RES on Power Operation: K-Means Clustering Over Elbow Approach

ALI ALMOHAMMEDI and MOHAMMED ABIDO
Department of Electrical Engineering
King Fahd University of Petroleum and Minerals
Academic Belt Road, Dhahran, 31261
SAUDI ARABIA

Abstract: - In this paper, a new optimum number of operation modes is developed for low, medium, and high penetrations of renewable energy source (RES). The data-driven approach is improved considering the unsupervised learning approach of the K-Means classifier with the assistance of the elbow method to decide the number of optimum cluster of the unknown data mode of the power grid. On the other hand, a combined approach of raw data and a 2-dimension Principle Component Analysis (PCA-2) feature reduction approach is developed to improve the performance in terms of the operation mode switching frequency (OMSF), but the high-dimensional variance (HDV) is degraded as the RES penetration increases. The simulation results show that 1-dimensional PCA (PCA-1) can maintain the power system operation regardless of the penetration level of RES and the theoretical equations of the power system can be maintained under any penetration level of RES. In simulation, the proposed approach of combining the elbow method results in three, five, and seven clusters for low, medium, and high penetration levels of RES, respectively. As expected, this results in OMSF exceptionally reduction. However, the HDV is degraded in some scenarios of RES penetration. On the other hand, the PCA-1 approach results in constant three clusters for all RES penetration levels and the OMSF and HSV indices are maintained at the lowest value. Also, the seasonal consistency (SC) index is outperformed other techniques at the highest value.

1 Introduction

The renewable energy source (RES) is widely adopted on a large scale throughout the global power grid with the motivation of cost-effective and environmental friendliness. Fig.1 shows a typical wind farm that uses wind energy to generate electricity.

The analysis of the power operation mode can be divided into three stages. The first stage involves power system operation simulation and data preprocessing. A massive daily power operation data such as power flow, load demand, wind, and solar resources, and power generation dispatch schedule is obtained. The raw data is processed using Principle Component Analysis (PCA) technique to convert the high dimension power operation raw data vector to lower dimension for obtaining reasonable operation mode.

In the second stage, an operation mode clustering and dimension reduction are performed. The reduced dimensionality of power system mode is represented in a well-known mode and pattern that can be visualized to human operators.

The third and last stage is the quantitative evaluation where the final pattern is evaluated using theoretical equations of power system mode.

The paper is outlined as follows: In section 2, related work and main contribution are discussed. Section 3 explains the problem statement and theoretical equations. In section 4 involves the development of the operation mode of the power grid. In section 5, we evaluate the proposed work using KERAS in the Python platform.

Figure 1: Wind farm onshore (West Baker in Egypt).

2 Related Work

In recent times, many researchers have concentrated on analyzing the possible challenges and solutions of the extreme penetration of renewable energy to the power grid [1]. The operation mode of such a new power system, when penetrated with high RES, can be discovered by finding the patterns of data measurement considering the grid parameters, this approach is referred to as a data-driven method which has three categories, namely:

- Prediction: Data-driven method has been proposed to uncover load and RES correlations and other factors such as weather and speed of wind [4].
- Pattern Recognition: Online state estimation proposed for the growing uncertainty of smart grids [5].
- Decision making: Data-driven has been proposed in the decision making of power system operation and planning [3].

Some theoretical equations have been proposed and used to quantify the proposed 2-dimensional PCA (PCA-2), these equations are the high-dimensional variance (HDV), the operation mode switching frequency (OMSF), and the seasonal consistency (SC) indices [1]. The PCA-2 approach outperformed the original data clustering approach in terms of the OMSF index, but the HDV index had a bad result in the high RES penetration.

In this work, we claimed that the dimensionality of data clustering can be reduced further to 1-dimension using the PCA approach (PCA-1) which represents the power operation modes perfectly and provide outstanding results when evaluated using the aforementioned theoretical equation.

2.1 Main Contributions

In this work, the overhead of figuring out the suitable number of clusters are solved using the elbow method. When the PCA-1 approach is used, the effect of high RES penetration can be ignored and no special deterioration on the power system operation. The main contributions of this work are stated as follows:

- Visual determination of cluster number using Elbow Method: The proposal of a data-driven method for the operation mode of the power grid and pattern recognition in the scenario of high penetration according to highly complex data dimensionality based on K-Means classifier with the assistance of elbow method.
- The combination method of raw features (4-dimension features) and reduced feature

using PCA-2: Although, the combined method increases the space dispersion variance (HDV), but the operation mode switching frequency (OMSF) is reduced in the case of high penetration renewable power. The proposed method selects the best OMSF results during the year based on the best historical results of the SC index.

- In the high RES penetration, the OMSF and HDV indices are reduced while the SC index is improved: The development of 1-dimension PCA (PCA-1) instead of PCA-2 allows for the same cluster number when the RES penetration levels vary which prevent tremendous variation in the power operation mode.

3 Methodology

To study the impact of high renewable penetration on the power grid and determine the operation modes, three claims must be addressed:

- Power system operation data acquisition: The power flow study is obtained as well as the load and generator dispatch for further operation mode analysis.
- Pattern identification: The optimum number of patterns (or clusters) is revealed when tremendous operation modes are present.
- Visualization and evaluation: The reduction of high dimension and complex features of operation modes to two or three dimensions for easy visualization by the human eyes.

3.1 Problem Statement

The daily power system operation mode is provided as [1]

$$P = [g_{1\times(|g|\times T_s)}, r_{1\times(|r|\times T_s)}, f_{1\times(|f|\times T_s)}, \\ d_{1\times(|d|\times T_s)}, d^c_{1\times(|d|\times T_s)}] \qquad (1)$$

where g and r represent the operation mode data of generation schedule and actual renewable generation, respectively, f is the branch power flow, d and d^c are the operation mode of the load demand and the load curtailment, respectively, T_s denotes the sample reading period per day, and $|g|$, $|r|$, $|f|$, and $|d|$ are the total sample number of generators, renewable energy source (RES), transmission line, and busload, respectively. This data packet structure is varied upon the operation interest. For instance, it can include the total capacity of the solar or wind infrastructure.

The issue with the aforementioned unknown data structure is that the power operation modes are unknown and must be revealed.

Table 1: Load and renewable energy statistics of Germany's power grid in 2012, 2015, and 2018.

Type of Energy (MW)	2012 (x10⁶)	2015 (x10⁶)	2018 (x10⁶)
Load	516	516	517
Solar	27	34	41
Wind Offshore	0	11	19
Wind Onshore	45	65	89
Solar %	5.4 %	6.69 %	7.97 %
Wind Offshore %	0 %	2.31 %	3.69 %
Wind Onshore %	8.79 %	12.6 %	17.3 %
RES %	14.2 %	21.6 %	28.9 %
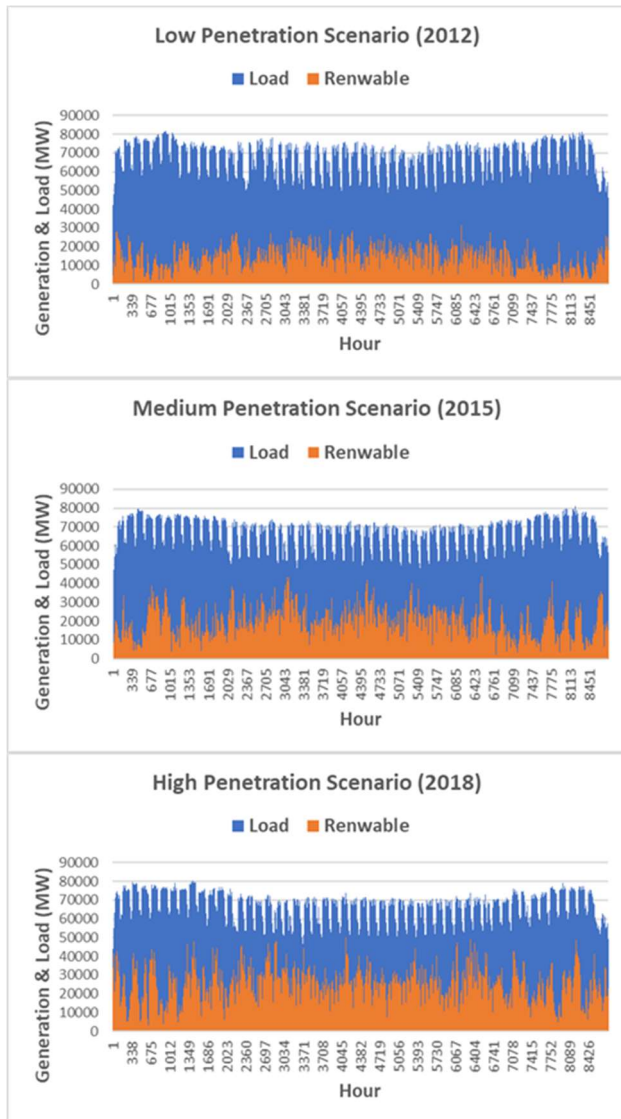
Figure 2: Hourly simulation of Germany power grid in the entire 2012, 2015, and 2018 for low, medium, and high RES penetrations, respectively.

3.2 Quantitative Evaluation

In this work, the analysis of power mode patterns is evaluated and determined using four criteria. Firstly, the high-dimensional variance (HDV) is given by

$$HDV = \frac{1}{M} Tr(\boldsymbol{P'} \boldsymbol{P'}^{T}) \tag{2}$$

where Tr denotes the trace operation and M represents the total sample number of daily power operation packet. When HDV gets high, the dispersion and diversity of power operation mode are high as well and vice versa.

Next, the operation mode switching frequency index (OMSF) is the measure of changes in operation mode patterns throughout the year. It is given by

$$OMSF = \sum_{i=1}^{M} I(p_i) \tag{3}$$

$$I(p_i) = \begin{cases} 1, & if\ p_i \in \Omega \ and\ p_{i+1} \notin \Omega \\ 0, & if\ p_i \in \Omega \ and\ p_{i+1} \in \Omega \end{cases}$$

where Ω corresponds to the operation mode p_i of the cluster. Hence, the operation modes have a large chance of change if the $OMSF$ index is high. The operators must be aware of the frequency of operation switching and ready to adjust the operation mode accordingly.

Eventually, the seasonal consistency (SC) is given by

$$SC_j = \frac{M_j - \sum_{i=1}^{M_j} I(p_i)}{M_j} \tag{4}$$

$$j \in \{spring, summer, autum, winter\}$$

where M_j is the total number of seasonal days.

Consequentially, the average seasonal consistency (ASC) is defined as

$$ASC = \frac{1}{4} \sum_j SC_j \tag{5}$$

$$j \in \{spring, summer, autum, winter\}$$

4 Power System Operation Mode

There are two types of unsupervised learning, namely:

- Clustering: It facilitates the identification of unknown structure in data.
- Dimensionality reduction: It retrieves the structural characteristics of the data.

In this sequel, we adopt the clustering approach to obtain the appropriate number of clusters for detecting patterns in an unlabeled dataset.

The two benefits of clustering that plays a significant role in this work are as follows:

Ali Almohammedi, Mohammed Abido

- Vector quantization: A short string of code that approximates the entire complex space is obtained.
- Finding meaningful structure in data: This helps in revealing salient grouping in data.

The following are the three common clustering algorithms:

- K-means (Known as Lloyd's Algorithm)
- EM for mixtures of Gaussian
- Agglomerative hierarchical clustering

We shall use the variant K-means algorithms as a powerful classifier. The cost function is mainly based on the Euclidean distance. The basic operations of K-Means are defined as follows:

Initialize:

- K-random samples are selected as the cluster centers.
- Alternate: For each sample, the Euclidean distance is calculated then assign the sample to the closest cluster center. After that, the average between cluster samples is recalculated for each cluster center and get updated.
- Stop Criteria: If there is no change in the sample assignment the operation is halted.

The problem in the aforementioned algorithm is how to recognize the correct number of clusters (or power system operation modes). This issue can be resolve by using the Elbow method called "Knee of a curve" which increases the K number of clusters gradually until the cost of Mean-Square-Error (MSE) is within a predefined threshold level [2].

In this work, different levels of RES penetration are evaluated in terms of MSE initially and the elbow method helps in varying the cluster numbers. The MSE is given by

$$MSE = \mathbb{E}\{|e(i)|^2\} = \frac{1}{N} \sum_{i=1, j \in K}^{N} (p_{i,j} - \mu_j)^2 \quad (2)$$

where $\mathbb{E}$ represents expectation operator, N is the total sample number of power operation packet during the year, μ_j is the mean of cluster j, and K is the total number of clusters. Furthermore, the performance of various RES penetrations is cross evaluated in the same plot to visualize the observations and decide the appropriate number of operation modes.

Table 2: Combination Method of 4D and 2D: summary and selection of SC in 2012, 2015, and 2018 (See Fig. 6)

Season	2012 (%)		2015 (%)		2018 (%)	
	4D	2D	4D	2D	4D	2D
Winter	.862	.861	.847	.854	.841	.827
Spring	.849	.850	.821	.832	.813	.819
Summer	.853	.853	.832	.84	.813	.815
Autumn	.856	.856	.857	.852	.834	.82

5 Simulation

In the simulations, we used KERAS in Python platform to validate the K-Mean unsupervised learning classifier. The power system data were obtained from a widely known benchmark website called the Open Power System Data (A free and open platform for power system modeling) as shown in Fig. 3. The selected data belongs to the power system of Germany (DE) for 2012, 2015, and 2018. The features of the data include actual load power (MW), actual solar generation (MW), actual wind offshore generation (MW), and actual wind onshore generation (MW).

The data obtained from each feature are sampled every hour and stacked in column vector of length 8,760 sample readings (24 hours per day X 365 days per year) in Megawatt (MW). The first scenario of the year 2012, the second scenario of the year 2015, and the third scenario of the year 2018 are referred to as the low, medium, and high RES penetrations, respectively, which are shown independently in Fig. 2.

The statistics of load generation and renewable energies (solar and wind resources) in DE are categorized in Table 1. The right number of clusters for each scenario is obtained with the assistance of the Elbow method as shown in Fig. 4. The pure raw data of four features (4-Dimensions) is reduced to two features (2-Dimensions) by invoking the PCA which is reduced further to one feature (1-Dimension) only. The number of clusters is varied from 1 to 10 clusters for simplicity. The error threshold level is fixed on the elbow edge to approximately 0.5×10^{12}, 0.3×10^{12}, and 0.1×10^{12} for the raw data, the reduced features PCA-2, and PCA-1, respectively. Notice that even though the error was raised to the order of 10^{12}, the actual raw data vector has every element in the order of 10^6, i.e., in MW, and any fractional mismatch between the estimated and actual output value would raise a value of 10^{12} by the definition of the MSE performance. To have a smaller value of error, we suggest normalizing the raw data vector.

Ali Almohammedi, Mohammed Abido

The best number of clusters that provide reasonable error for the low (the year 2012), medium (the year 2015), and high (the year 2018) renewable penetrations are 3, 5, and 7 clusters, respectively, for raw data (origin) and PCA-2. While for PCA-1, the selected number of clusters is 3 for all scenarios of RES penetration.

The cross-clustering of raw data has four features: the load power, the solar, the wind offshore, and the wind onshore features are demonstrated using the K-means classifier as shown in Fig. 5 for three scenarios of RES penetration. Fig.7 and 8 illustrate the clustering of the reduced feature using PCA-2 and PCA-1 with two and one features respectively. The resultant partition had independent and disjoint clusters without any overlap using PCA.

As a result, the operation mode of the power system with/without preprocessing the data using PCA were simulated in Figs. 6, 7, and 8.

Figs. 9, 10, and 11 illustrate the results of the $OMSF$, the HDV, and the SC indices respectively. The results of the preprocessed data using PCA-2 are compared to that of the raw data result which gives better OMSF result in medium penetration 2015 and poor result in high penetration 2018. But, the HDV results were better in medium and high penetrations.

The PCA-2 and PCA-1 outperformed other techniques in terms of $OMSF$, HDV, and SC in medium and high penetrations in 2015 and 2018 respectively. Also, a combined method of 4-D original raw data and PCA-2 was considered to switch between modes (4-D original data and PCA-2) as illustrated in Figs. 10 and 11. The $OMSF$ outperformed 4-D original data and PCA-2, but the HDV index was degraded when compared to PCA-2. The switch between modes is based on the highest result of SC seasonally. A particular year has four seasons (winter, spring, summer, and autumn) and the SC is illustrated in Fig. 9 for three RES penetrations. In 2012, the winter had 0.862% and 0.861 % SC results for 4-D and PCA-2 methods. To reduce the OMSF index, we picked the highest SC results which is 0.862% of the 4-D method in the combined method, and proceed in the same manner as shown in Table 2. The underline scores of SC for each season and every year were selected as the preferred clusters of the combined method.

6 Conclusion

This paper proposed a new method for the optimum number of operation modes in the scenario of high RES penetration based on the elbow method of k-means.
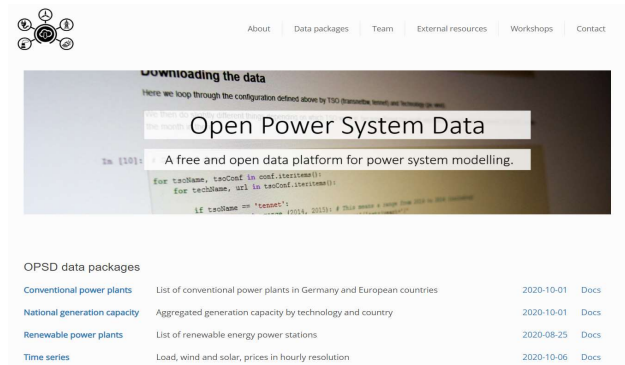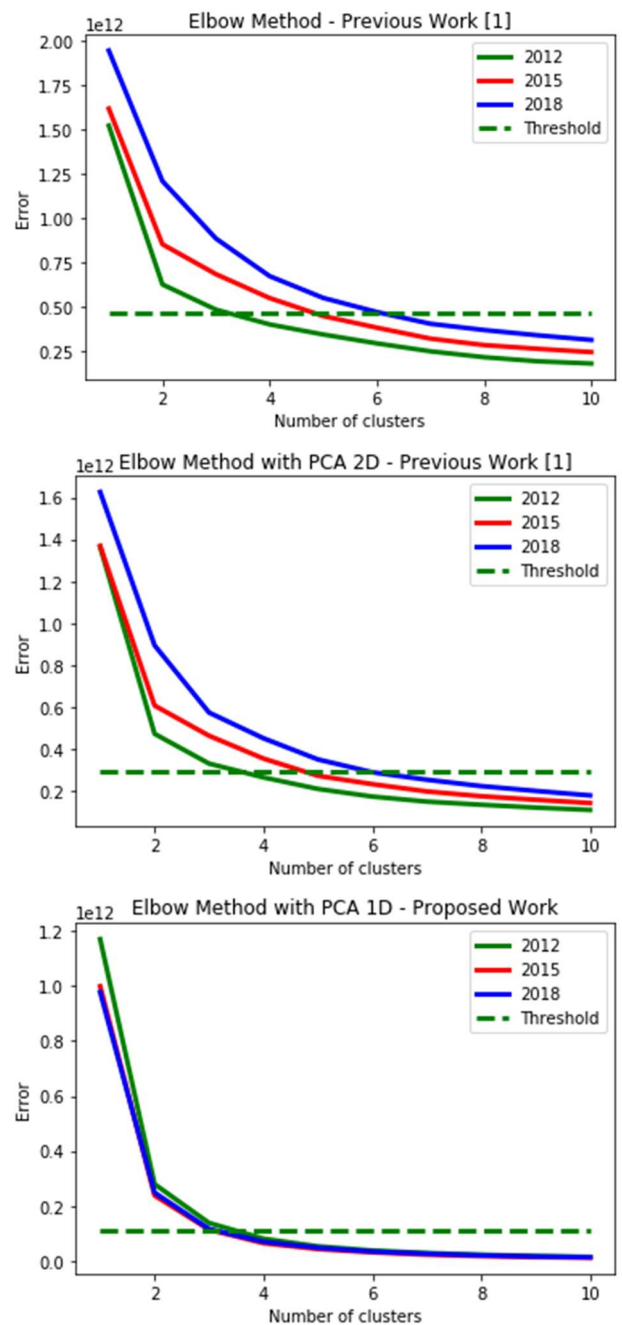

Figure 3: Open power system data website.

Figure 4: Elbow method of k-means clustering (Top), with PCA-2D (Middle), and with PCA-1D (Bottom).
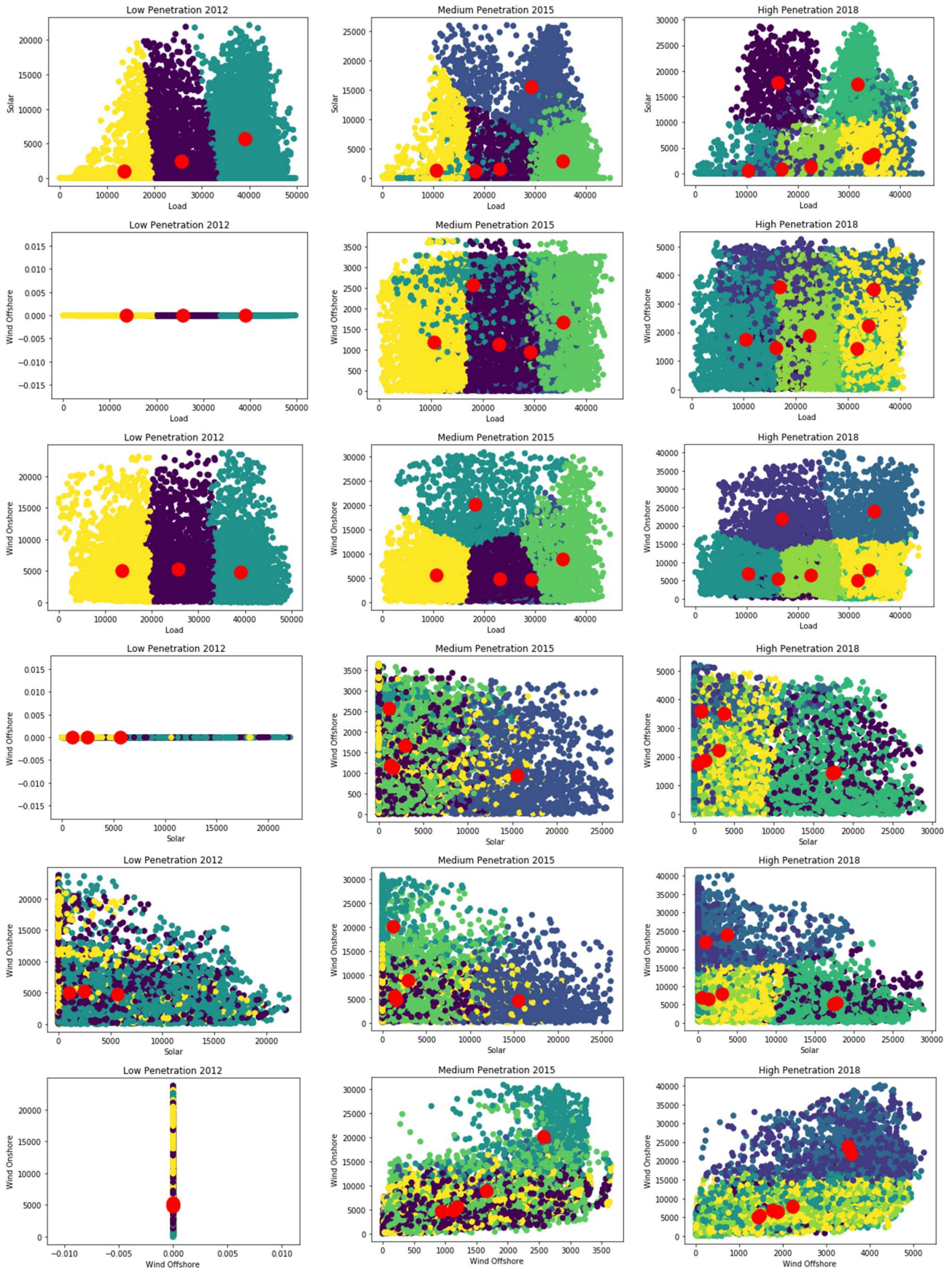
Figure 5: Clustering of the operation modes under low, medium, and high RES penetrations with 3, 5, and 7 clusters, respectively.
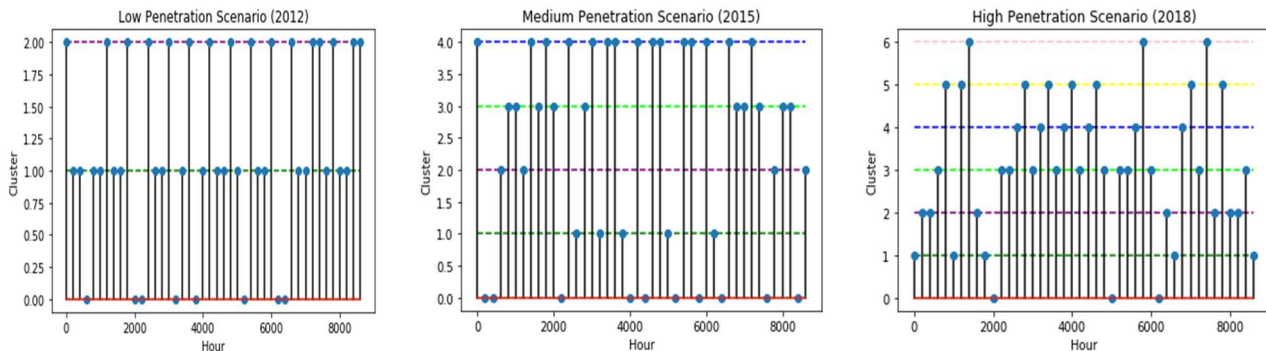
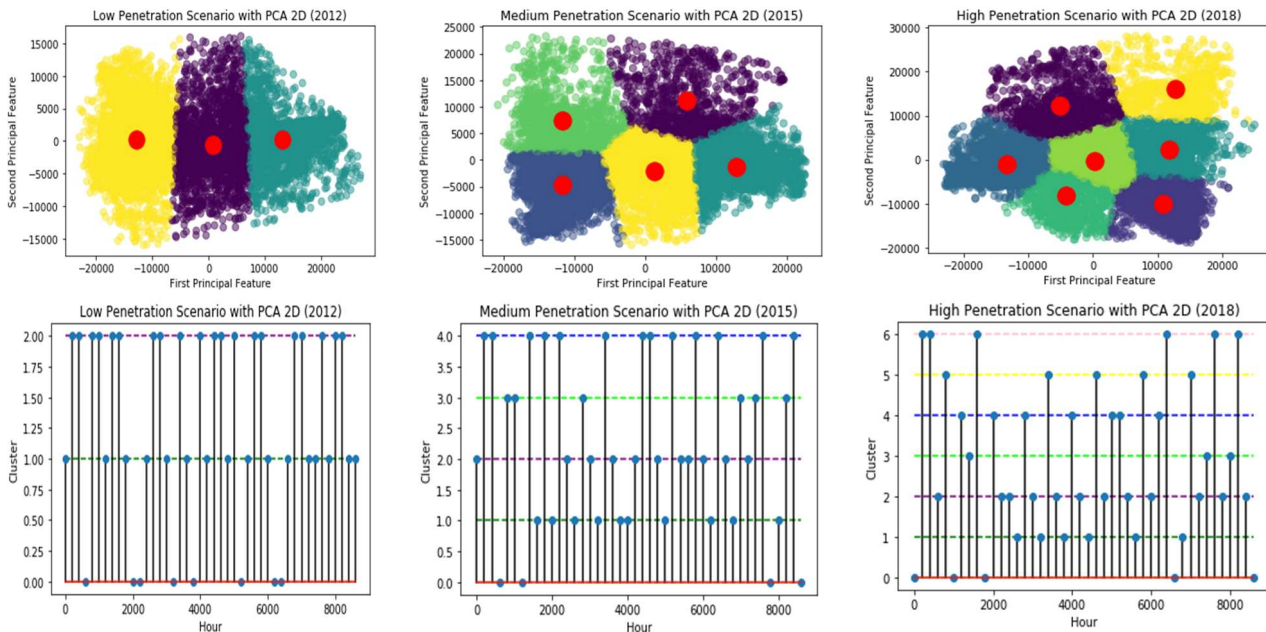Figure 6: The operation modes of DE power grid.

Figure 7: The operation modes of DE power grid using reduced feature PCA-2 with 3, 5, and 7 clusters.
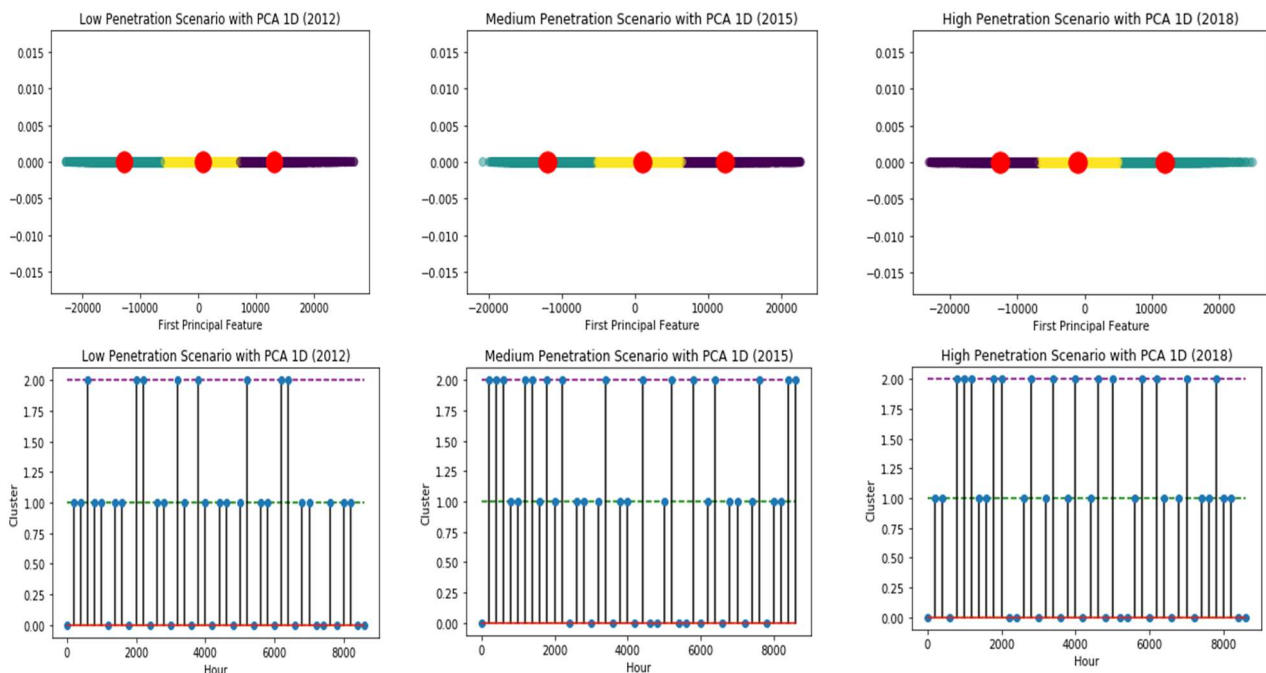

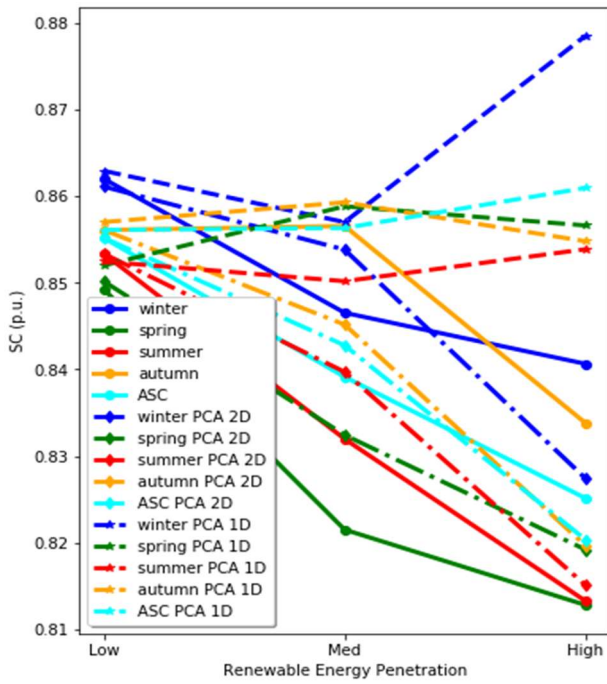Figure 8: The operation modes of DE power grid using reduced feature PCA-1 with constant 3 clusters.

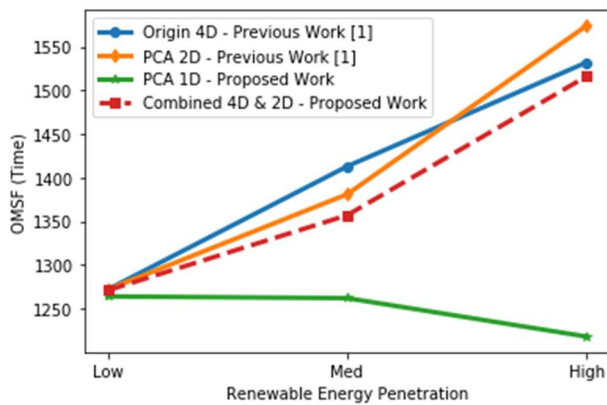Figure 9: SC of the operation modes of DE power grid when RES penetration varies.

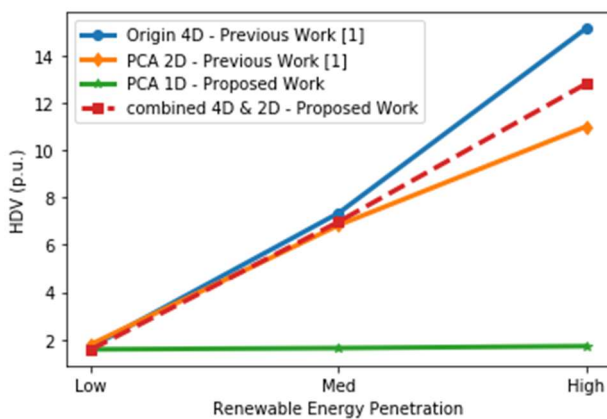Figure 10: OMSF of the operation modes of DE power grid when RES penetration varies.

Figure 11: HDV of the operation modes of DE power grid when RES penetration varies.

To reduce the dimensionality of the raw power operation vector, the combined method has been proposed based on the PCA-2 approach. It demonstrated a better OMSF index on the cost of bad HDV and SC indices.

Also, detailed simulations were carried out using the K-Means classifier where the operation mode of the high penetration of RES had 3, 5, and 7 clusters for low, medium, and high penetration of RES, respectively. Moreover, the combination method showed the lowest *OMSF* index but the SC index had a poor result as the RES penetration varies.

The proposal of the PCA-1 approach outperformed other techniques and all theoretical equations, i.e. OMSF, HDV, and SC indices, providing superior results. The results had constant results regardless of the level of RES penetration to the HDV index and inverse relation to the OMSF and SC indices.

Acknowledgments:

References:
[1] Q. Hou, E. Du, N. Zhang, and C. Kang, Impact of High Renewable Penetration on the Power System Operation Mode: A Data-Driven Approach, in *IEEE Transactions on Power Systems*, Vol. 35, No. 1, pp. 731-741, Jan. 2020
[2] X. Song, W. Li, D. Ma, Y. Wu and D. Ji, An Enhanced Clustering-Based Method for Determining Time-of-Day Breakpoints Through Process Optimization, in *IEEE Access*, Vol. 6, pp. 29241-29253, 2018.
[3] Mingyang Sun, Jochen Cremer, Goran Strbac, A Novel Data-Driven Scenario Generation Framework for Transmission Expansion Planning with High Renewable Energy Penetration, Applied Energy, Vol. 228, pp. 546-555, 2018.
[4] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li and Q. Xia, Data-Driven Probabilistic Net Load Forecasting With High Penetration of Behind-the-Meter PV, in *IEEE Transactions on Power Systems*, Vol. 33, No. 3, pp. 3255-3264, May 2018
[5] Y. Weng, R. Negi, C. Faloutsos and M. D. Ilić, Robust Data-Driven State Estimation for Smart Grid, in *IEEE Transactions on Smart Grid*, Vol. 8, No. 4, pp. 1956-1967, July 2017