited on 12th Feb. 2017).

[30] I. Rüedi and W. Finsterle, "The World Radiometric Reference and its quality system," in *Proc. WMO Tech. Conf. on Meteorological and Environmental Instruments and Methods of Observation (TECO-2005), Instruments and Observing Methods, Bucharest, Romania, Government of Romania, Rep*, 2005, pp. 434-436.

[31] L. Wang, G. A. Salazar, W. Gong, S. Peng, L. Zou, and A. Lin, "An improved method for estimating the Ångström turbidity coefficient β in Central China during 1961-2010," *Energy,* vol. 81, pp. 67-73, 2015.
[32] K. Scharmer, J. Page, L. Wald, M. Albuisson, G. Czeplak, B. Bourges, *et al.*, "The European solar radiation atlas Vol. 1: Fundamentals and maps," 2000.
[33] T. Hove and E. Manyumbu, "Estimates of the Linke turbidity factor over Zimbabwe using ground-measured clear-sky global solar radiation and sunshine records based on a modified ESRA clear-sky model approach," *Renewable Energy,* vol. 52, pp. 190-196, 2013.
[34] F. Kasten, "The Linke turbidity factor based on improved values of the integral Rayleigh optical thickness," *Solar Energy,* vol. 56, pp. 239-244, 1996.
[35] "NASA surface meteorology and solar energy," [Online]. Available: https://eosweb.larc.nasa.gov/cgi-bin/sse/grid.cgi. (Visited on 12th Feb. 2017).
[36] P. Hedelin and J. Skoglund, "Vector quantization based on Gaussian mixture models," *IEEE Transactions on Speech and Audio Processing,* vol. 8, pp. 385-401, 2000.
[37] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, "A robust EM clustering algorithm for Gaussian mixture models," *Pattern Recognition,* vol. 45, pp. 3950-3961, 2012.
[38] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognition,* vol. 38, pp. 1857-1874, 2005.
[39] U. Mori, A. Mendiburu, and J. A. Lozano, "Similarity Measure Selection for Clustering Time Series Databases," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, pp. 181-195, 2016.
[40] R. C. de Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Information Sciences,* vol. 324, pp. 126-145, 2015.
[41] H. Izakian, W. Pedrycz, and I. Jamal, "Fuzzy clustering of time series data using dynamic time warping distance," *Engineering Applications of Artificial Intelligence,* vol. 39, pp. 235-244, 2015.
[42] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition,* vol. 44, pp. 678-693, 2011.
[43] C. Zhu and D. Gao, "Multiple matrix learning machine with five aspects of pattern information," *Knowledge-Based Systems,* vol. 83, pp. 13-31, 2015.
[44] J. F. Kolen and T. Hutcheson, "Reducing the time complexity of the fuzzy c-means algorithm," *IEEE Transactions on Fuzzy Systems,* vol. 10, pp. 263-267, 2002.
[45] C. Bouveyron, S. Girard, and C. Schmid, "High-dimensional data clustering," *Computational Statistics & Data Analysis,* vol. 52, pp. 502-519, 2007.
[46] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis,* vol. 71, pp. 52-78, 2014.
[47] S. Agrawal, B. Panigrahi, and M. K. Tiwari, "Multiobjective particle swarm algorithm with fuzzy clustering for electrical power dispatch," *IEEE Transactions on Evolutionary Computation,* vol. 12, pp. 529-541, 2008.
[48] X. Huang, Y. Ye, and H. Zhang, "Extensions of kmeans-type algorithms: a new clustering framework by integrating intracluster compactness and intercluster separation," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 25, pp. 1433-1446, 2014.
[49] C. S. Lai and M. D. McCulloch, "Sizing of Stand-Alone Solar PV and Storage System with Anaerobic Digestion Biogas Power Plants," *IEEE Transactions on Industrial Electronics,* vol. 64, pp. 2112-2121, 2017.



Chun Sing Lai (S'11) was born in Staffordshire, U.K. He received the B.Eng. (1st Class Hons.) degree in electrical and electronic engineering from Brunel University London, U.K. in 2013.

He is currently working toward the D.Phil. degree in engineering science at the University of Oxford, U.K and is also a research associate at the School of Automation, Guangdong University of Technology, Guangzhou, China. He is chair of the IEEE Student Branch and IEEE Power & Energy Society Student Branch Chapter at University of Oxford. He was a recipient of the IEEE Systems, Man and Cybernetics (SMC) 2014 student travel grant. He is a visiting Ph.D. student at The Hong Kong Polytechnic University, Hong Kong, China. He reviews papers regularly for journals such as IEEE Trans. on Industrial Informatics and IEEE Trans. on SMC: Systems. His current interests are data analytics and energy economics for renewable energy and storage systems.

Youwei Jia (S'11-M'15) received his B.Eng. and Ph.D. degrees from Sichuan University, China, in 2011, and The Hong Kong Polytechnic University, Hong Kong, in 2015, respectively. He is now with The Hong Kong Polytechnic University, China and also with the Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou, China, as a postdoctoral fellow. His research interests include power system security analysis, cascading failures, complex network and artificial intelligence application in power engineering.

Malcolm D. McCulloch (S'88-M'89) was born in South Africa. He received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the University of Witwatersrand, Johannesburg, South Africa, in 1986 and 1990, respectively.

In 1993, he joined the University of Oxford, U.K., where he started up the Electrical Power Group (EPG), and where he is currently an Associate Professor. His work addresses transforming existing power networks, designing new power networks for the developing world, developing new technology for electric vehicles, and developing approaches to integrated mobility. He has over 100 journal and refereed conference papers, 15 patents, and four spinout companies.

Zhao Xu (M'06-SM'12) is an Associate Professor with The Hong Kong Polytechnic University, where he is currently Leader of Smart Grid research. He was previously with the Centre for Electric Power and Energy, Technical University of Denmark. He is currently an Editor for IEEE POWER ENGINEERING LETTERS, and Electric Power Components and Systems journal. He is serving as Deputy Chairman of IEEE PES joint Chapter, Hong Kong. His research interests include grid integration of renewable energies and EVs, electricity market planning and management, and big data and AI applications in power engineering.