Clustering of PV Load Patterns Based on Any Colony Centroid Model

Amr Munshi†
aaamunshi@uqu.edu.sa
† Computer Engineering Department, Umm Al-Qura University, Saudi Arabia

Summary
There has been a significant growth in global population and industrialization, as a consequence demand for electricity is increasing rapidly and the power systems need to increase the electricity generation. Currently, most of generated electricity is generated from fossil fuels. However, there are many financial and environmental concerns associated with the generation of electricity from such resource. Photovoltaic)PV) solar as a renewable resource is promising. The power output of PV systems is mainly affected by the solar irradiation and ambient temperature. This paper attempts at reducing the burden and improving the accuracy of the extensive simulations related to integrating PV systems into the electrical grid.

Keywords:
Ant colony clustering, electric grid, photovoltaic, solar panel

1. Introduction

Due to the growth of the global population and industrialization, the demand for electricity is increasing rapidly. Consequently, the power systems need to increase the electricity generation. Recent studies predict that the world's net electricity generation is projected to reach 8000 tera watt-hours (TWh) in 2030 [1]. Currently, most of generated electricity is generated from fossil fuels. However, there are many financial and environmental concerns associated with the generation of electricity from such resource. To overcome much of these concerns, renewable energy resources can assist in the electricity generating. There has been an increase in generation electricity from renewable resources, such as solar photovoltaic (PV) panels and wind turbines. The total installed wind and solar PV capacity is predicted to surpass fossil fuels by 2024. Solar PV alone accounts for 60% of all renewable capacity additions through 2025 [2].

A rigorous amount of research has been dedicated to integrate renewable resources into the electrical grid. Many studies cluster time series power patterns and obtaining representatives for these clusters during the last few years. Models based on clustering techniques were used to group electrical load patterns of customers in order to assist tariff formation [3]–[6], short-term forecasting [7], and demand response programs to support management decisions [8]–[10]. Also, power load clustering has been used for the classification of load profiles for ship electric consumers [11] and for estimating the power load of warships [12]. In [13], aggregate modeling of wind farms has been proposed based on the wind farm's layout and the clustering of wind speed patterns. A method to improve the management decisions of wind farms was also proposed in [14] by applying a clustering method on wind power loads.

The power output of PV systems is mainly affected by the solar irradiation and ambient temperature [15]. This essentially results in operational problems and instability in the output power generated from PV systems. Consequently, integration PV systems requires extensive simulations of lengthy historical data. However, dealing with such data is time consuming and computationally expensive. For this purpose, the main focus of this paper is enhancing the process by reducing the burden and improving the accuracy of the extensive simulations related to integrating PV systems into the electrical grid.

The remainder of the paper is structured as follows. Section 2, presents the background information. Section 3, presents the methodology followed to reduce the burden of the extensive simulations related to integrating PV systems into the electrical grid. The simulation results of applying the methodology are presented in Section 4. The conclusions are drawn is Section 5.

2. Background

This section presents the preliminary definitions and the Ant Colony algorithm modified to support the utilization of the algorithm for clustering purposes.

2. 1. Preliminary Definitions

This subsection illustrates the definitions and notations of the clustering algorithms and validity indices used in the context of PV power load (PVPL) clustering. The initial data are a set of N daily PVPL referring to a specified period of time (i.e., the fall season for the past few years). Each daily PVPL contains d time-series observations (features). The row vector $x_n = [x_{n1}, \ldots, x_{nd}]$ represents the nth PVPL for $x = 1,\ldots, N$. The PVPL data set is represented by the matrix $\boldsymbol{X} = [x_{1,}\ldots, x_N]$. The clustering process creates a partitioning of the N PVPL into K clusters with non-overlapping PVPL through an iterative process. Each cluster is represented by a centroid $C_k = [C_{k1},\ldots, C_{kd}]$, for $k = 1,\ldots, K$. The set of centroids is represented by the matrix $\boldsymbol{C} = [C_1,\ldots, C_K]$.

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.11

2. 2. Ant Colony Clustering Algorithm

Ant Colony clustering is a swarm-based approach that attempts to mimic the behavior of real ants to find the shortest path between their nest and prey. The ants communicate and exchange information about the paths by means of pheromone trails. As more ants trace a certain path and deposit their pheromone, the more attractive this path becomes and is followed by other ants. Consequently, this collaborative behaviour leads to the establishment of the shortest route path [16].

The algorithm can be divided into three main stages: initialization, first iteration, and successive iterations.

Initalization: In the initializtion stage, the number of clusters K and the number of ants A are defined. Then the initial set of centroids $\boldsymbol{C}^{(0)}$ are randomly chosen from the data set. An initial $N \times K$ phermone matrix $\phi^{(0)}$ is constructed by computing the distances between each data point and each centroid. The resulting null distances are replaced by a relatively small value ε to avoid division by zero:

$$r_{ik}^{(0)} = \max\left\{d(x_i, C_k), \varepsilon\right\} \tag{1}$$

Then, auxiliary variables based on the squared inverse of distances in $\phi^{(0)}$ are calculated:

$$\varphi'^{(0)}_{ik} = \left(r_{ik}^{(0)}\right)^2 \tag{2}$$

The components $\varphi'^{(0)}_{ik}$ of the pheromone matrix are normalized to avoid the continuous growth of pheromone components in the iterative stage. This normalizing is accomplished by dividing each auxiliary variable by the sum of auxiliary variables occurring in the same corresponding row:

$$\varphi_{ik}^{(0)} = \varphi'^{(0)}_{ik} \Big/ \sum_{k=1}^{K} \varphi'^{(0)}_{ik} \tag{3}$$

First iteration: For the number of ants $a = 1,\ldots, A$, each ant generates a solution path vector $S_a^{(1)}$ based on a probablistic criterion using the pheromone matrix components $\varphi_{ik}^{(0)}$. The $\boldsymbol{S}^{(1)}$ matrix is an $N \times A$ matrix that contains the solution (clusters) to which each data point is assigned for ant a. The generated solutions are determined by using the biased roulette wheel selection criterion with the probability of choice proportional to row values of the pheromone matrix $\phi^{(0)}$. The pseudo code to implement the biased roulette wheel is as follows [17]:

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

The set of centroids $\boldsymbol{C}_a^{(1)}$ is now obtained for each $S_a^{(1)}$ vector by averaging the data points assigned to a specified cluster. Hence, A clustering solution vectors and centroid sets are obtained. Each clustering solution is evaluated by a fitness function based on the sum of square errors:

$$\psi_a^{(m)} = \sum_{k=1}^{K} \sum_{x_i \in c_{ak}}^{N} \| x_i - C_{ak} \|^2, \quad \text{for } a = 1,\ldots, A \quad , \tag{4}$$

where m is the iteration number.

In this fitness function, lower values indicate better clustering solutions. Thus, the set of $S_a^{(m)}$ and $\boldsymbol{C}_a^{(m)}$ leading to the lowest fitness values are considered to be the best sets, defined as $\tilde{\boldsymbol{S}}_a^{(m)}$ and $\tilde{\boldsymbol{C}}_a^{(m)}$, respectively, and these replace the initial ones. Finally, the pheromone matrix is updated to $\phi^{(1)}$ by:

$$r_{ik}^{(m)} = \max\left\{d(x_i, \tilde{\boldsymbol{C}}_{ak}^{(m)}), \varepsilon\right\} \tag{5}$$

Differently from the initialization stage, the auxiliary variables $\varphi'^{(1)}_{ik}$ are calculated by adding a pheromone reinforcement term to (2):

$$\varphi'^{(m)}_{ik} = \varphi'^{(m-1)}_{ik} + \left(r_{ik}^{(m)}\right)^2 \quad . \tag{6}$$

The components of $\varphi'^{(m)}_{ik}$ of the pheromone matrix are then normalized with $m = 1$ to avoid continuous growth of pheromone components in the iterative stage:

$$\varphi_{ik}^{(m)} = \varphi'^{(m)}_{ik} \Big/ \sum_{k=1}^{K} \varphi'^{(m)}_{ik} \tag{7}$$

Successive iterations*:* To avoid losing the best solution sets, the solution and centroid set for the first ant (a=1) is set to equal $\tilde{\boldsymbol{S}}_a^{(m)}$ and $\tilde{\boldsymbol{C}}_a^{(m)}$. For the successive ants $a = 2,\ldots, A$, at each iteration m, solution vectors $S_a^{(m)}$ are generated based on the roulette wheel criterion as indicated in the first iteration.

The following operations are the same as the ones mentioned in the first iteration, with the obtaining of the set of clusters $\boldsymbol{C}_a^{(m)}$ and evaluating each ant's solution then

obtaining the best solution vector $\tilde{\boldsymbol{S}}_a^{(m)}$ and set of centroids $\tilde{\boldsymbol{C}}_a^{(m)}$. At the end of each successive iteration, the pheromone matrix $\phi^{(m)}$ is updated by following (5) through (7).

Stop criterion: An effective criterion in heuristic methods is to stop when there is no noticable improvement in the fitness function after a specified number of successive iterations. For the purpose of preventing excessive computation time, a user defined maximum number of iteration is adopted here.

Final clustering results: The assignment of data points to clusters is achieved by taking the index of the highest value in each row of the pheromone matrix ϕ. Hence, the final centroids can be obtained by averaging the data points assigned to each cluster.

2. 3. Validity Indices

In order to evaluate the resulted clusters, two validity indies namely, the Silhouette index (SI) and Clustering Dispersion Indicator (CDI) are used to identify compact and separate grouping of clusters that presents the optimal clustering quality [18].

2. 3. 1. Silhouette Index (SI)

The SI [19] calculates the silhouette width for each data point, average silhouette width for each cluster, and the average silhouette width for the entire data set.

For a given cluster C_k, this approach assigns a quality measure to each data point in C_k, known as the silhouette width. The silhouette width is a confidence indicator on the membership of the ith data point in cluster C_k and is defined by the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}, \tag{8}$$

where $a(i)$ is the average distance between the ith data point and all data points in the same cluster of C_k, and $b(i)$ is the minimum average distance between the ith data point and all data points not included in the same cluster. The $s(i)$ value will vary between $-1 \le s(i) \le 1$. A value close to 1 indicates that the data point i is classified to the right cluster, whereas a value close to -1 indicates the misclassification of that data point. A value close to 0 indicates that a data point contained within one cluster is at an equal distance away 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 calculated by:

$$S_j = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{9}$$

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

$$GS = \frac{1}{K}\sum_{i=1}^{K} S_j . \tag{10}$$

In order to choose the optimal number of clusters using the SI index, the clustering that presents the maximal GS is chosen.

2. 3. 2. Clustering Dispersion Indicator (CDI)

The CDI [19] is the ratio of the mean intra-set distance between data points in the same cluster ($\hat{d}(\Omega_k)$) and the intra-set distance between the cluster centroids ($\hat{d}(C)$):

$$CDI = \frac{1}{\hat{d}\left(C\right)}\sqrt{\sum_{i=1}^{K} \hat{d}\left(\Omega_k\right)} \tag{11}$$

Lower CDI values indicate better clustering results. However, an increasing number of clusters decreases the CDI value. A knee point can define the optimum number of clusters.

3. Methodology

The clustering of PVPLs is achieved by applying a machine learning methodology on historical time series data. This historical data consists of solar irradiance and ambient temperature at a certain site for past years. The time resolution should be able to capture the short-term fluctuations in the solar irradiance and ambient temperature. The data is then converted to daily PVPLs. The next step is to group together the PVPLs that have similar features and a representative for each group is determined. The representative PVPLs can then be used instead of the whole data set.

3. 1. Data Pre-processing

Input: Historical solar irradiance and ambient temperature for a location with proper time steps.

Output: Noise-suppressed daily time series of irradiance and ambient temperature.

Description: The solar irradiance and ambient temperature data for the past few years are divided into segments where each segment represents a day. The daily time series patterns of solar irradiance and ambient temperature are examined for normality.

3. 2. Data Conversion

Input: Noise-suppressed daily time series solar irradiance and ambient temperature.

Output: Time series of the corresponding AC power of the PV system (PVPL).

Description: The AC power output time series of the PV system can be estimated from the solar irradiance and ambient temperature time series data by using adapting an appropriate model.

3. 3. Data Segmentation

Input: Time series of PVPLs.

Output: Categorical segments of daily PVPLs.

Description: Each year can be divided into categorical segments (e.g., seasonal categories). The similar categories for each year are segmented together.

3. 4. Clustering of PVPLs

Input: Categorical segments of PVPLs.

Output: Different groupings of PVPLs from each clustering method.

Description: Each category of data is clustered by each clustering method. The results are different groupings of data, clustered according to the perspective of the applied clustering method.

3.5 Validation of Clustering

Input: Grouping results of each clustering method.

Output: Validity index values.

Description: The results of the clustering methods are evaluated by properly defined metrics and indicators (validity indices). An evaluation value with respect to the utilized validity index is an indicator of how well the clustering method grouped the data.

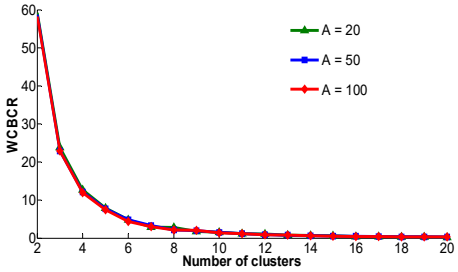

Figure 1: Values with respect to A = {20, 50, 100} for Ant Colony for two to 20 clusters.

Table 1: Validity indices of clustering algorithms for eight clusters of Fall.

Validity index	SI	CDI
Ant Colony	0.250	2.038

Table 2: Comparison of Ant Colony clustering algorithm w.r.t compactness, separation, and CPU for eight clusters on Fall data.

Algo.	Comp.	Sep.	CPU time(second)		
			Best	Worst	Average
Ant Colony	450.102	855.981	9.55	10.03	9.71

4. Simulation Results

The methodology was applied on data concerning three consecutive past years (2010-2012) with ten-minute time-steps of irradiation and ambient temperature from the Solar Radiation Research Laboratory [20]. The location of the obtained data has a latitude of 39.74ºN and a longitude of 105.18ºW. The irradiance data with this high time resolution (ten minutes) can lead to better accuracy due to the autocorrelation coefficients that will have higher positive values as compared to those obtained for data with lower time resolutions. Thus, the 10-minute time resolution will result in 144 observations per day.

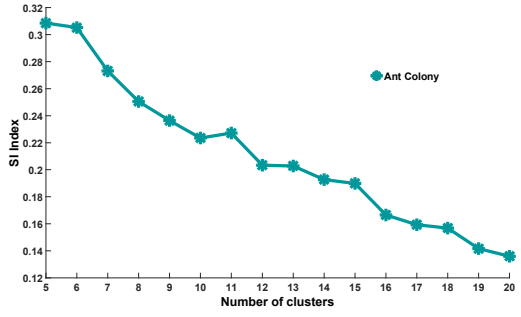

Figure 3: The best results of each clustering method for the fall data set of PVPLs for 5 to 20 clusters on the SI validy index.

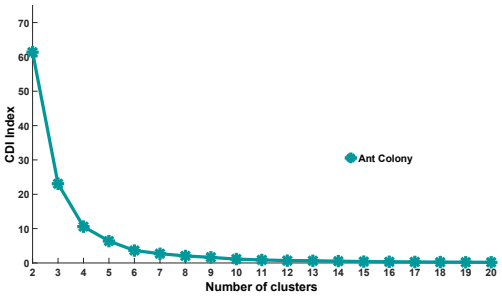

Figure 4: The best results of each clustering method for the fall data set of PVPLs for 2 to 20 clusters on the CDI validy index.

The results of the parametric analysis were determined on the number of ants in the initialization and successive steps. The algorithm has been performed with 50 repetitions with $A = \{20, 50, \text{and } 100\}$ and the solutions giving the best validity values were recorded. An overall best value when the number of ants increases has not been demonstrated (Fig. 1). However, increasing the number of ants increases the number of fitness evaluations, directly proportional to the number of ants and iterations.

For SI and CDI indices, the performance improved as the number of clusters increased. In addition, the Ant Colony cluster algorithm had relatively similar measures with respect to those indices. It can be observed that the utilization of CDI is slightly better than SI as it combines the distances of input data from the representative clusters and distance between clusters, which covers the mean-square error and SI characteristics. Consequently, the knee-point at the CDI index plot is of concern. It can be observed from Fig. 4 that eight clusters can be an optimum number of clusters to represent the fall PVPL data. The comparison of validity indices values of the clustering algorithms for eight clusters is shown in Table 1 and Table 2. It can be observed that the Ant Colony clustering algorithm with the CDI validity index presented the best performance on clustering the PVPL data and presenting representative PVPLs can then be used in studies related to integrating PV systems into the electric grid.

5. Conclusions

In this study a methodology to cluster PVPLs was presented in detail. The Ant Colony optimization algorithm approach was used in a dedicated formulation for clustering PVPLs. The main purpose of this study was to enhance the results of the clustering PVPLs. For the simulation results, the best combination that can present the optimum number of clusters was the Ant colony clustering algorithm with the CDI validity index. Together, significantly highly separated and well-compacted clusters were presented. Consequently, the resulted cluster representatives can be utilized in PV power system integration studies reducing the burden of extensive studies related to integrating PV systems into the electrical grid.

References

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

[2] IEA (2020), Renewables 2020 Analysis and forecast to 2025, IEA, Paris [Online]. Available: https://iea.blob.core.windows.net/assets/3350006e-c203-4b21-aa45-faefd36f22ad/Renewables2020-ExecutiveSummary.pdf

[3] 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.

[4] 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.

[5] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *Proc. Inst. Elect. Eng., Gener. Transm., Distrib.*, vol. 151, no. 3, pp. 395–400, 2004.

[6] A. Munshi, and Y. A.-R. I. Mohamed, "Comparisons among Bat algorithms with various objective functions on grouping photovoltaic power patterns," Solar Energy, vol. 144, pp. 254-266, 2017.

[7] G. Chicco, R. Napoli, and F. Piglione, "Load pattern clustering for short-term load forecasting of anomalous days," *Proc. IEEE PowerTech Conf.*, Porto, 10-13 Sept 2001, vol. 2.

[8] 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.

[9] 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," *Proc. WSEAS EPESE,* Corfu Island, Greece, 2008.

[10] 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 and Systems*, vol. 12, no. 7, pp. 1090-1104, 2008.

[11] G. J. Tsekouras, I.K. Hatzilau, J. Prousalidis, "A new pattern recognition methodology for classification of load profiles for ships electric consumers," *Journal of Marine Engineering and Technology,* no. A14, pp. 45-58, 2009.

[12] G. Tsamopoulos, N. Giannitsas, F. D. Kanellos, and G. J. Tsekouras, "Load estimation for war-ships based on pattern recognition methods," *Journal of Computations and Modeling*, vol. 4, no. 1, pp. 207-222, 2014.
[13] 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.
[14] F. J. Duarte, J. M .M. Duarte, S. Ramos, A. Fred, and Z. Vale, "Daily wind power profiles determination using clustering algorithms," *Proc. IEEE Power Syst. Tech.*, pp. 1-6, Oct. 30-Nov. 2 2012.
[15] G. Farivar, B. Asaei, N. Haghdadi, and H. Iman-Eini, "A novel temperature estimation method for solar cells," *Proc. PEDSTC*, pp. 336-341, 16-17 Feb. 2011.
[16] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, "An ant colony approach for clustering," *Analytica Chimica Acta*, vol. 504, pp. 187–195, 2004.
[17] A. P. Engelbrecht, *Computational Intelligence: An Introduction*, 2nd Edition, John Wiley & Son, 2007, pp. 135-137.
[18] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data", *Signal Processing*, vol. 83, pp. 825-833, 2002.
[19] *R. Jain, and R. Koronios, "Innovation in the cluster validating techniques," Fuzzy Optimization and Decision Making, vol. 7, no. 3, pp. 233–241, 2008.*
[20] Solar Radiation Research Laboratory (BMS) available online at: http://www.nrel.gov/midc/srrl_bms.