Received May 24, 2020, accepted June 7, 2020, date of publication June 10, 2020, date of current version June 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001437

Clustering of Wind Power Patterns Based on Partitional and Swarm Algorithms

AMR A. MUNSHI, **(Member, IEEE)**
Computer Engineering Department, College of Computer and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia
e-mail: aaamunshi@uqu.edu.sa

ABSTRACT Wind power pattern clustering can potentially supply information about the effect of incorporating wind farms in smart electrical grid without in-depth analysis and studies of lengthy data. The present study investigates the most effective clustering technique and optimum number of clusters for wind power pattern data through various unsupervised clustering techniques. It also presents the introduction of Ant Colony and Bat, swarm optimization strategies in clustering wind power patterns. Three clustering algorithms from two different unsupervised techniques were concerned. A total of eight validity indices were used; Davies Bouldin, mean square error, mean index adequacy, ratio of within-cluster sum-of-squares to between-cluster-variation, Dunn, Silhouette, Xie-Beni, and clustering dispersion indicator for evaluation of the unsupervised clustering algorithms in inclusive manner. Findings depicted that Bat bio inspired clustering is comparative to K-means clustering and the most effective combination of clustering algorithm and validity index was K-means and Silhouette index, respectively. Secondly, in order to achieve improved clustering of WPP, the best clustering algorithm (K-means with Silhouette index) was modified by integrating the Silhouette index as an objective function for K-means. To check the potency of the produced wind power pattern representatives during a wind system simulation, a short wind generation prediction model is presented. The results of those cluster representatives presented promising short-term prediction results and suggest that the produced wind power pattern cluster representatives can potentially be used in other wind power pattern simulations.

INDEX TERMS Clustering, wind power, swarm methods, power patterns.

NOMENCLATURE

Cp	power coefficient
ρ	air density
v	wind speed
x_i	Wind Power Pattern
C_k	Centroid k
$\boldsymbol{C}$	Set of Centroids
N	data points
K	Number of clusters
A	number of ants
$\boldsymbol{C}^{(0)}$	initial set of centroids
Φ	pheromone matrix
a	Ant counter
S	solution vector
$\boldsymbol{S}$	solution matrix
φ	pheromone matrix component
v_i	Bat velocity
y_i	Bat position
f	frequency
λ	wavelength
r	pulse emission rate
b	Bat counter
$\tilde{\boldsymbol{\psi}}$	objective function
$\tilde{\boldsymbol{S}}$	Best solution
$\tilde{\boldsymbol{y}}$	Best position value
$\hat{d}(\Omega_i)$	average distances
$d(C_k)$	intra-cluster distance
GS	Global silhouette-width
t	Time step
Cp	power coefficient

The associate editor coordinating the review of this manuscript and approving it for publication was Jagdish Chand Bansal.

I. INTRODUCTION

The electrical systems are compelled to increase their power generation due to increase demand for electric power after the expansion in population and industry worldwide. It is anticipated that the global electricity demand is likely to reach 30-thousand Tera-watt-hours in 2030, and to reach

37-thousand Tera-watt-hours in 2040 [1]. Currently, most of this electricity is generated from fossil-fuel power resources. However, there are several monetary, environmental, and accessibility problems related to the increasing generation of electricity from fossil fuels. To mitigate the effect of these problems, renewable energy resources are capable of generating electricity. The utilization of renewable energy resources as an alternate for electricity generation is of research interest [2], [3]. The motivations behind this are to mitigate the effect of the environmental and accessibility problems with fossil fuels power resources [4]–[8]. Wind energy is considered a promising renewable resource of power and has been an area of research interest [9]–[12]. Wind turbine systems are capable of transforming the wind energy into electric power. In electrical grid systems, the advances in the wind turbine technologies, including reliability and efficiency, encourages including wind power as a major resource of power.

The quantity of perceived wind affects the power output of wind turbines. This result in operational issues and instability within the power output generated from these systems. Moreover, there is need of intensive studies and simulations of lengthy historical data will timely observations to integrate wind power in the electrical grid system. However, analyzing such lengthy data is time consuming and computationally costly. For that, enriched knowledge on the impacts of integrating those systems with the electric grid is provided by grouping wind power patterns (WPP) without intensive analysis and studies. Therefore, the present study aims to present a method for reducing burden of in-depth lengthy studies and simulations that are linked with the integration of wind power as a renewable resource in the electrical grid. For this, various unsupervised clustering techniques are used for grouping WPPs with similar patterns. Further, a representative WPP from each cluster is utilized within the simulations. Moreover, statistical information and informed decisions can be made by utilizing those wing power clustering representative patterns, and consequently, assist within the operation and integration of such systems with the electrical grid.

Intensive efforts are exerted since the last decade to cluster power patterns and produce representatives, that represent the entire power pattern data. Unsupervised learning techniques were utilized to group electrical load patterns for customer tariff formations [13]–[21], power demand prediction functions [22] and in demand side management programs [23]–[25]. Also, unsupervised grouping of power loads has been used for classifying load profiles of electricity consumers [26] and for estimating electrical loads of ships [27], [28]. In [29], planning and modeling the layout of wind farms by utilizing wind speed data to optimize wind power generation. An unsupervised learning method was developed to enhance the management of wind farms and dispatch planning in [30] and for short-term wind power prediction in [31]–[33]. In solar power studies, unsupervised learning techniques were used for planning and operating Photovoltaic produced power. These studies include mitigating effects of output fluctuation on integration of solar systems in electrical grid [34], also, in location planning for solar power plants [35]. Models to predict solar power generation by grouping historical time-series data by using representative patterns are of interest, in [36] and [37] a prediction model was developed by grouping historical data. The results of [31]–[37] indicated the accuracy of the utilized cluster representatives determines the prediction results. Therefore, major interest is exerted on investigating clustering algorithms that potentially group and present cluster representatives. The potential of using wind power as a renewable energy resource [38] motivates the investigation of applying unsupervised learning techniques to analyze WPPs. Recently, unsupervised grouping algorithms such as, K-means Ward's minimum variance, Fuzzy C-means, along with Ant Colony and Bat from bio-inspired swarm techniques were utilized for grouping Photovoltaic power. The results on Photovoltaic power showed that K-means and Bat grouping algorithms presented the overall best results on Photovoltaic power.

Various Unsupervised learning algorithms are applied to investigate which of them are appropriate for a particular task. In order to compare those algorithms, validity indices based on various metrics are utilized. Then determining the formation and number of clusters is application dependent. The present study investigates the technique that is most appropriate for clustering the wind power load by applying K-means and swarm technique algorithms. The reason for including the K-means algorithm, is due to its intensive utilization in grouping load patterns in literature. Swarm optimization techniques, which mimic swarm behavior, are recently immerging, and investigating its efficiency in grouping wind power loads is of interest. Although, every particle of the swarm moves independently, the swarm as group collaborates to attain a specified optimization goal. Swarm algorithms have proven their feasibility in various optimization problems. However, the application of such algorithms on power loads is of interest. For instance, in [20] Ant Colony swarm technique was used to group electrical load patterns and discover unusual load patterns. In addition, Bat algorithm was used to group Photovoltaic power patterns to study the effects of solar radiation on the electric grid [39]. For this, the efficiency in grouping wind power loads is investigated using K-means partitional algorithm and swarm intelligence techniques, Ant Colony and Bat algorithm. The formation of groupings of each technique is assessed by eight validity indices.

The contributions of this paper to the research field can be summarized as:

(1) Comparative study and application of swarm methods to cluster wind power patterns. This includes the first-time application of Ant Colony and Bat clustering methods to wind power loads.
(2) Systematic determination of the number of clusters for wind power patterns data, and the utilization of a comprehensive set of validity indices for comprehensive evaluation of different clustering methods.

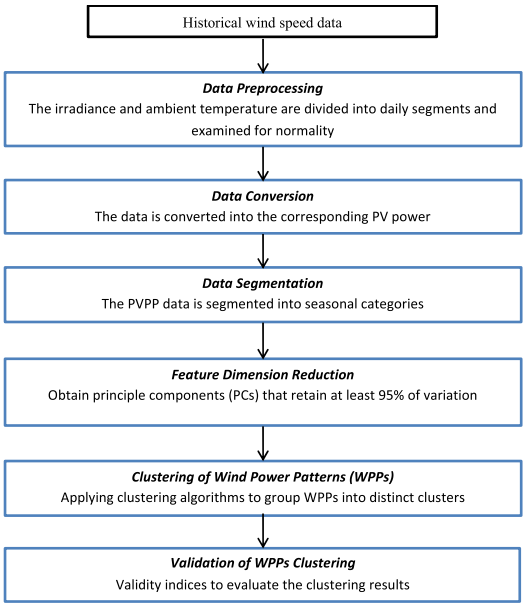

FIGURE 1. Methodology flowchart flowchart.

(3) Presenting representative wind power pattern loads that could be used in simulations and studies of interconnecting wind power systems into the electric grid with reduced data.

The rest of the study is organized as; Section II presents the layout of methodology. Section III introduces the clustering methods and Section IV introduces the validity indices assessing the clustering algorithms' results. Whereas, section V presents and discusses the clustering results in details. Section VI presents results of using the clustering representatives in a short-term prediction model. Furthermore, section VII improving the clustering results by integrating a validity index as an objective function of a clustering algorithm and using the clustering results in the short-term prediction model to improve the results. While, section VIII lists the final conclusions.

II. METHODOLOGY

The grouping of wind power patterns is achieved by applying unsupervised learning techniques on historical data. This historical data consists of wind speeds at a specified location for past years with a low-time resolution. The reason for including time-series data with low-time resolution is to capture the short fluctuations within the wind speed. In power load related applications, a 10-minute time-resolution is commonly used [34]. Fig. 1 shows the procedure of grouping the time-series wind power load data through distinct groupings.

A. DATA PREPROCESSING

The wind speed time-series data are separated into different segments with each representing a period of 24-hours. Those daily wind patterns are examined for normality, to modify or remove incorrect observations. Missing time-series observations were averaged.

B. DATA CONVERTION

The wind speed time-series data is converted to wind power load patterns at this stage. The wind power time-series loads are obtained by estimating the power that could be generated from the wind turbines by following the formula [40]:

$$P = \frac{1}{2}\rho A v^3 C_p \tag{1}$$

where Cp is the power coefficient, ρ is air density, v is wind speed, and A is swept area of turbine computed from the length of the turbine blades using the following formula:

$$A = \pi r^2 \tag{2}$$

where r is radius equal to blade length.

C. DATA SEGMENTATION

The daily output power of the wind turbines is obtained, after the conversion step. In this case, 365 WPPs are available for each year for the 24-hour period of output power. Initially, each year is divided into different seasonal categories, that helps to group data with close profiles. Secondly, each daily WPP is divided into two segments: day and night. The day segment corresponds to the time interval from 06:00 to 17:59, where the night segment corresponds to the time interval from 18:00 to 05:59. This day and night segmentation reflects the nature of wind patterns, as wind speeds during night are stronger and can be segmented as a group, whereas wind speeds are less during the day time. Hence, this segmentation can lead to subsets of data that can be grouped more efficiently.

D. FEATURE DIMENSION REDUCTION

The application of feature generation technique called principal component analysis (PCA) reduces daily features of the data [41]. For example, if the wind speed data was sampled at 10-minute time steps, there will be 1440 observations (features) per day. Those features can be reduced by applying the PCA technique. The number of the features or principal components (PC) should retain at least 95% of the variation of the original data after applying the PCA to achieve more accurate representation of the original data.

E. CLUSTERING OF WIND POWER PATTERNS

Three clustering algorithms are utilized for each subset of the wind time series data, to assign the wind power patterns into distinct groups. In such a case, there is similarity in the wind power patterns in the same cluster, and significant difference to other clusters. It is possible to obtain a representative wind power pattern (centroid) from each cluster. The entire dataset is represented using the set of centroids.

F. CLUSTERING VALIDATION

Each clustering algorithm from the previous stage forms different groupings of wind power patterns, and thus different cluster representatives. Therefore, evaluating the resulted

groupings is essential to find the most suitable algorithm for clustering the WPPs.

III. UNSUPERVISED GROUPING METHODS

The preprocessed and converted data is a set of N half-daily (i.e., day or night) WPPs that is concerned with specific time period. Each half-daily WPP contains d observations (data points). The row vector $x_i = [x_{i1}, \ldots, x_{id}]$ represents ith WPP for $i = 1, \ldots, N$. Using iterative process, the unsupervised grouping methods divide the N WPPs into different distinct clusters K with non-overlapping WPPs. Each distinct cluster could be represented by its centroid $C_k = [c_{k1}, \ldots, c_{kd}]$, for $k = 1, \ldots, K$. The matrix $\boldsymbol{C} = [C_1, \ldots, C_K]$ represents the set of centroids.

A. K-MEANS CLUSTERING

K-means algorithm is a popular unsupervised clustering technique utilized in numerous disciplines. Essentially, the K-means algorithm uses repetitive procedure to group a set of N data points into K clusters. K represents number of clusters, which is a pre-specified parameter. The representative data point presents mean of data points in a group, which is known as centroid representing the entire group. The objective of K-means clustering algorithm is to reduce a pre-specified function.

The K-means algorithm is summarized as follows [42]:

1. Initialize K data points either on some previous knowledge $\boldsymbol{C} = [C_1, C_2, \ldots, C_K]$ or arbitrarily.
2. Distances between each data point x and the centroid C is computed and each data point is assigned to the closest centroid.

$$x_i \in C_w, \text{ if } ||x_i - C_w|| < ||x_i - C_j|| \quad \text{for } i = 1, \ldots, N, j \neq w, \quad \text{and } j = 1, \ldots, K \tag{3}$$

3. Centroid of each cluster is recomputed.

$$C_k = \frac{1}{N_k} \sum_{x \in C_k} x \tag{4}$$

where N_k is the number of data points in C_k.

Steps 2 and 3 are repeated till there is no change for each cluster.

B. ANT COLONY SWARM CLUSTERING

Ant Colony algorithm is an intelligent swarm-based technique that seek out the shortest path between their nest and prey by mimicking the behavior of real ants. The ants of the swarm use a scented chemical also known as pheromone to communicate and exchange information concerning the food routes. As ants trace the same route and deposit pheromone, the route becomes more attractive and essentially followed by more ants. This cooperative behavior ends up by establishing optimized shortest route path [43]. A dedicated formulation with respect to other applications of Ant Colony clustering is used to implement Ant Colony clustering approach [20], [43], [44]. The implementation is carried out in three steps; initialization, first iteration, and successive iterations.

1: Let $i = 1$, where i denotes the row index of the normalized pheromone matrix
2: $sum = \varphi_{ik}^{(0)}$
3: Generate $rand \sim U(0, 1)$
4: **while** $sum < rand$ **do**
5: $i = i + 1$, i.e. advance to the next index;
6: $sum = sum + \varphi_{ik}^{(0)}$;
7: **end while**
8: Return i as the selected cluster for $S_{an}^{(1)}$;

FIGURE 2. Pseudo-code of biased roulette wheel.

1) INITIALIZATION

The number of clusters K and the number of ants A are defined at the initialization stage. The initial set of centroids $\boldsymbol{C}^{(0)}$ are randomly chosen from the data, based on the number of clusters K. The computation of the distances between each data point and centroid help in constructing an initial $N \times K$ pheromone matrix $\Phi^{(0)}$. The situations of division by zero are avoided by replacing null distances by a relatively small value ε:

$$r_{ik}^{(0)} = \max \{d(x_i, C_k), \varepsilon\} \tag{5}$$

Then, calculations are done for auxiliary variables based on the squared inverse of distances (between each data point and centroid) in $\Phi(0)$:

$$\varphi\prime_{ik}^{(0)} = 1/\left(r_{ik}^{(0)}\right)^2 \tag{6}$$

At the iterative stage, the components of the pheromone matrix are normalized for preventing the continuous growth of pheromone components. This normalization is achieved by dividing each auxiliary variable with the sum of auxiliary variables in the corresponding row:

$$\varphi_{ik}^{(0)} = \varphi\prime_{ik}^{(0)} / \sum_{k=1}^{K} \varphi\prime_{ik}^{(0)} \tag{7}$$

2) FIRST ITERATION

Each ant produces a solution route vector $S_{a^{(1)}}$ based on a probabilistic criterion using the pheromone matrix $\varphi_{ik}^{(0)}$ for the ants $a = 1, \ldots, A$. The $\boldsymbol{S}^{(1)}$ matrix is an $N \times A$ matrix containing solution and here each data point is assigned for ant a. This produced solution corresponds to the cluster that the data point is grouped to. The biased-roulette wheel criterion with the probability of selection proportional to the row values of the pheromone matrix Φ is used to determine the produced solutions. Fig. 2 represents the pseudo-code of the biased roulette wheel [44].

The average of data points assigned to same cluster helps in obtaining the set of centroids $\boldsymbol{C}_{a^{(1)}}$ for each $S_{a^{(1)}}$ vector.

Consequently, this helps in obtaining A clustering solution vectors and centroids. An objective function based on the sum of square errors is used to evaluate each clustering solution:

$$\psi_a^{(m)} = \sum_{k=1}^{K} \sum_{x_i \in c_{ak}} ||x_i - C_{ak}||^2, \quad \text{for } a = 1, \ldots, A \tag{8}$$

where m is the iteration number.

The lower values indicate well-formed clusters of data in this objective function. The best sets include set of $S_{a^{(m)}}$ and $\boldsymbol{C}_{a^{(m)}}$ that lead to the lowest values of objective function. These sets are defined as $\tilde{\boldsymbol{S}}_a^{(m)}$ and $\tilde{\boldsymbol{C}}_a^{(m)}$ respectively that replaces the initial ones. Then, the pheromone matrix is updated to $\Phi^{(1)}$ using the following equation:

$$r_{ik}^{(m)} = \max \left\{ d(x_i, \tilde{\boldsymbol{C}}_{ak}^{m}) \right\} \tag{9}$$

The auxiliary variables $\varphi'^{(1)}_{ik}$ are computed by adding a pheromone reinforcement term to (6) that is entirely different from the initialization stage:

$$\varphi\prime_{ik}^{(m)} = \varphi\prime_{ik}^{(m-1)} + 1/\left(r_{ik}^{(m)}\right)^2 \tag{10}$$

The pheromone components are prevented from increasing continuously in the iterative stage by normalizing the components of $\varphi'^{(m)}_{ik}$ of the pheromone matrix:

$$\varphi_{ik}^{(m)} = \varphi\prime_{ik}^{(m)} / \sum_{k=1}^{K} \varphi\prime_{ik}^{(m)} \tag{11}$$

3) SUCCESSIVE ITERATIONS

The solution and centroid set for the first ant ($a = 1$) are recorded as $\tilde{\boldsymbol{S}}_a^{(m)}$ and $\tilde{\boldsymbol{C}}_a^{(m)}$, respectively to avoid losing the best solution sets. Whereas, the solution vectors $S_{a^{(m)}}$ are produced as in the first iteration, based on the roulette-wheel criterion for the successive ants $a = 2, \ldots, A$, at each iteration m.

The following operations are similar to the ones mentioned in the first iteration, including the set of clusters $\boldsymbol{C}_a^{(m)}$, evaluation of each ant solution and getting the best solution vector $\tilde{\boldsymbol{S}}_a^{(m)}$ and set of centroids $\tilde{\boldsymbol{C}}_a^{(m)}$. The pheromone matrix $\Phi^{(m)}$ is updated by following (9)-(11) at the end of each successive iteration.

4) STOP CRITERION

An effective criterion is to stop when no noticeable improvement in the objective function of heuristic methods such as swarm methods. Also, this algorithm adopts a pre-defined maximum number of iterations for preventing excessive computation time.

5) FINAL CLUSTERING SOLUTION

Recording the index of the highest value in each row of the pheromone matrix Φ helps in obtaining the final assignment of data points to clusters. Whereas, averaging the data points assigned to that cluster determines the final centroids for each formed cluster.

C. BAT CLUSTERING

The Bat algorithm is considered as an intelligent swarm technique inspired by the echolocation behavior of micro-bats. A loud sound-pulse is emitted by micro-bats, who then listen for the echo to bounce back from surrounding objects. Micro-bats rely on this echo for computing the distance of an object. There is increase and reduction in the sound-pulse loudness as it approaches towards its prey. The Bat algorithm mimics the micro-bats' behavior as it searches for prey. The idealization proposed by [46] is used to model this algorithm that is summarized as follows: each virtual micro-bat flies randomly with a velocity v_i at position y_i with a fixed frequency f, varying wavelength λ and varying loudness A at step t. They adjust their wavelength, loudness from a positive value A_0 to a minimum constant value A_{min} and adjust their pulse emission rate $r\epsilon$ [0,1] as they find prey. The search is intensified by exploitation and the selection of the best current position continues until a pre-defined criterion is obtained. Bat clustering algorithm [47] is divided into three stages:

1) INITIALIZATION

At initialization stage, the number of clusters K and number of bats, $b = 1, \ldots, B$, are defined. Each bat (b) is then allocated an emission rate r, a frequency value $f_b \in [f_{min}, f_{max}]$, loudness value $A_b \epsilon\ [A_0, A_{min}]$ and a random solution vector (S_b) to represent the cluster at which each data point is assigned. Accordingly, an $N \times B$ matrix $\boldsymbol{S}$ is formed; where N is the number of data points and each column vector represents the bth bat solution vector. The initial set of centroids $\boldsymbol{C}_{b^{(0)}}$ are computed from each S_b vector by taking average of the data points assigned to that cluster. The set of centroids $\boldsymbol{C}$ corresponds to the bat's position y.

2) EXPLOITATION

Based on the sum of square errors, each solution vector S_b is assessed by an objective function:

$$\psi_b^{(t)} = \sum_{k=1}^{K} \sum_{x_i \in c_{bk}} ||x_i - C_{bk}||^2, \quad \text{for } b = 1, \ldots, B \tag{12}$$

In the previous objective function, lower values indicate well-formed clusters of data. The lowest value is defined as $\tilde{\psi}$ and the corresponding S_b and y_b are considered to be the best sets, defined as $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{y}}$, respectively. Now, the frequency f_b is adjusted, the velocity v_b and the positions y_b of the bats is updated to generate new B solution vectors:

$$f_b = f_{min} + (f_{min} - f_{max})\,\beta \tag{13}$$

$$v_b^{(t)} = v_b^{(t-1)} + \left[y_b^{(t)} - \tilde{\boldsymbol{y}} \right] f_b \tag{14}$$

$$y_b^{(t)} = y_b^{(t-1)} + v_b^{(t)} \tag{15}$$

where β denotes a random number between [0,1].

A new random variable $\beta 2$ between [0, 1] is generated and a new search solution is generated around $y_b^{(t)}$ if it is greater than the pulse rate r:

$$y_b^{(t)} = \tilde{\boldsymbol{y}} + \varepsilon \mathcal{A} \tag{16}$$

where; ε is a small value that directs and strengthens the random walk (exploitation). $\mathcal{H}$ is a randomly generated normal distribution vector of the same size as y.

Now, evaluations are carried out to determine the distance between each data point and its position, along with each data point that is assigned to the lowest distance solution (i.e., the nearest centroid). Then the corresponding objective function value is computed $\psi_b^{(t)}$. Finally, a random number $\beta 3$ between [0, 1] is generated and the new position for that bat and, increase r and reduce A is accepted, if it is less than the loudness A and the computed fitness $\psi_b^{(t)}$ is less than $\psi_b^{(t-1)}$, accept.

3) UPDATING CLUSTERING RESULTS

In this stage, $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{y}}$ are updated, if one of the generated positions improves the objective function.

4) STOP CRITERION

The process of stop criterion continues until a pre-defined maximum number of iterations M is obtained. The final set of centroids $\boldsymbol{C}$ is obtained from $\tilde{\boldsymbol{y}}$ for each grouping of data points.

IV. CLUSTERING VALIDITY INDICES

The clusters obtained by the aforementioned clustering techniques require validation. To accomplish this, there is need to evaluate formation of the clusters produced by the clustering algorithms using the validity indices. Validity indices examine the compactness and separation of the produced clusters from a certain perspective and present an index value that expresses how well the clusters are partitioned [48]. The optimum number of clusters and the algorithm that performs best on WPPs is investigated for assessing the results of the aforementioned clustering algorithms. A number of validity indices based on various metrics and indicators are used for investigating to determine the optimum number of clusters. The validity index's best value at the elbow-point expresses the utilized technique of finding the optimum number of clusters. Following section present the definitions of eight validity indices are following:

A. DAVIES-BOULDIN INDEX (DBI)

The function of the ratio of the sum-of-within-cluster-scatter to between-cluster-separation is known as Davies-Bouldin index [49]. It identifies clusters with low inter-connectivity and high intra-connectivity and is defined as:

$$DBI = \frac{1}{K}\sum_{i=1}^{K}\max_{j,i\neq j}\frac{\hat{d}(\Omega_i)+\hat{d}(\Omega_j)}{d(C_i,C_j)} \tag{17}$$

where $\hat{d}(\Omega_i)$ and $\hat{d}(\Omega_j)$ are the average distances between the data points in cluster i and j to their respective cluster centroid. The distance between the centroids of clusters i and j, respectively is $d(C_i, C_j)$. The compact clusters and large distances between cluster centroids are implied by a lower Davies-Bouldin value.

B. DUNN INDEX

The Dunn index is obtained by identifying clustering scheme as a ratio between the minimal inter-cluster distances and maximal intra-cluster distance [50]:

$$Dunn = \min_{1\leq i\leq C}\left\{\min_{\substack{1\leq i\leq C\\ j\neq i}}\left\{\frac{d\left(C_i,C_j\right)}{\max_{1\leq k\leq C}\left(d\left(C_k\right)\right)}\right\}\right\} \tag{18}$$

where $d(C_k)$ is the intra-cluster distance of cluster k. Compact and well-separated clusters are implied by large values. The largest value as an optimum number of clusters is presented, based on the number of clusters selected for an algorithm.

C. SILHOUETTE INDEX

The Silhouette index (SI) [51] assigns a quality measure to each data point in the cluster C_k, which is called the silhouette-width. In cluster C_k, the silhouette-width is a confidence indicator on the membership of the ith data point. The approach helps in calculating the silhouette-width for each data point, average silhouette-width for each cluster, and the average silhouette-width for the whole data. The following formula defines he silhouette-width:

$$s(i) = \frac{b(i)-a(i)}{\max\left(a(i),b(i)\right)} \tag{19}$$

where $a(i)$ is the average distance between the ith data point and all data points in the same cluster of cluster C_k. $b(i)$ is the minimum average distance between the ith data point and all data points not included in the same cluster. The $s(i)$ value varies between $-1 \leq s(i) \leq 1$. A value closer to 1 implies that the data point i is classified to the right cluster, whereas a value close to -1 implies the misclassification of assigning that data point. A value close to 0, indicates that a data point contained within one cluster is at an equal distance from another cluster and could be contained within either cluster. The average silhouette-width that represents the heterogeneity of a given cluster C_k is computed by:

$$S_j = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{20}$$

where n is the number of data points in $s(i)$. The global silhouette-width (GS) is computed by:

$$GS = \frac{1}{K}\sum_{i=1}^{K} S_j \tag{21}$$

The clustering formation that presents the maximal GS value can be chosen as the optimal number of clusters.

D. XIE-BENI INDEX

The Xie-Beni (XB) index includes the U matrix and data set for assessing the partitioning of fuzzy algorithms. It is also utilized for validating clustering of crisp partitioning algorithms. The XB index can be defined as the ratio of the

total variation to the minimum separation of the clusters that is calculated using following formula [48]:

$$XB = \frac{1}{N}\left(\frac{\sum_{j=1}^{K}\sum_{i=1}^{N} U_{ij}||C_j - x_i||}{\min\limits_{i\neq j}||C_i - C_j||}\right) \tag{22}$$

The compact and well-separated clusters are indicated by lower values of XB.

E. MEAN SQUARE ERROR OR ERROR FUNCTION (J)

The distance of each data point from its cluster centroid with the same weight values is expressed via J function [13]:

$$J = \frac{1}{N}\sum_{i=1}^{N} d\left(x_i, C_j\right), \quad \text{for } j = 1, \ldots, K \tag{23}$$

F. MEAN INDEX ADEQUACY

The mean distances between each data point assigned to the same cluster (Ω_k) and its centroid are computed through mean index adequacy (MIA) [13]:

$$MIA = \sqrt{\frac{1}{K}\sum_{i=1}^{K} d\left(\Omega_k, C_k\right)} \tag{24}$$

G. CLUSTERING DISPERSION INDICATOR

The ratio of the mean intra-distance between data points in the same cluster ($\hat{d}\left(\Omega_k\right)$) and the intra-distance between the cluster centroids ($\hat{d}\left(C\right)$) is defined as the clustering dispersion indicator (CDI) [13]:

$$CDI = \frac{1}{\hat{d}(C)}\sqrt{\sum_{i=1}^{K} d\left(\Omega_k\right)} \tag{25}$$

H. RATIO OF WITHIN-CLUSTER SUM-OF-SQUARES TO BETWEEN-CLUSTER-VARIATION (WCBCR)

The sum of squared distances between each data point in the data set and its centroid, and the distances between centroids is computed through WCBCR [53]:

$$WCBCR = \frac{\sum_{k=1}^{K}\sum_{x_i\in C_k} d\left(x_i, C_k\right)}{\sum_{1\leq q<p}^{K} d\left(C_p, C_q\right)} \tag{26}$$

The monotonously decreasing indices with the increase in number of clusters represent the main characteristics of J, MIA, CDI and WCBCR. An additional adequacy measure is considered in this study to calculate the aforementioned validity indices, including the number of dead clusters, for which the sets are empty. It is important to note that the validity indices are only computed for partitions without dead clusters in this study.

V. SIMULATION STUDIES

The methodology used in this study was applied on data concerning past years from 2012 to 2014 with 10-minute time steps of wind speeds from the National Wind Technology Center (NREL) [54]. The obtained data was located at latitude of 39.91°N and longitude of 105.235°W. Accuracy is likely to improve through the wind speed data with such time resolution (10-minute) due to the auto-correlation coefficients resulting in greater positive values as compared to those obtained from data with lower time resolutions. A total of 144 observations per day are obtained through the 10-minute time resolution. Data examination was conducted for normality to either amend or remove unusual and error recorded observations of wind speeds. For instance, three days of the year 2012 were removed as the wind speed values were "-9999" for the entire day. Also, there were several days that had missing observations and were removed. The data resulted from the data pre-processing step were 1085 10-minute daily wind speed data points for the three consecutive past years. Later, the data was converted to power time-series with respect to the KW3 wind turbine data sheet [55], which resulted in 1085 row vectors of WPPs. In the next step, the data was segmented into seasonal categories that include fall, winter, spring, and summer. Each daily time-series pattern was separated into day and night segments i.e., 72 observations for day and 72 observations for night (a total of 144 observations/day). This results in achieving data sets with close profiles. The data was normalized and the PCA method [41] was used to reduce the dimensionality of the eight on-hand data sets while retaining at least 95% of the total variance to reduce the dimensionality of the data sets (Table 1). 50 executions on each data set were carried-out from 2 up-to 20 clusters, for each of the aforementioned clustering methods. Among the 50 executions of each clustering algorithm, the best result for each validity index was recorded.

TABLE 1. The four seasons' data after applying The PCA method 'to reduce the dimensionality of the data while retaining at least 95% of the total variance.

Season	Day (72)	Night (72)
Fall	18	18
Winter	17	19
Spring	27	21
Summer	24	21

A. IMPLEMENTATION OF K-MEANS

The K-means clustering algorithm was applied with 50 replicates, where for each replicate a new set of initial centroids were chosen. Recordings were obtained for the cluster formation with the lowest intra-cluster distances.

B. IMPLEMENTATION OF ANT COLONY

The parameters on the number of ants in the initialization step and successive steps were done according to [39]. 50 repetitions were carried on to implement the algorithm with $A = 50$ and the clustering formations (solutions) to record the best validity indices values.

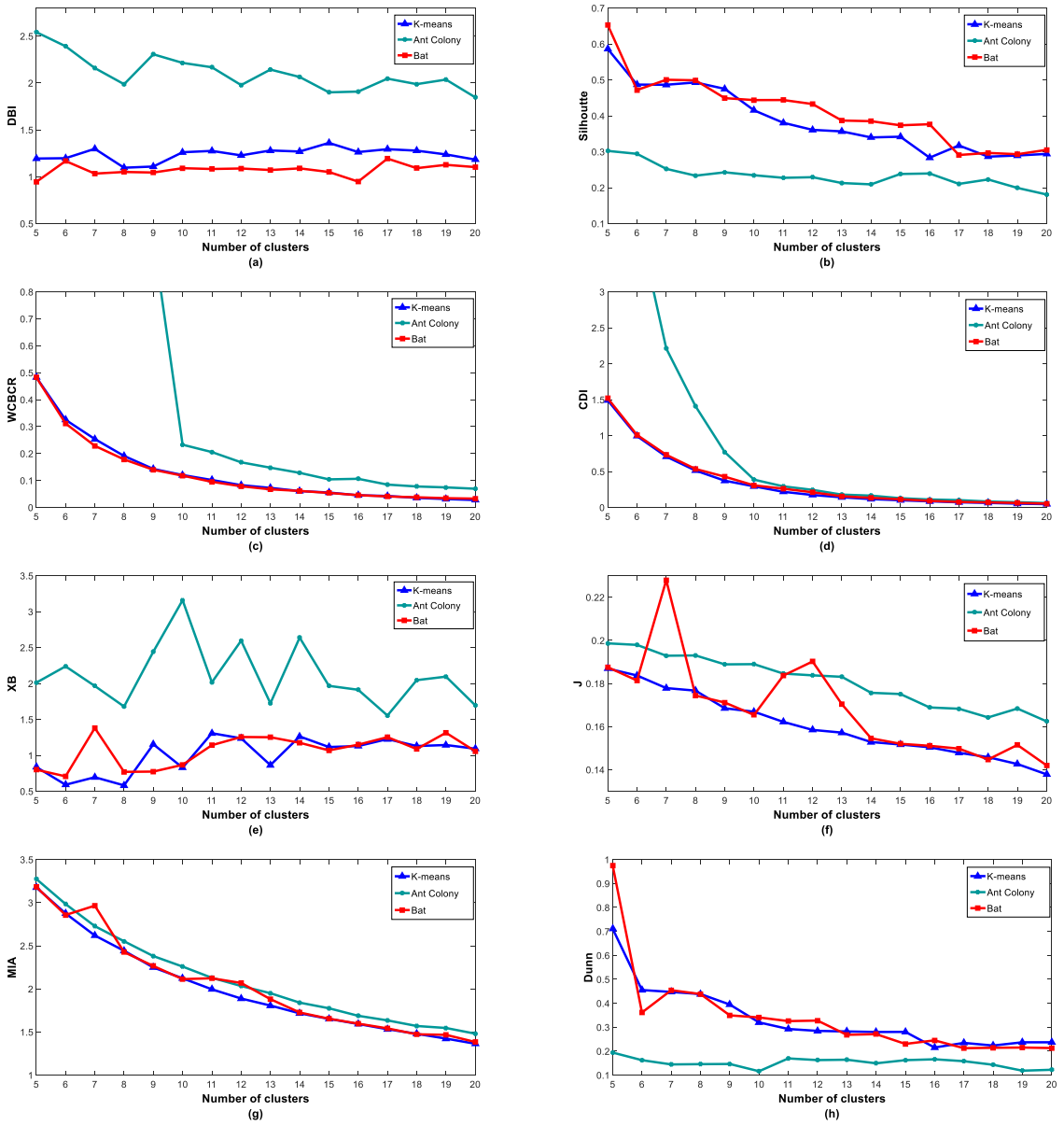

FIGURE 3. The best results of each clustering method for the spring day time data set of WPPs for 5 to 20 clusters. (a) DBI. (b) SI. (c) WCBCR. (d) CDI. (e) XB. (f) J. (g) MIA. (h) Dunn.

C. IMPLEMENTATION OF BAT

The Bat algorithm was executed with $B = 50$, $A = 0.5$, $r = 0.5$, $f_{min} = 0$, $f_{max} = 0.9$ and $M = 50$ [39]. The algorithm was performed by carrying out 50 repetitions and the solutions giving the best validity values were recorded.

D. COMPARISONS OF CLUSTERING ALGORITHMS AND VALIDITY INDICES

Fig. 3 and Fig 4 present the best results for each of the aforementioned clustering algorithms regarding the spring data. The K-means had the smallest values for the mean-square-error J. Also, it had the best values for Dunn on most partitions and competitive values for the other validity indices. While Ant Colony presented the overall worst results, moreover, it was not included for some partitions in the CDI and WCBCR index plots as it produced abnormal values compared to the other clustering algorithms. Therefore, Ant Colony will not be considered for further analysis. It should be noted that Ant Colony methods have been successfully used in clustering other types of time-series load data. Bat algorithm presented a competitive behavior with K-means on DBI, Silhouette, WCBCR, CDI and MIA.

The optimum number of clusters cannot be explicitly observed, by comparing the measures of all validity indices with each other. The observation required validity indices that present adequate measures for K-means and Bat algorithm. Also, the task of determining the best clustering algorithm and optimum number of clusters is complicated

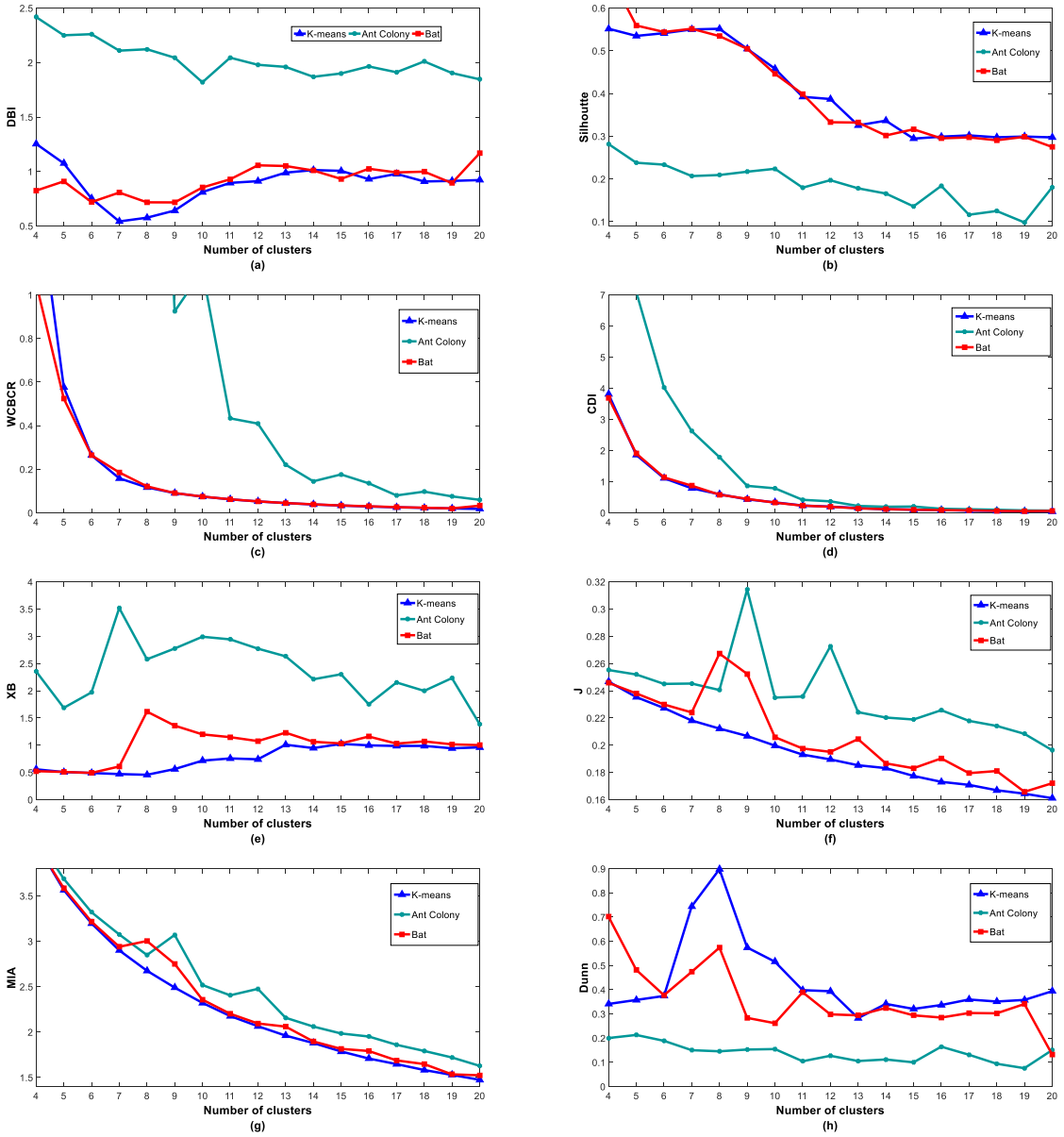

FIGURE 4. The best results of each clustering method for the spring night time data set of WPPs for 5 to 20 clusters. (a) DBI. (b) SI. (c) WCBCR. (d) CDI. (e) XB. (f) J. (g) MIA. (h) Dunn.

by the elbow points present in DBI, Dunn and XB. For that, considering validity indices that have adequate measures for K-means and Bat algorithm and less elbow points will result in considering the Silhouette and DBI validity indices. The remaining validity indices that include J, MIA, CDI, and WCBCR) are decreasing functions, monotonically. The J and MIA only consider the compactness of the formed clusters without taking the separation into account; therefore, they were excluded. They were only examined in the presence of an explicit elbow point that needs to be observed. The performance of CDI and WCBCR indices enhanced as there is increase in the number of clusters. Moreover, with respect to indices, both clustering algorithms (K-means and Bat) had relatively close measures. The use of WCBCR is slightly better as compared to CDI as the result of involving the distances of input data from the representative clusters and distance between clusters covering the CDI and J characteristics [14]. For that, the WCBCR will be considered for further analysis. To this extent, the best clustering algorithm and optimum number of clusters cannot be sufficiently determined, considering the Silhouette, DBI and WCBCR validity indices. For this, the compactness and separation of the partitioning of each clustering algorithm is investigated to find out the best combination of clustering algorithm and validity index producing the utmost compact and separate partitioning of WPPs.

There is correspondence in the number of optimum clusters and the elbow point of the respective curve [19], [26], [27].

TABLE 2. Validity indices, Compactness and separation values for K-means and Bat on Elbow-Points (Spring day time).

Validity index (VI)	5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	1.194	0.349	1.675	1.198	0.348	1.486	1.298	0.352	1.388	1.096	0.347	1.357	1.109	0.315	1.397	1.261	0.331	1.241
SI	0.586	**0.346**	1.614	0.487	0.362	1.389	0.486	0.380	1.381	0.493	0.357	1.349	0.475	0.381	1.266	0.415	0.364	1.289
WCBCR	0.483	0.349	1.675	0.325	**0.275**	1.589	0.254	0.342	1.437	0.191	**0.296**	1.392	0.143	0.315	1.397	0.120	0.304	1.307
Bat																		
DBI	0.944	0.351	**1.707**	1.168	0.346	1.585	1.033	**0.294**	**1.612**	1.051	0.299	**1.475**	1.045	0.344	1.325	1.091	**0.294**	1.334
SI	0.653	0.351	**1.707**	0.472	0.287	**1.691**	0.500	0.352	1.437	0.499	0.349	1.391	0.449	0.324	1.407	0.444	0.306	**1.343**
WCBCR	0.483	0.349	1.675	0.312	0.287	**1.691**	0.229	**0.294**	**1.612**	0.178	0.299	**1.475**	0.140	**0.302**	**1.429**	0.118	**0.294**	1.334

However, elbow point cannot be explicitly detected as shown in Fig. 3. The figure clearly shows that the elbow points are in the range of 5 to 10 for the spring day time WPP data. For that, evaluations are done for the compactness and separation from 5 to 10 clusters for K-means and Bat algorithms. It should be noted that lower values of DBI and WCBCR indicate better formation of clusters, whereas, greater Silhouette values indicate better clustering. The validity index value and the associated compactness and separation for the day-time and night-time data sets for spring are illustrated in Table 2 and III. Whereas, the detailed results for validity indices from two to twenty clusters are shown in Appendix A. The best compactness and separation values for each partitioning (column-wise) are in bold, and the best value for each validity index (row-wise) is highlighted in yellow. It is consistently observed that the compactness is low and separation is high on those elbow-point partitions, when the Silhouette validity index value is high. Moreover, the compactness and separation for Silhouette were mostly best with K-means. Thus, the higher Silhouette validity index values on K-means indicate well compact and separated clusters of WPPs. To validate that the optimum number of clusters has been chosen, a systematic method should be used. For that in this work, the L-method [56] is adopted. The L-method finds the possible lines that fit the curve of the number of clusters and validity index. Then the intersection of the two lines with the least RMSE are chosen and the optimum number of clusters (elbow point) is determined. Fig. 5, illustrates the how number of clusters for K-means with respect to the Silhouette validity index was determined as five. The observations on the spring night data set were similar (Fig. 4), and the elbow points utilizing the L-method were observed to be in the range of 4 to 9 clusters with respect to the validity indices. It can be observed that the optimum number of clusters using the L-method presents the overall best combination of compactness and separation among all cluster partitioning (Table 2–Table 7). Also, the observations on the results for the remaining data sets of the WPP data are presented in Table 4 to Table 7, Table 18, Table 19, and are similar to those discussed. Accordingly, K-means and Silhouette are considered as the best combination of clustering algorithm and validity index producing the overall best results of compactness and separation between WPP clusters, respectively.

The eight cluster representatives from the K-means for the spring night time data are presented in Fig. 6(a).

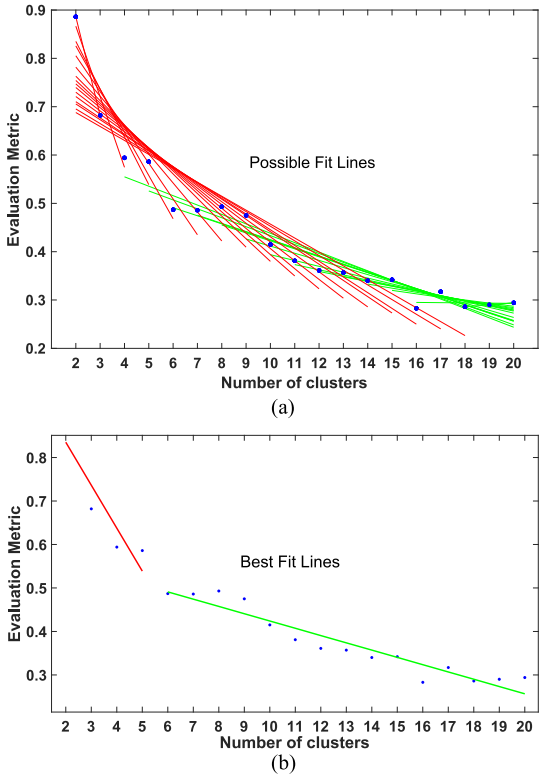

FIGURE 5. Finding the optimum number of clusters using the L-method. (a) all posible lines that fit the curve. (b) best fit.

Fig. 7 presents the representative WPPs with their confidence limits of the variations for the eight clusters of spring day-time produced by K-means. It also represents the intermediate area between the confidence limits with a probability of occurrence with 70%, which assumes normal distribution. The WPPs within a cluster have close power profiles. For instance, cluster #7 contains WPPs with poor output power. Also, clusters #1 to #4 contain single WPPs that are uncommon.

Fig. 6 shows the statistical information about the frequency of occurrence of clusters having power patterns with interesting profiles. In wind farm related simulations, the cluster representatives represent all other WPPs in that cluster group. Moreover, WPPs are used to conduct thorough analysis of a particular cluster when required. This help to plan installation of wind turbines and predict the performance of wind

TABLE 3. Validity indices, Compactness and separation values for K-means and bat on Elbow-Points (Spring night time).

Validity index (VI)	4 Clusters			5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	1.254	0.747	1.713	1.075	0.541	2.024	0.752	0.423	**2.269**	0.542	**0.186**	**2.397**	0.576	**0.197**	**2.302**	0.642	**0.236**	**2.177**
SI	0.552	0.747	1.713	0.535	0.541	2.024	0.541	0.423	**2.269**	0.550	**0.186**	**2.397**	0.552	**0.197**	**2.302**	0.505	**0.236**	**2.177**
WCBCR	1.536	0.747	1.713	0.577	0.541	2.024	0.264	0.423	**2.269**	0.159	**0.186**	**2.397**	0.117	**0.197**	**2.302**	0.091	**0.236**	**2.177**
Bat																		
DBI	0.824	**0.586**	**2.077**	0.910	**0.338**	**2.177**	0.720	**0.272**	2.240	0.807	0.449	1.986	0.717	0.231	2.243	0.716	**0.236**	**2.177**
SI	0.683	**0.586**	**2.077**	0.559	**0.338**	**2.177**	0.544	0.511	1.927	0.552	0.253	2.250	0.534	0.239	2.231	0.505	**0.236**	**2.177**
WCBCR	1.056	0.601	2.148	0.525	**0.338**	**2.177**	0.264	0.423	**2.269**	0.186	0.253	2.250	0.122	0.231	2.243	0.091	**0.236**	**2.177**

TABLE 4. Validity indices, Compactness and separation values for K-means and bat on Elbow-Points (Winter Day Time).

Validity index (VI)	5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	1.020	**0.560**	**1.979**	1.030	0.517	1.805	0.856	0.399	2.067	0.879	**0.319**	**2.073**	0.841	0.337	**1.982**	0.906	0.325	1.871
SI	0.707	**0.560**	**1.979**	0.613	0.674	1.534	0.597	**0.388**	2.025	0.604	**0.319**	**2.073**	0.539	0.337	**1.982**	0.518	0.422	1.796
WCBCR	0.685	**0.560**	**1.979**	0.441	**0.493**	1.797	0.217	0.398	**2.069**	0.148	**0.319**	**2.073**	0.112	**0.327**	1.978	0.087	**0.272**	**1.964**
Bat																		
DBI	1.124	0.681	1.442	1.082	0.534	**1.810**	0.954	0.511	1.705	0.953	0.531	1.734	0.998	**0.327**	1.978	1.059	0.499	1.613
SI	0.695	0.735	1.504	0.632	0.679	1.544	0.597	0.531	1.712	0.597	0.528	1.780	0.531	0.541	1.688	0.489	0.418	1.697
WCBCR	0.987	0.723	1.639	0.457	0.534	**1.810**	0.290	0.518	1.814	0.214	0.528	1.780	0.112	**0.327**	1.978	0.121	0.428	1.706

TABLE 5. Validity indices, Compactness and separation values for K-means and bat on Elbow-Points (Winter Night Time).

Validity index (VI)	4 Clusters			5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	1.167	0.499	1.085	1.072	0.504	**1.121**	1.171	0.478	1.072	1.216	0.462	1.022	1.151	0.390	1.149	1.150	0.369	1.099
SI	0.647	0.499	1.085	0.620	0.504	**1.121**	0.619	0.488	1.079	0.567	0.446	0.914	0.510	0.369	1.030	0.511	0.390	1.130
WCBCR	1.736	0.499	1.085	0.937	0.504	**1.121**	0.637	0.480	1.094	0.488	0.456	1.041	0.292	0.390	1.149	0.216	0.369	1.154
Bat																		
DBI	0.995	**0.372**	1.138	1.162	**0.475**	0.958	1.086	**0.393**	**1.136**	1.125	0.455	1.067	1.063	0.368	1.044	1.050	0.398	1.181
SI	0.664	**0.372**	1.138	0.633	**0.475**	0.958	0.638	**0.393**	**1.136**	0.583	0.390	**1.136**	0.514	**0.323**	1.099	0.514	0.328	1.056
WCBCR	1.612	0.385	**1.172**	1.012	0.501	1.085	0.637	0.480	1.094	0.418	**0.373**	1.130	0.287	0.389	**1.162**	0.205	**0.309**	**1.187**

TABLE 6. Validity indices, Compactness and separation values for K-means and bat on Elbow-Points (Fall Day Time).

Validity index (VI)	5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	0.855	**0.616**	**1.536**	1.107	0.619	1.439	0.941	0.628	1.481	0.912	0.518	**1.607**	0.995	0.484	1.510	0.924	**0.416**	**1.636**
SI	0.723	**0.616**	**1.536**	0.693	0.619	1.439	0.693	0.628	1.481	0.692	0.519	1.569	0.533	0.484	1.510	0.536	**0.416**	**1.636**
WCBCR	1.866	**0.616**	**1.536**	1.134	0.619	1.439	0.650	0.628	1.481	0.363	0.518	**1.607**	0.275	0.484	1.510	0.170	**0.416**	**1.636**
Bat																		
DBI	1.215	0.661	1.227	1.017	0.520	1.469	0.916	0.517	1.518	1.020	**0.476**	1.475	1.087	0.520	1.464	1.040	0.423	1.468
SI	0.744	0.686	1.176	0.744	0.521	1.427	0.694	0.538	1.360	0.670	0.540	1.393	0.625	**0.480**	1.400	0.622	0.465	1.387
WCBCR	2.277	0.639	1.404	1.131	**0.518**	**1.476**	0.559	**0.507**	**1.599**	0.403	0.478	1.504	0.269	0.490	**1.548**	0.215	0.503	1.476

TABLE 7. Validity indices, Compactness and separation values for K-means and bat on Elbow-Points (Fall Night Time).

Validity index (VI)	5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	0.881	0.350	1.815	0.861	0.258	**2.060**	0.887	0.269	**1.829**	0.839	0.222	1.815	0.917	0.231	1.673	0.945	0.248	1.578
SI	0.643	0.304	1.701	0.671	0.258	**2.060**	0.612	0.269	**1.829**	0.610	0.222	1.815	0.606	0.231	1.673	0.585	0.231	1.567
WCBCR	0.709	0.377	1.902	0.302	**0.221**	2.046	0.215	0.269	**1.829**	0.149	0.222	1.815	0.116	0.231	1.673	0.092	0.248	1.578
Bat																		
DBI	0.655	**0.177**	**2.384**	0.712	0.231	1.691	0.702	**0.174**	1.961	0.761	0.222	1.815	0.634	0.176	1.782	0.786	**0.182**	1.670
SI	0.677	0.477	1.965	0.654	**0.221**	2.046	0.648	0.178	1.978	0.610	0.222	1.815	0.607	0.225	1.655	0.610	0.194	1.672
WCBCR	0.477	**0.177**	**2.384**	0.302	**0.221**	2.046	0.205	0.174	1.961	0.149	**0.222**	**1.816**	0.107	**0.176**	**1.782**	0.084	0.183	**1.674**

farms, in addition, to preparing for implementing appropriate corrective measures.

VI. SHORT-TERM PREDICTION OF WIND POWER

It is a great challenge to integrate wind farms into the smart electric grid due to the uncertainty of wind. Studies related to the integration of wind farms into the smart grid are of interest [8]–[10]. This problem can be solved by accurate prediction of short-term wind power generation that would help in optimizing and operating the power systems control. However, the meteorological and climatic conditions define the accuracy of wind power prediction that makes it a challenging task.

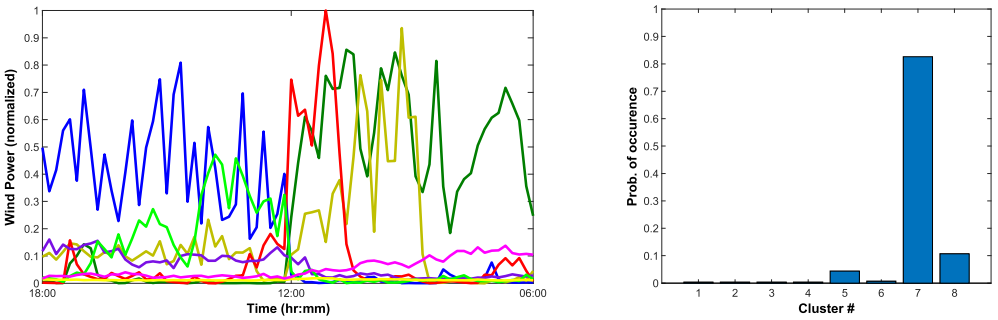

FIGURE 6. (a) The eight cluster representatives for the spring night time data. (b) Probability of occurrence for the 8 clusters of (a).

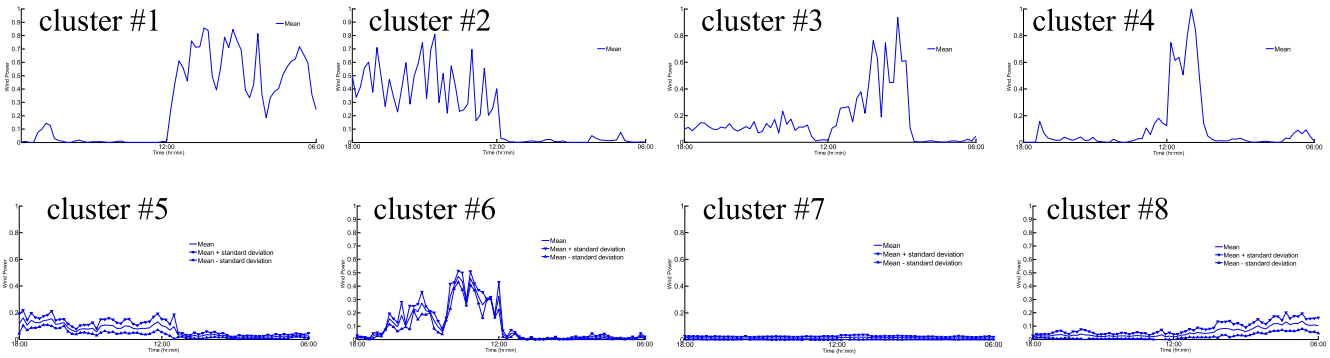

FIGURE 7. WPPs with respective confidence intervals using K-means with assuming 8 clusters (normalized).

TABLE 8. Accuracy measures between the actual and predicted data.

	Past time (min)	Predicted time (min)	RMSE	MAE	Corr.
Night time	*K-means*				
	120	60	6.8654e+04	1.8600e+03	0.383
	60	60	6.8521e+04	1.8703e+03	0.457
	Bat				
	120	60	2.8993e+05	4.4016e+03	0.146
	60	60	3.0985e+05	4.2154e+03	0.151
Day time	*K-means*				
	120	60	6.7523e+04	1.6710e+03	0.635
	60	60	6.9285e+04	2.6530e+03	0.640
	Bat				
	120	60	7.2333e+05	2.2021e+03	0.569
	60	60	7.3641e+05	3.5665e+03	0.612

A. WIND POWER PREDICTION MODEL

The wind power prediction model utilizes dedicated formulation to study the efficiency of the WPP cluster representatives. This model takes advantage of the clustering representatives for short-term wind power prediction.

At time (t), past observations of wind speeds (t-1,t-2,…, t-n), and representative WPPs is used to predict the future wind power generation (t+1,t+2,…,t+f). The prediction is based on the classification of the past WPP time-step observations to the representative WPPs, then the future values are obtained from the closest pattern of the representative WPPs. The flowchart of this approach is presented in Fig. 8 and the steps are as follows:

1- The sequence of wind speeds are obtained prior to the interval to be predicted (t-1,t-2,… , t-n).
2- The wind power output for (t-1,t-2,…,t-n) is computed by using the model of Section II.
3- The distance between the obtained sequence and the corresponding time sequence of each representative WPP is computed.
4- The closest two WPPs and found and the average distance between them was computed, resulting in three WPPs.
5- The distance between the obtained sequence from step two and the corresponding time sequence of the three WPPs from step four is computed.
6- The future wind power values (t+1,t+2,…,t+f) are obtained from the closest WPP.

B. APPLICATION OF POWER PREDICTION MODEL

The 10-minutes ahead is predicted using the short-term wind power prediction model applied on a real data set. The cluster representatives are obtained for the previous section

TABLE 9. Validity indices, Compactness and separation values for K-means with silhouette objective function on Elbow-Points (Fall).

Validity index (VI)	4 Clusters			5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means Night Time																		
DBI	1.107	0.541	2.029	0.804	0.421	2.316	1.003	0.430	2.074	0.888	0.443	1.995	1.025	0.463	1.886	1.014	0.464	1.926
SI	0.710	0.541	2.029	0.703	0.418	2.294	0.592	0.436	2.050	0.589	0.428	2.041	0.590	0.445	1.946	0.591	0.463	1.943
WCBCR	0.937	0.565	2.054	0.448	0.421	2.316	0.302	0.430	2.086	0.200	0.431	2.062	0.148	0.445	1.946	0.106	0.448	1.943
	4 Clusters			5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters		
K-means Day Time																		
DBI	1.135	0.616	1.536	1.128	0.631	1.337	1.010	0.628	1.481	0.936	0.510	1.617	0.984	0.490	1.514	0.904	0.357	1.675
SI	0.741	0.616	1.536	0.712	0.619	1.439	0.713	0.628	1.481	0.715	0.510	1.617	0.700	0.439	1.706	0.660	0.357	1.675
WCBCR	0.918	0.616	1.536	0.516	0.619	1.439	0.274	0.507	1.599	0.177	0.518	1.607	0.117	0.439	1.706	0.086	0.357	1.675

TABLE 10. Validity indices, Compactness and separation values for K-means with silhouette objective function on Elbow-Points (Spring).

Validity index (VI)	4 Clusters			5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means Night Time																		
DBI	0.978	0.712	2.465	0.837	0.451	2.564	0.752	0.351	2.695	0.767	0.245	2.690	0.581	0.212	2.650	0.767	0.235	2.480
SI	0.658	0.712	2.465	0.653	0.710	2.286	0.630	0.351	2.695	0.578	0.447	2.487	0.628	0.242	2.652	0.478	0.267	2.495
WCBCR	0.792	0.691	2.529	0.420	0.451	2.564	0.246	0.351	2.695	0.161	0.254	2.699	0.113	0.212	2.650	0.086	0.252	2.483
	5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters		
K-means Day Time																		
DBI	1.009	0.349	1.675	1.247	0.350	1.482	1.129	0.292	1.515	1.011	0.300	1.477	0.887	0.223	1.495	0.907	0.244	1.462
SI	0.497	0.348	1.599	0.533	0.353	1.487	0.494	0.295	1.524	0.493	0.300	1.477	0.481	0.317	1.440	0.478	0.246	1.483
WCBCR	0.278	0.353	1.680	0.192	0.350	1.493	0.124	0.295	1.523	0.088	0.255	1.516	0.064	0.224	1.499	0.048	0.246	1.483

TABLE 11. Validity indices, Compactness and separation values for K-means with silhouette objective function on Elbow-Points (Summer).

Validity index (VI)	4 Clusters			5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means Night Time																		
DBI	1.189	0.324	0.957	1.239	0.330	0.855	1.262	0.335	0.798	1.121	0.349	0.837	1.220	0.343	0.797	1.126	0.308	0.786
SI	0.523	0.325	0.957	0.525	0.329	0.844	0.466	0.321	0.774	0.477	0.275	0.865	0.477	0.263	0.860	0.486	0.303	0.762
WCBCR	0.802	0.324	0.957	0.483	0.369	0.912	0.326	0.344	0.799	0.188	0.258	0.935	0.138	0.265	0.908	0.109	0.302	0.870
	5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters		
K-means Day Time																		
DBI	1.196	0.241	0.905	1.318	0.252	0.834	1.176	0.208	0.968	1.186	0.221	0.854	1.041	0.172	0.862	1.004	0.148	0.845
SI	0.529	0.241	0.905	0.465	0.248	0.826	0.477	0.200	0.889	0.480	0.248	0.784	0.469	0.179	0.872	0.471	0.186	0.849
WCBCR	0.494	0.241	0.905	0.327	0.252	0.834	0.187	0.208	0.968	0.148	0.221	0.854	0.106	0.179	0.872	0.082	0.177	0.847

TABLE 12. Validity indices, Compactness and separation values for K-means with silhouette objective function on Elbow-Points (Winter).

Validity index (VI)	7 Clusters			8 Clusters			9 Clusters			10 Clusters			11 Clusters			12 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means Night Time																		
DBI	0.986	0.542	1.693	0.928	0.519	1.582	0.843	0.373	1.736	0.779	0.322	1.793	0.790	0.356	1.774	0.777	0.381	1.712
SI	0.660	0.542	1.693	0.636	0.515	1.635	0.611	0.448	1.634	0.662	0.311	1.796	0.620	0.356	1.774	0.621	0.374	1.738
WCBCR	0.266	0.542	1.693	0.186	0.584	1.434	0.122	0.373	1.736	0.086	0.304	1.810	0.066	0.356	1.774	0.052	0.352	1.706
	6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters			11 Clusters		
K-means Day Time																		
DBI	0.797	0.394	2.053	0.865	0.400	1.946	0.809	0.334	1.990	0.768	0.288	1.979	0.679	0.248	1.957	0.641	0.223	1.960
SI	0.590	0.399	2.067	0.566	0.405	1.954	0.549	0.335	1.983	0.555	0.277	1.968	0.562	0.246	1.959	0.559	0.218	1.956
WCBCR	0.101	0.398	2.069	0.070	0.403	1.956	0.048	0.335	1.983	0.035	0.282	1.980	0.027	0.254	1.975	0.021	0.223	1.960

using the K-means clustering algorithm. The 22 day-time representative WPPs and 20 night-time representative WPPs resulted due to clustering of three consecutive years. The wind speeds were converted to wind power for a following year and the test data was obtained by selecting eight random days from each season. Accordingly, daily 36 WPPs were obtained. The results of predicting 60-minutes ahead from the sequence of the past 60-minutes and 120-minutes are illustrated in Table 8. The clustering representatives of Bat algorithm were also put into the prediction model and the results are shown in Table 8, to prove that the presented methodology including that algorithm with better validity indices values should obtain more accurate results in the simulations. Calculations were performed for RMSE, MAE, and correlation coefficient between the actual data and predicted data. A superior prediction performance of the model is implied using smaller values of RMSE and MAE. The correlation of data is presented through a larger positive correlation coefficient value. Table 8 shows that increasing the past sequence does not improve the prediction. While, the comparison between the actual and predicted wind power for day-time using the previous 60-minutes for K-means is shown in Fig. 8. From the results, it can be observed that there is a significant error in the prediction and that the prediction results are unsatisfactory for wind power data. The reason for those poor results can be due to: 1) the problem of dealing

TABLE 13. CPU time in Seconds for clustering algorithms from two to twenty clusters on WPP night time spring data.

Algo.	***K-means***																		
Clusters	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Best	0.012	0.012	0.013	0.014	0.015	0.015	0.016	0.017	0.018	0.019	0.019	0.021	0.021	0.022	0.023	0.024	0.024	0.025	0.026
Worst	0.033	0.038	0.018	0.017	0.026	0.034	0.022	0.022	0.025	0.033	0.025	0.025	0.025	0.026	0.029	0.028	0.029	0.034	0.032
Average	0.015	0.015	0.014	0.015	0.017	0.018	0.018	0.018	0.020	0.021	0.021	0.022	0.023	0.024	0.025	0.025	0.026	0.028	0.028
Algo.	***Ant Colony***																		
Clusters	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Best	3.077	3.295	3.428	3.521	3.634	3.779	3.844	3.961	4.093	4.239	4.333	4.458	4.526	4.605	4.808	4.916	5.010	5.140	5.227
Worst	7.800	3.504	5.668	5.048	4.264	7.580	5.490	6.486	4.303	9.281	9.481	6.546	7.087	5.271	5.118	5.705	9.661	7.588	5.829
Average	3.452	3.351	3.624	3.678	3.729	4.024	4.103	4.188	4.176	4.652	4.673	4.627	4.773	4.788	4.890	5.038	5.377	5.396	5.388
Algo.	***Bat***																		
Clusters	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Best	3.67	4.60	5.49	6.46	7.40	8.26	9.23	10.10	11.01	11.95	12.83	13.77	14.71	15.63	16.51	17.43	18.34	19.29	20.18
Worst	7.34	4.84	7.82	7.58	11.05	9.26	9.64	10.95	19.29	15.37	16.60	16.20	16.57	16.80	23.84	21.30	19.12	23.27	23.83
Average	4.16	4.66	5.80	6.69	7.89	8.48	9.35	10.26	12.54	12.63	13.49	14.28	15.23	15.83	17.83	18.31	18.58	20.09	20.92
Algo.	***K-means-SI***																		
Clusters	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Best	5.25	5.92	6.44	7.15	7.59	8.19	8.77	9.27	9.86	10.42	10.89	11.81	12.15	12.71	13.21	13.77	14.46	14.84	15.57
Worst	14.74	9.72	16.67	10.16	13.95	14.54	12.85	19.61	13.20	20.85	14.67	21.72	23.16	23.73	17.55	24.62	27.11	21.90	26.54
Average	7.02	7.63	8.68	8.58	9.28	10.08	10.53	11.91	11.84	13.44	12.80	14.64	14.87	15.56	15.17	16.55	18.27	17.49	19.62

TABLE 14. Accuracy measures between the actual and predicted data.

	Past time (min)	Predicted time (min)	RMSE	MAE	Corr.
Night	*K-means*				
	120	60	5.4329e+04	1.1557e+03	0.611
	60	60	5.9964e+04	1.4496e+03	0.732
Day	*K-means*				
	120	60	5.1973e+04	1.2071e+03	0.614
	60	60	5.3342e+04	2.0270e+03	0.634

with continuous fluctuation in the wind speeds. 2) the results of the clustering algorithms were in-efficient with this type of data, although it presented satisfactory results on Photovoltaic solar power and residential loads data in previous studies.

The CPU time for the applied clustering algorithms is presented in Table 13. This time includes data conversion, data segmentation, dimension reduction, clustering and cluster formation evaluation. Each clustering algorithm was executed twenty times. It can be observed that both swarm methods (Ant Colony and Bat) require more time on average than the partitional K-means algorithm. This is due to the nature of the swarm algorithms, where each "ant" or "bat" generates a different solution that needs to be evaluated. Accordingly, the best solution sometimes requires more time to be achieved depending on the collaborative behavior of the swarm.

VII. K-MEANS WITH SILHOUETTE (K-MEANS-SI) OBJECTIVE FUNCTION

From the previous sections it was observed that the best combination of clustering algorithm and validity index were K-means and SI, respectively. In this section, an attempt to improve the clustering results of the K-means algorithm by integrating the SI index as an objective function (K-means-SI). The success of this formation of the K-means algorithm is expressed by having higher global Silhouette values instead of lower mean square distances.

A. IMPLEMENTATION OF K-MEANS-SI

Similar to Section V-A, the K-means-SI method is applied with 50 replicates. A new set of initial centroids are chosen at each step of replication. Recordings were obtained for solutions with the highest global Silhouette values.

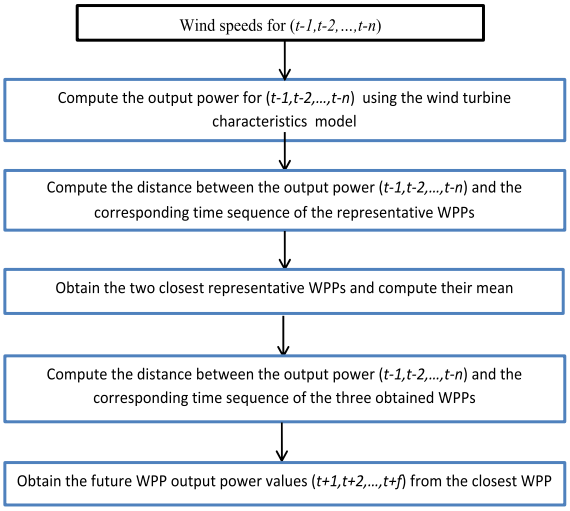

FIGURE 8. Flowchart of the model.

B. APPLICATION OF K-MEANS-SI

The K-means-SI clustering algorithm was applied on the same datasets. Tables 9-12, presents the results of day and night times for all four seasons. It can be observed that the overall results with regards to the compactness and separation improved by integrating the Silhouette function into the K-means algorithm. The CPU time of the K-means-SI is presented in Table 13. The integration of the SI index into the K-means algorithm increased the CPU significantly. This is due to the increased computations in the SI objective function. The K-mean-SI clustering representatives are used in the short-term wind power prediction model of Section VI to further test the results of the K-means-SI.

The 26 day-time representative WPPs and 25 night-time representative WPPs resulted due to clustering of the three consecutive years for the data. The wind speeds for a following unseen year were changed to wind power and eight

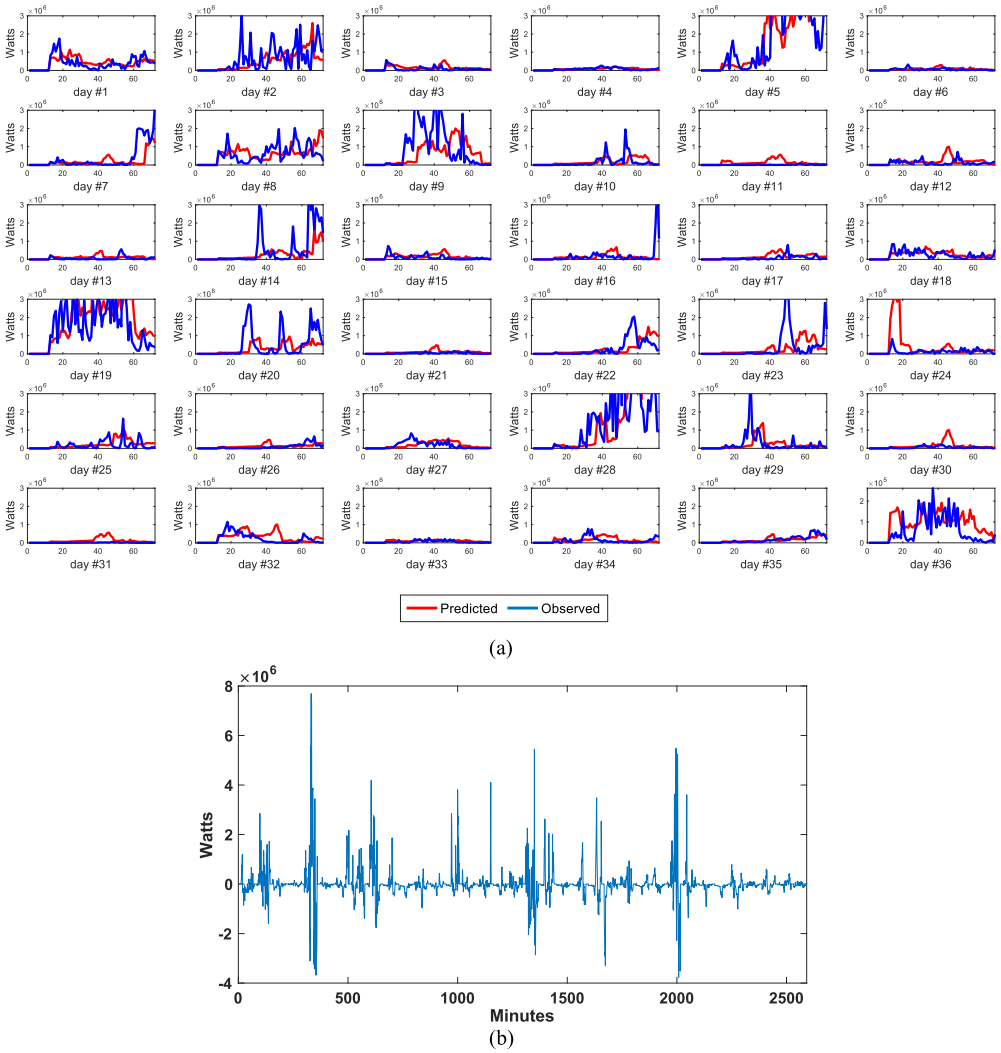

FIGURE 9. (a) Comparison between the actual and predicted WPP for predicting 60 minutes from the past 60 minutes (day time). (b) Error for day time (72 observations) for the 36-days.

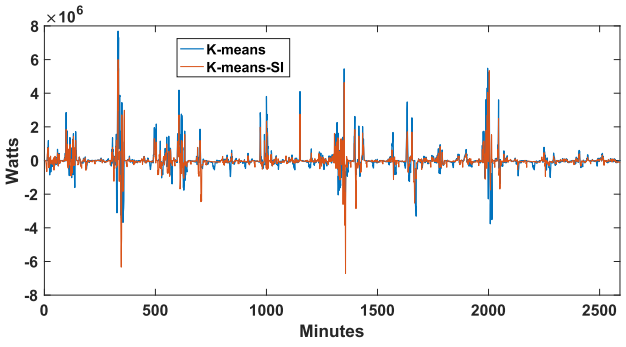

FIGURE 10. Error for day time (72 observations) for the 36-days.

random days from each season were chosen to test data was obtained. Accordingly, 36 daily WPPs were acquired. Table 13 shows the results of predicting 60-minutes ahead from the sequence of the past 60-minutes and 120-minutes. The wind power output at each time observation could reach up to 4.3250e+07. Accordingly, the results of RMSE and MAE are superior given that observations could reach up to 4.3250e+07 and that the error was computed for at least 60 min (60 time-observations). By comparing the Tables 8 and 14, it can be observed that the clustering results of integrating the SI objective function into the K-means algorithm improved the cluster representatives and hence, the prediction improved around 17% and 23% on average for night-time and day-time, respectively. Fig. 10, shows the prediction error results of K-means versus K-means-SI on the test data of day-time observations for the 36-days. It can be observed that the K-means-SI (Fig. 10 in red) reduced the prediction error at many time observations.

VIII. CONCLUSION

The present study investigates the most appropriate method for establishing the wind power pattern (WPP) clustering process using three clustering algorithms from two different categories. The applied clustering algorithms included K-means, Ant Colony, and Bat. The eight validity indices,

TABLE 15. SI validity index values for K-means, Bat and ant colony for two-to-twenty clusters (Spring).

SI Ind.	Day Time																		
Cluster#	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K-means	0.886	0.682	0.594	**0.586**	0.487	0.486	0.493	0.475	0.415	0.381	0.361	0.357	0.340	0.342	0.283	0.317	0.286	0.290	0.294
Bat	0.734	0.684	0.594	**0.653**	0.472	0.500	0.499	0.449	0.444	0.444	0.433	0.387	0.385	0.374	0.377	0.291	0.297	0.293	0.305
Ant	0.528	0.406	**0.324**	0.302	0.294	0.252	0.233	0.243	0.235	0.227	0.229	0.213	0.209	0.238	0.239	0.211	0.223	0.199	0.181
SI Ind.	**Night Time**																		
Cluster#	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K-means	0.828	0.684	0.552	0.535	0.541	0.550	**0.552**	0.505	0.458	0.392	0.387	0.326	0.337	0.294	0.299	0.302	0.298	0.299	0.294
Bat	0.762	0.680	**0.683**	0.559	0.544	0.552	0.534	0.505	0.446	0.399	0.333	0.332	0.302	0.316	0.295	0.297	0.291	0.299	0.305
Ant	0.514	0.459	**0.282**	0.238	0.234	0.207	0.209	0.217	0.224	0.180	0.197	0.178	0.166	0.136	0.184	0.116	0.125	0.098	0.181

TABLE 16. DBI validity index values for K-means, Bat and ant colony for two-to-twenty clusters (Spring).

DBI Ind.	Day Time																		
Cluster#	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K-means	0.107	0.930	1.091	1.194	1.198	1.298	**1.096**	1.109	1.261	1.276	1.228	1.280	1.269	1.361	1.263	1.294	1.278	1.238	1.183
Bat	1.390	0.827	1.091	**0.944**	1.168	1.033	1.051	1.045	1.091	1.083	1.088	1.072	1.089	1.051	0.948	1.194	1.092	1.127	1.103
Ant	2.155	2.259	2.304	2.543	2.392	2.159	**1.987**	2.306	2.215	2.168	1.978	2.143	2.064	1.900	1.909	2.047	1.988	2.036	1.847
DBI Ind.	**Night Time**																		
Cluster#	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K-means	0.952	1.081	1.254	1.075	0.752	**0.542**	0.576	0.642	0.811	0.896	0.911	0.989	1.013	1.006	0.930	0.978	0.909	0.914	0.921
Bat	1.714	1.107	0.824	0.910	0.720	0.807	0.717	**0.716**	0.853	0.929	1.057	1.051	1.009	0.932	1.025	0.992	0.999	0.894	1.168
Ant	2.299	2.331	2.418	2.251	2.261	**2.110**	2.122	2.045	1.819	2.045	1.980	1.961	1.869	1.900	1.965	1.911	2.011	1.904	1.846

TABLE 17. WCBCR validity index values for K-means, Bat and ant colony for two-to-twenty clusters (Spring).

WCBCR	Day Time																		
Cluster#	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K-means	3.056	1.302	0.757	0.483	0.325	0.254	0.191	**0.143**	0.120	0.102	0.083	0.072	0.061	0.055	0.045	0.043	0.035	0.031	0.027
Bat	34.31	1.298	0.757	0.483	0.312	0.229	0.178	**0.140**	0.118	0.095	0.079	0.067	0.061	0.054	0.046	0.041	0.038	0.034	0.033
Ant	217.5	64.10	27.67	15.27	**7.960**	4.670	1.759	1.002	0.233	0.206	0.167	0.148	0.128	0.104	0.106	0.085	0.078	0.074	0.069
WCBCR	**Night Time**																		
Cluster#	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K-means	11.34	2.979	1.536	0.577	0.264	0.159	0.117	**0.091**	0.074	0.062	0.052	0.045	0.038	0.034	0.029	0.026	0.023	0.020	0.018
Bat	42.16	4.115	1.056	0.525	0.264	0.186	0.122	**0.091**	0.075	0.062	0.053	0.045	0.039	0.034	0.030	0.026	0.024	0.020	0.033
Ant	244.4	60.34	32.43	16.89	9.816	**6.248**	3.418	0.925	1.111	0.434	0.410	0.221	0.145	0.176	0.136	0.080	0.097	0.076	0.060

TABLE 18. Validity indices, Compactness and separation values for K-means and bat on Elbow-Points (Summer Day Time).

Validity index (VI)	5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters			10 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	1.251	0.262	0.948	1.255	**0.246**	**0.896**	1.062	**0.209**	0.949	1.185	**0.208**	0.876	1.249	0.249	0.780	1.209	0.220	0.812
SI	0.516	0.240	0.878	0.503	**0.246**	**0.896**	0.510	0.247	0.817	0.461	0.251	0.787	0.446	0.246	0.733	0.444	0.220	0.812
WCBCR	0.767	0.239	**0.990**	0.544	**0.246**	**0.896**	0.344	**0.209**	0.949	0.272	**0.208**	0.876	0.224	0.209	0.818	0.176	0.220	0.812
Bat																		
DBI	1.287	**0.237**	0.892	1.211	0.248	0.831	1.377	0.254	0.843	1.226	0.210	0.876	1.176	**0.196**	0.784	1.096	**0.158**	**0.875**
SI	0.505	0.238	0.893	0.503	**0.246**	**0.896**	0.458	0.248	0.764	0.461	0.211	**0.890**	0.452	0.223	0.742	0.466	0.197	0.808
WCBCR	0.833	0.241	0.896	0.544	**0.246**	**0.896**	0.404	0.254	0.843	0.269	0.211	**0.890**	0.221	0.223	**0.843**	0.165	**0.158**	**0.875**

DBI, Dunn, J, XB, Silhouette, MIA, CDI and WCBCR were used to evaluate the clustering results of each algorithm to find the optimum number of clusters that best fits the WPP data, and further investigate the most efficient clustering algorithm and validity index.

Further, this work introduced swarm clustering techniques to establish the clustering of WPPs. The comparison of the clustering algorithms of different categories and characteristics illustrated that bio-inspired swarm Bat clustering algorithm is comparable to K-means. However, it corresponds to increased complexity as the number of parameters should be calibrated. To determine the optimum number of clusters, the L-method was adopted. The L-method was able to systematically find the elbow point that presents the overall best compactness and separation of data points. The best clustering results were chosen based on the Silhouette value as the best combination of clustering algorithm and validity index were K-means and Silhouette, respectively. Furthermore, to improve the results presented by the K-means and Silhouette, the SI was integrated as an objective function for K-means to present the K-means-SI algorithm. The clustering results produced by K-means-SI improved the clusters' formation and produced more compact and separated clusters of data. However, integrating the Silhouette index as an objective function for K-means increased the CPU time significantly. This is due to the additional computations of the SI objective function in partitioning the data. Those observations were applicable to all the other seasons' data sets used in the study.

A short-term wind power prediction model was presented. This model utilizes the WPP representatives to predict future wind power. This this model tested the efficiency of the representative WPPs resulting from the clustering methodology. In a first application, the clustering representatives

TABLE 19. Validity indices, Compactness and separation values for K-means and bat on Elbow-Points (Summer Night Time).

Validity index (VI)	4 Clusters			5 Clusters			6 Clusters			7 Clusters			8 Clusters			9 Clusters		
	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.	VI	Comp.	Sep.
K-means																		
DBI	1.026	0.223	1.036	1.057	0.224	0.948	1.026	**0.231**	**0.955**	1.037	0.221	0.887	1.013	**0.180**	0.863	1.081	**0.181**	0.819
SI	0.573	**0.215**	0.950	0.500	**0.215**	0.963	0.504	**0.231**	**0.955**	0.475	0.222	**0.888**	0.453	0.216	0.841	0.435	**0.181**	0.819
WCBCR	0.493	0.223	1.036	0.339	**0.215**	0.963	0.237	**0.231**	**0.955**	0.184	0.221	0.887	0.144	**0.180**	0.863	0.116	0.186	0.833
Bat																		
DBI	1.021	0.228	**1.048**	0.972	0.239	**1.007**	1.019	**0.231**	**0.955**	1.098	0.221	0.887	0.955	0.189	0.896	1.064	0.189	**0.839**
SI	0.563	0.218	0.997	0.541	0.239	**1.007**	0.504	**0.231**	**0.955**	0.487	0.240	0.852	0.460	0.218	0.831	0.454	0.211	0.804
WCBCR	0.493	0.223	1.036	0.335	0.239	**1.007**	0.237	**0.231**	**0.955**	0.184	**0.220**	**0.888**	0.139	0.189	**0.896**	0.116	0.189	**0.839**

resulted from the K-means and SI were used for short-term prediction. The prediction results suggest that more research should be conducted to efficiently cluster WPPs. For that, in another application the cluster representatives of the K-means-SI were used in the short-term prediction model. The prediction results were improved by around 17% and 23% on average for night-time and day-time, respectively, and suggest that the presented cluster formation could be used in other WPP studies.

Based on the presented results in this work, some of the studies that can be carried out in the future are: 1) the investigation of clustering algorithms and validity indices that could potentially improve the clustering of wind power data. 2) the examination of the use of other features to improve the accuracy of WPP cluster representatives. 3) the construction on models that can improve the prediction using only cluster representatives.

APPENDIX A

See Table 15, 16, and 17 here.

APPENDIX B

See Table 18 and 19 here.

REFERENCES

[1] World Energy Outlook, Paris. (2019). *International Energy Association*. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2019

[2] A. Qazi, F. Hussain, N. A. Rahim, G. Hardaker, D. Alghazzawi, K. Shaban, and K. Haruna, "Towards sustainable energy: A systematic review of renewable energy sources, technologies, and public opinions," *IEEE Access*, vol. 7, pp. 63837–63851, 2019.

[3] X. Chen, K.-C. Leung, and A. Y. S. Lam, "Power output smoothing for renewable energy system: Planning, algorithms, and analysis," *IEEE Syst. J.*, vol. 14, no. 1, pp. 1034–1045, Mar. 2020.

[4] M. Marwede and A. Reller, "Estimation of life cycle material costs of cadmium telluride-and copper indium gallium diselenide-photovoltaic absorber materials based on life cycle material flows," *J. Ind. Ecol.*, vol. 18, no. 2, pp. 254–267, Apr. 2014.

[5] A. Goodrich, P. Hacke, Q. Wang, B. Sopori, R. Margolis, T. L. James, and M. Woodhouse, "A wafer-based monocrystalline silicon photovoltaics road map: Utilizing known technology improvement opportunities for further reductions in manufacturing costs," *Sol. Energy Mater. Sol. Cells*, vol. 114, pp. 110–135, Jul. 2013.

[6] K. Kushiya, "Key near-term R&D issues for continuous improvement in CIS-based thin-film PV modules," *Sol. Energy Mater. Sol. Cells*, vol. 93, no. 6, pp. 1037–1041, 2009.

[7] X. Yang, Y. Yang, Y. Liu, and Z. Deng, "A reliability assessment approach for electric power systems considering wind power uncertainty," *IEEE Access*, vol. 8, pp. 12467–12478, 2020.

[8] X. Lyu, Y. Jia, Z. Xu, and J. Ostergaard, "Mileage-responsive wind power smoothing," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5209–5212, Jun. 2020.

[9] Y. Guo and H. Gao, "Data-driven online system equivalent for self-adaptive droop voltage control of wind power plants," *IEEE Trans. Energy Convers.*, vol. 35, no. 1, pp. 302–305, Mar. 2020.

[10] P. Sinha, "Life cycle materials and water management for CdTe photovoltaics," *Sol. Energy Mater. Sol. Cells*, vol. 119, pp. 271–275, Dec. 2013.

[11] W. D. Cyrs, H. J. Avens, Z. A. Capshaw, R. A. Kingsbury, J. Sahmel, and B. E. Tvermoes, "Landfill waste and recycling: Use of a screening-level risk assessment tool for end-of-life cadmium telluride (CdTe) thin-film photovoltaic (PV) panels," *Energy Policy*, vol. 68, pp. 524–533, May 2014.

[12] Y. Li, J. Wang, D. Zhao, G. Li, and C. Chen, "A two-stage approach for combined heat and power economic emission dispatch: Combining multi-objective optimization with integrated decision making," *Energy*, vol. 162, pp. 237–254, Nov. 2018.

[13] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 381–387, Feb. 2003.

[14] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005.

[15] D. Gerbec, S. Gašperič, I. Šmon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *IEE Proc.—Gener., Transmiss. Distrib.*, vol. 151, no. 3, pp. 395–400, May 2004.

[16] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 1232–1239, May 2004.

[17] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.

[18] S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. G. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Trans. Power Syst.*, vol. 21, no. 4, pp. 1672–1682, Nov. 2006.

[19] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, Aug. 2007.

[20] G. Chicco, O.-M. Ionel, and R. Porumb, "Electrical load pattern grouping based on centroid model with ant colony clustering," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1706–1715, May 2013.

[21] A. Molina-García, A. Fernández-Guillamón, E. Gómez-Lázaro, A. Honrubia-Escribano, and M. C. Bueso, "Vertical wind profile characterization and identification of patterns based on a shape clustering algorithm," *IEEE Access*, vol. 7, pp. 30890–30904, 2019.

[22] G. Chicco, R. Napoli, and F. Piglione, "Load pattern clustering for short-term load forecasting of anomalous days," in *Proc. IEEE PowerTech Conf., Porto*, vol. 2, Sep. 2001, p. 6.

[23] A. Gabaldon, A. Guillamon, M. C. Ruiz, S. Valero, C. Alvarez, M. Ortiz, and C. Senabre, "Development of a methodology for clustering electricity-price series to improve customer response initiatives," *IET Gener., Transm., Distrib.*, vol. 4, no. 6, pp. 706–715, 2010.

[24] G. J. Tsekouras, C. A. Anastasopoulos, F. D. Kanellos, V. T. Kontargyri, I. S. Karanasiou, A. D. Salis, and N. E. Mastorakis, "A demand side management program of vanadium redox energy storage system for an interconnected power system," in *Proc. WSEAS EPESE*, Corfu Island, Greece, 2008, pp. 26–28.

[25] G. J. Tsekouras, F. D. Kanellos, V. T. Kontargyri, I. S. Karanasiou, A. D. Salis, and N. E. Mastorakis, "A new classification pattern recognition methodology for power system typical load profiles," *WSEAS Trans. Circuits Syst.*, vol. 12, no. 7, pp. 1090–1104, 2008.

[26] Y. Xiang, J. Hong, Z. Yang, Y. Wang, Y. Huang, X. Zhang, Y. Chai, and H. Yao, "Slope-based shape cluster method for smart metering load profiles," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1809–1811, Mar. 2020.
[27] G. J. Tsekouras, I. K. Hatzilau, and J. M. Prousalidis, "A new pattern recognition methodology for classification of load profiles for ships electric consumers," *J. Mar. Eng. Technol.*, vol. 8, no. 2, pp. 45–58, Jan. 2009.
[28] G. Tsamopoulos, N. Giannitsas, F. D. Kanellos, and G. J. Tsekouras, "Load estimation for war-ships based on pattern recognition methods," *J. Comput. Model.*, vol. 4, no. 1, pp. 207–222, 2014.
[29] M. Ali, I.-S. Ilie, J. V. Milanovic, and G. Chicco, "Wind farm model aggregation using probabilistic clustering," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 309–316, Feb. 2013.
[30] Z. Sun, S. Zhao, and J. Zhang, "Short-term wind power forecasting on multiple scales using VMD decomposition, K-Means clustering and LSTM principal computing," *IEEE Access*, vol. 7, pp. 166917–166929, 2019.
[31] R. Rahmani, R. Yusof, M. Seyedmahmoudian, and S. Mekhilef, "Hybrid technique of ant colony and particle swarm optimization for short term wind energy forecasting," *J. Wind Eng. Ind. Aerodyn.*, vol. 123, pp. 163–170, Dec. 2013.
[32] L. Xiao, F. Qian, and W. Shao, "Multi-step wind speed forecasting based on a hybrid forecasting architecture and an improved bat algorithm," *Energy Convers. Manage.*, vol. 143, pp. 410–430, Jul. 2017.
[33] F. J. Duarte, J. M. M. Duarte, S. Ramos, A. Fred, and Z. Vale, "Daily wind power profiles determination using clustering algorithms," in *Proc. IEEE Int. Conf. Power Syst. Technol. (POWERCON)*, vol. 2, Oct./Nov. 2012, pp. 1–6.
[34] W. A. Omran, M. Kazerani, and M. M. A. Salama, "A clustering-based method for quantifying the effects of large on-grid PV systems," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2617–2625, Oct. 2010.
[35] N. Haghdadi, B. Asaei, and Z. Gandomkar, "Clustering-based optimal sizing and siting of photovoltaic power plant in distribution network," in *Proc. EEEIC*, May 2012, pp. 266–271.
[36] H. Mori and M. Takahashi, "Application of preconditioned generalized radial basis function network to prediction of photovoltaic power generation," in *Proc. IEEE ISGT*, Oct. 2012, pp. 1–6.
[37] Y. Hosoda and T. Namerikawa, "Short-term photovoltaic prediction by using H_∞ filtering and clustering," in *Proc. SICE*, Aug. 2012, pp. 119–124.
[38] Global Wind Report. (2014). *Global Wind Energy Council*. [Online]. Available: http://www.gwec.net/wp-content/uploads/2015/03/GWEC_Global_Wind_2014_Report_LR.pdf
[39] A. A. Munshi and Y. A.-R.-I. Mohamed, "Photovoltaic power pattern clustering based on conventional and swarm clustering methods," *Sol. Energy*, vol. 124, pp. 39–56, Feb. 2016.
[40] W. Tong, *Wind Power Generation and Wind Turbine Design*. Southampton, U.K.: WIT Press, 2010.
[41] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.
[42] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.
[43] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, "An ant colony approach for clustering," *Anal. Chim. Acta*, vol. 509, no. 2, pp. 187–195, May 2004.
[44] L. Jiang, L. Ding, Y. Peng, and C. Zhao, "An efficient clustering approach using ant colony algorithm in multidimensional search space," in *Proc. FSKD*, vol. 2, Jul. 2011, pp. 1085–1089.
[45] A. P. Engelbrecht, *Computational Intelligence: An Introduction*, 2nd ed. Hoboken, NJ, USA: Wiley, 2007, pp. 135–137.
[46] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature Inspired Cooperative Strategies for Optimization (NICSO)*. Berlin, Germany: Springer, 2010, pp. 65–74.
[47] R. Tang, S. Fong, X.-S. Yang, and S. Deb, "Integrating nature-inspired optimization algorithms to k-means," in *Proc. ICDIM*, Aug. 2012, pp. 116–123.
[48] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013.
[49] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[50] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, Jan. 1973.
[51] R. Jain and A. Koronios, "Innovation in the cluster validating techniques," *Fuzzy Optim. Decis. Making*, vol. 7, no. 3, pp. 233–241, Sep. 2008.
[52] Z. Liang, P. Zhang, and J. Zhao, "Optimization of the number of clusters in fuzzy clustering," in *Proc. ICCDA*, vol. 3, Jun. 2010, pp. 580–584.
[53] D. Hand, H. Manilla, and, P. Smyth, *Principles of Data Mining*. Cambridge, MA, USA: MIT Press, 2001.
[54] NREL. *National Wind Technology Center*. Accessed: Nov. 2015. [Online]. Available: http://www.nrel.gov/midc/nwtc_m2/
[55] Kingspan Wind. (2011). *KW3, Datasheet*. [Online]. Available: http://www.kingspanwind.com/media/1068/ks-wind-brochure-12pp-may15-lr.pdf
[56] S. Salvador and P. Chan, "Learning states and rules for detecting anomaly in time series," *Appl. Intell.*, vol. 23, no. 3, pp. 241–255, 2005.

AMR A. MUNSHI (Member, IEEE) received the B.Sc. degree in computer engineering from Umm Al-Qura University, Makkah, Saudi Arabia, in 2008, and the M.Sc. and Ph.D. degrees in computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2014 and 2019, respectively. He is currently an Assistant Professor with the Computer Engineering Department, Umm Al-Qura University. His research interests include artificial intelligence, data mining, and big data analytics.

Dr. Munshi is a member of the Golden Key International Honor Society and currently serves as an Editor of *The Alberta Academic Review* journal.