Sustainable Energy, Grids and Networks 31 (2022) 100773

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

journal homepage: www.elsevier.com/locate/segan

An unsupervised learning schema for seeking patterns in rms voltage variations at the sub-10-minute time scale

Younes Mohammadi, Postdoctoral researcher [a,*], Seyed Mahdi Miraftabzadeh, Postdoctoral researcher [b], Math H.J. Bollen [a], Michela Longo [b]

[a] Department of Engineering Sciences and Mathematics, Luleå University of Technology, Skellefteå campus, Forskargatan 1, 93187 Skellefteå, Sweden
[b] Department of Energy, Politecnico di Milano, Via Lambruschini 4, 20156 Milano, Italy

ARTICLE INFO

Article history:
Received 2 February 2022
Received in revised form 22 April 2022
Accepted 14 May 2022
Available online 21 May 2022

Keywords:
Power-quality monitoring
Voltage variations
Seeking patterns
Time series clustering
Kernel PCA (KPCA)

ABSTRACT

This paper proposes an unsupervised learning schema for seeking the patterns in rms voltage variations at the time scale between 1 s and 10 min, a rarely considered time scale in studies but could be relevant for incorrect operation of end-user equipment. The proposed framework employs a Kernel Principal Component Analysis (KPCA) followed by a k-means clustering. The schema is applied on 10-min time series with a 1-s time resolution obtained from 44 different periods of a location south of Sweden. Then, ten patterns are obtained by reconstructing the 10-min time series from each cluster center. The results of the proposed schema show a good separation of cluster centers. Moreover, some statistical power-quality indices are applied to the whole dataset, showing voltage variation between (0.5–3) V over a 10-min window. Obtaining the most suitable indices and applying them to the ten obtained cluster centers and their belonging time series shows that the existing statistical indices may not be enough to show a complete picture of the sub-10 min actual variations. This outcome shows the necessity of extracting 10-min patterns through our proposed schema besides the existing statistics to quantify the voltage variations, levels, and patterns together. Findings of this paper are: Not forgetting the sub-10-min time scale; The necessity of employing both statistics and the proposed schema; Extraction of ten typical patterns; The need for the statistics and patterns that are justified as changes in equipment connected to the grid; and compressing a huge amount of data from power-quality monitoring. The proposed schema is applied to a much less understood phenomena/disturbance type so that this work will result in general knowledge beyond the specific case study.

1. Introduction

Voltage magnitude (rms value) deviation from its nominal voltage varies over a range of time scales. Power-quality monitoring (collecting measurement data) is an important part of the performance evaluation of power systems and a significant aspect of power-quality studies for network operators. An agreement for data collection with limited capacity has been restricted to 10-min values as defined in the IEC 61000-4-30 standard [1,2]. The overview of voltage-quality regulation in [3] also shows that 10 min is the most common value used. Moreover, there is a lack of performance necessities for voltage quality and knowledge on the levels for sub-10-min values (a time scale between 1 s and 10 min), and power-quality monitoring programs seldom include it. In contrast, this time scale should not be neglected because equipment susceptibility to sub-10-min variations changes. Moreover, tripping of PV installations due to overvoltages is seen for the values belonging to the 10-min time scale. Fast voltage variations and light flicker are also due to variations in this time scale [4–7]. Besides, many different types of equipment (generation or load) show voltage variations in this time scale: PV power installation [6,8,9], wind power installations [10,11], EV charging [12], and electric heat pumps [13].

A very short variation (VSV) index over the 10-min widow was calculated for a single-customer nanogrid in [14]. The measurement-based definition of individual rapid voltage changes (voltage steps) as standardized by IEC 61000-4-30 resulted in a number of publications discussing this voltage-quality event [8,15]. However, voltage steps are only one aspect of the sub-10-min variations. Later, some research has been done on statistical indices and actual levels for quantifying variations of rms voltage [6,16] and harmonic voltage [17] in a sub-10-min time scale. Such statistics are defined, for example, the 95th percentile of 1-s values of the rms or harmonic voltage over the 10-min window minus a 10-min rms value. As single-window or single-site, these statistical indices are appropriate for quantifying some

* Corresponding author.
E-mail addresses: Younes.mohammadi@ltu.se (Y. Mohammadi), seyedmahdi.miraftabzadeh@polimi.it (S.M. Miraftabzadeh), Math.bollen@ltu.se (M.H.J. Bollen), Michela.longo@polimi.it (M. Longo).

https://doi.org/10.1016/j.segan.2022.100773

Table 1
Single-window existing statistics used in this research.

Indices	Symbol	Explanation
Quantifying the range in value	R100	Highest 1-s value minus lowest value
	R98	99th percentile minus 1st percentile
	R90	95th percentile minus 5th percentile
	R80	90th percentile minus 10th percentile
Quantifying deviations from the rms (overdeviation)	P100	Highest value minus 10-min rms value
	P99	99th percentile minus 10-min rms value
	P95	95th percentile minus 10-min rms value
	P90	90th percentile minus 10-min rms value
Quantifying deviations from the rms (underdeviation)	P0	Lowest value minus 10-min rms value
	P1	1st percentile minus 10-min rms value
	P5	5th percentile minus 10-min rms value
	P10	10th percentile minus 10-min rms value
10-min very short variations	VSV	10-min sliding-window rms on 1-s very short variations
Standard deviations	Std.	10-min non-sliding-window rms on 1-s very short variations

kind of voltage variations but do not result in typical patterns of variations versus a 10-min time window. A complete picture of the sub-10-min time range needs to quantify not only the levels but also the patterns.

Power-quality monitoring can result in large amounts of data, especially where it concerns measurements at multiple locations over a long period. For example, one year of monitoring rms values (3 voltages, 3 currents, 1 s values) results in about 190 million data points per location, which will take a very high amount of data. Although manual analysis is possible, it is too time-consuming for multiple location measurements. Automatic analysis methods enable a continuous assessment of the power quality and other operational aspects without time-consuming human intervention. Recent developments in machine learning could identify such hidden patterns. In general, two sets of methods can be used for training as supervised and unsupervised [18,19]. The initial approaches, as supervised, needed a pre-labeled dataset. In [20–22], expert systems classify power-quality disturbances. Artificial intelligent-based methods used expert classifiers like support vector machines [23–25], ensemble learnings [26], and neural networks [27,28]. Automatic extraction of input features has been done as one step before the supervised classifiers in literature [29,30]. Seeking patterns, so-called time series clustering is part of unsupervised problems since labeling/assigning cluster numbers to the input dataset (e.g., time series of signal variations) is not possible/too time-consuming along with the errors. As observed in [29,30], the automatic extraction of principal features has a normally better role than manually extracted features (e.g., statistical indices) [1, 31,32] to group a dataset. There are many works done on time series clustering. e.g., clustering on the areas of big data in [33], multivariable time series clustering in [34–36], and clustering by using different tools than k-means and the Euclidean distance measurement criterion addressed in [37] as shape-based clustering and in [38] as fuzzy-based by using Distance Time Wrapping (DTW) as similarity measure criteria. However, a limited number of applications in power quality data measurement analysis have been found, e.g., a time-series clustering methodology for knowledge extraction in energy consumption data in [39], a clustering method for probabilistic evaluation of harmonic load flow in [40] and in [41] a k-means clustering for identification of distributed generation contribution. A deep autoencoder followed by a k-means clustering was applied in voltage harmonics with a 1-day time window by a 10-min resolution in [42–44] to seek the daily patterns. They are concerned with a rather well-understood phenomenon (daily variation in harmonic voltage) so that their method did not create any new general knowledge. Among the few unsupervised machine learning schemas applicable for power quality measurement analysis, none of them have been applied yet to seek patterns for rms/harmonic voltage fast variations in the sub-10-min scale, which is different from daily variational patterns.

On following the refs. [6,7,16,17] to learn more about the sub-10-min time scale, this paper proposes an unsupervised learning schema, scalable and computationally cheap, to seek the sub-10-min patterns in rms voltage variations at the time scale between 1 s and 10 min. The upper limit of the window (10 min) is defined in the power-quality monitoring standard, IEC61000-4-30, and is commonly used in power quality monitoring. The lower limit of the window (1 s) is not part of any standard nor commonly used. The 1-s period is partly set by the available measurement data, partly by the computation effort needed, and partly by the fact that standards and regulations exist for time scales below a few seconds. In this paper, it was decided to go for unsupervised learning because (a) the time window 1 s-10 min is still largely unexplored, (b) there is still a lack of knowledge and insight about the window (c) relevant information for labeling is missing. The unsupervised schema employs a Kernel Principal Component Analysis (KPCA) to automatically reduce the input feature size and generate principal features. A k-means clustering is employed to distinguish between the principal features into K clusters. A t-SNE (t-Distributed Stochastic Neighbor Embedding) is employed to visualize the principal and clustered feature vectors into two dimensions. The method is applied to measured 10-min time series with a 1-s time resolution obtained from 44 different periods of an apartment in a city in southern Sweden. The apartment is part of a 12-story apartment building with two elevators, 230-kV, and 50 Hz. Then, ten typical 10-min patterns are obtained by reconstructing the cluster centers. A statistical analysis of the obtained results shows that the proposed schema is effective in seeking sub-10-min patterns and confirms that a full representation of voltage variations at the sub-10-min scale needs both results of statistical indices and extracted patterns. In comparison with [42–44], our proposed schema is applied to a much less understood phenomena/disturbance type so that the work will result in general knowledge beyond the specific case study.

Section 2 of this paper presents the single-window existing statistical indices shortly. Section 3 describes our unsupervised learning schema in detail. Section 4 presents the measurement dataset and shows some examples of variations on the time scale below 10 min. A statistical analysis of the whole dataset, the results of the proposed schema, and statistical analysis of obtained cluster centers and their belonging samples are given in Section 5. Section 6 discusses the paper and suggests future works, and finally, Section 7 concludes the paper.

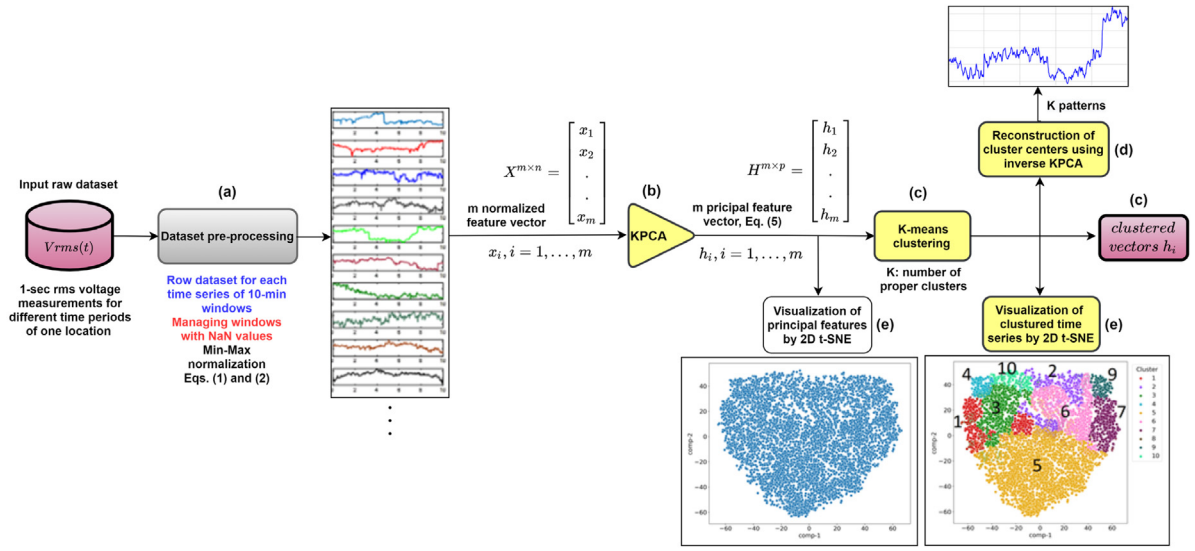


Fig. 1. Process of the proposed unsupervised learning schema for clustering 10-min time series.

2. Single-window existing statistical indices

In order to confirm the further results of our unsupervised schema and to have a full picture of the variations of rms voltage in a sub-10-min time scale, Table 1 explains the single-window existing indices [16,17] used in this paper. The indices quantify the range of 1-s values within a 10-min window and somehow show a picture of the variation of values. 1-s very short variations are a difference between 1-s rms voltages and the 10-min rms voltage. Also, a 10-min rms voltage is the rms of 1-s values within the 10-min window.

3. Proposed unsupervised learning schema

The process of the proposed unsupervised learning scheme consists of four modules optimized by the grid search, as shown in Fig. 1.

(a) Pre-processing measurement dataset; (b) KPCA on the feature vectors (normalized high-dimensional) x_i, which results in the vectors h_i with p principal features; (c) Using k-means clustering to group the principal features h_i; (d) Reconstruction of cluster centers using an inverse KPCA; (e) Visualizing the size reduced principal and clustered features in 2D space using t-SNE.

(a) Pre-processing the measurement dataset

The initial part of the proposed methodology is pre-processing of the dataset. First, the 1-s rms voltages are shaped within 10-min windows. Therefore, an input dataset matrix $X^{m\times n}$ (1) concludes, in which each row x_i includes a 10-min feature vector with 1-s resolution including n = 600 dimensions ($600\times 1\,s = 600$ s/10 min).

$$X^{m\times n} = [x_1, x_2, \ldots, x_m]^T,$$
$$x_i = [x_{i1}, x_{i2}, \ldots, x_{in}],\ i = 1, 2, \ldots, m, n = 600 \tag{1}$$

n is the dimension of each sample (time series or 10-min windows) and m is the number of samples. Second, the windows including missing data (NaN values), are managed by two different removing, and replacing actions. In this way, the windows with a high number ($\geq 50\% \times 600 = 300$ values) of NaN values are removed (24 windows equal to 0.32% of all windows in our measurement dataset), and the NaN values in windows with NaN values less than 300 are replaced to the window' mean value.

The proposed model is based on unsupervised learning algorithms (KPCA+k-means), which use the distance between data points to determine the similarity between samples. Considering a not normalized dataset, the higher weightage may be given to higher-magnitude features; consequently, the ML-based model will be biased toward such features, and the performance deteriorates. Hence, a Min-Max normalization (2) is applied to each column of matrix X from several possible approaches. In this way, each of the 600 features is considered as an independent coordinate which means that those samples with very high (low) 1s rms voltages will have values close to 1 (0). Each time series is scaled between [0,1] at the end of this operation.

$$x_{ij[0,1]} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})},$$
$$j = 1, 2, \ldots, 600, i = 1, 2, \ldots, m \text{ for each } j \tag{2}$$

(b) KPCA

After using Principal Component Analysis (PCA) [45] and sparse PCA [46] on the dataset, a KPCA algorithm was chosen and employed as a non-linear method. KPCA takes high-dimensional data sequences (x_i^{600D}) from $X^{m\times 600}$, maps data space $x_i (i = 1, 2, \ldots, m)$ to a higher dimension space $\Phi(x_i), (i = 1, 2, \ldots, m)$ and makes a non-linear function (kernel matrix $\boldsymbol{K}$ (3)). After checking the results of different kernels (Polynomial, RBF, Cosine, and sigmoid) [47], this study chose a cosine kernel (4). The next step is reducing dimension linearly. Hence, KPCA does an eigen analysis and projects the feature vectors on the first p dominant eigenvectors (principal components). Finally, the output of KPCA is determined to map the input to low-dimensional principal feature vectors ($h_i^{pD} = f_{for}(x_i)$) into $H^{m\times p}$ (5).

$$\boldsymbol{K}^{m\times m} = \begin{bmatrix} [\Phi(x_1), \Phi(x_1)], & \cdots & [\Phi(x_1), \Phi(x_m)] \\ \vdots & \ddots & \vdots \\ [\Phi(x_m), \Phi(x_1)] & \cdots & [\Phi(x_m), \Phi(x_m)] \end{bmatrix} \tag{3}$$

$$[\Phi(x_i), \Phi(x_j)] = \frac{\Phi(x_i)\Phi(x_j)^T}{\|\Phi(x_i)\| \, \|\Phi(x_j)\|} \tag{4}$$

$$H^{m\times p} = [h_1, h_2, \ldots, h_m]^T, h_i = [h_{i1}, h_{i2}, \ldots, h_{ip}], i = 1, 2, \ldots, m \tag{5}$$

In addition to the size feature reduction, this step may help in better initialization of centroids for k-means clustering [48]. The best mapping from the input of KPCA to the output is obtained after many runs of the function of *KernelPCA* by the *"cosine"* kernel in *Phyton* under parameter settings, where the best mapping is found as a minimum loss.

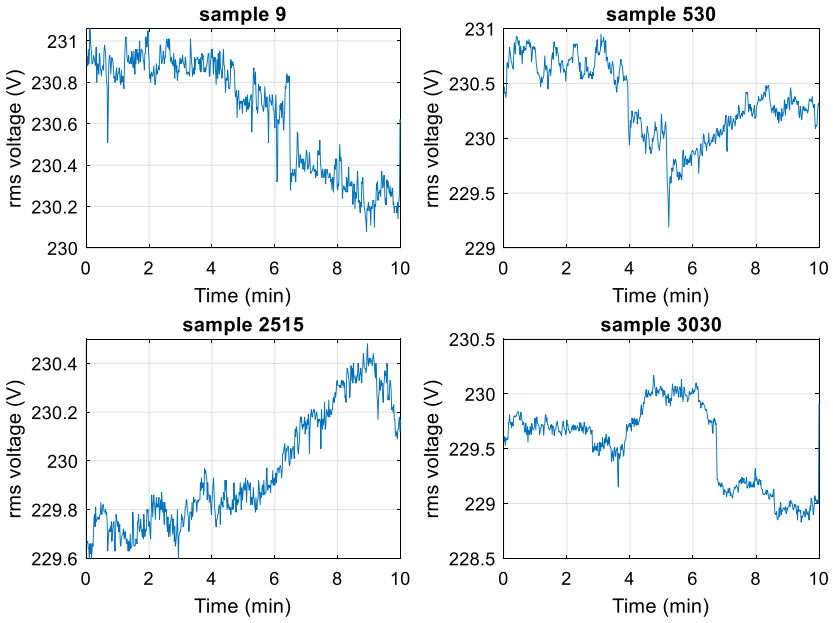


Fig. 2. Four examples of variations in rms voltage within a 10-minute window.

(c) K-means clustering

Principal feature vectors (h_i^{pD}) from the output of KPCA are inputted to the k-means clustering block. The *k-means++* initialization scheme [49] finds out K initial centroids (cluster centers) μ_j in an effective way. The k-means clustering aims to group the vectors h_i into K clusters (in our case, $K = 10$). Each feature vector is assigned to the cluster with the shortest 'distance' to one of the cluster centers. Centroids are then updated once all feature vectors are assigned. The k-means minimizes (6) the sum of the Euclidian distances of each h_i to its cluster centroid. This inertia is then trained by alternatively applying the following steps (7) until convergence:

$$\min \sum_{j=1}^{K} \sum_{i=1}^{m} \omega_{ij} \left\| h_i - \mu_j \right\|^2 \tag{6}$$

$$\omega_{ij} = \begin{cases} 1, & if\ j = argmin_j \left\| h_i - \mu_j \right\|^2 \\ 0, & otherwise \end{cases}, \quad \mu_j = \frac{\sum_{i=1}^{m} \omega_{ij} h_i}{\sum_{i=1}^{m} \omega_{ij}} \tag{7}$$

where the first part of (7) assigns each feature vector h_i to its closest μ_j, and the second part updates the μ_j by averaging all feature vectors within the jth cluster.

(d) Reconstruction of cluster centers

To further analyze the properties of clustered feature vectors, especially the representative data from each cluster, feature vectors from the centroids are fed to the inverse KPCA function to reconstruct the representative data sequence of each cluster, $x_j^{rec} = f_{inv}(\mu_j)$. Where x_j^{rec} is the reconstructed data sequence (600 dimensions) for the feature vector μ_j (jth cluster center with the dimension p). These reconstructed data sequences are used as the representative data patterns for the individual clusters.

(e) Visualization of principal and clustered features by t-SNE

To visualize the principal feature vectors (h_i^{pD}) (non-clustered) from $H^{m\times p}$ and the clustered principal vectors (clustered h_i), a t-SNE method [50] is used. T-SNE is another embedding method for converting high-D Gaussian distributed feature points into low-D (two/three) points in a t-student distribution. First, the similarity between two feature vectors, i and j belong to spaces non-clustered/clustered h_i are modeled by p_{ij} and q_{ij} in the input and output of t-SNE, respectively. The mapping is then obtained by minimizing the KL divergence between those two distributions:

$$\mathrm{KL(P \parallel Q)} = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{8}$$

In our case, using t-SNE is only for 2D visualization of feature vectors, principal and clustered principal to see how the proposed unsupervised methodology extracts K different patterns.

4. Measurement dataset

Time series of the 1-s rms voltage were obtained from recorded measurements of a location, which was an apartment in a city in southern Sweden. The apartment is part of a 12-story apartment building with two elevators. During a non-continuous eight-month period between 2017 and 2018, all measurements were performed at a wall outlet, 230-V, 50-Hz. A Metrum PQsmart portable monitor was used for the measurements. All measurements were following IEC 61000-4-30 Class A. 44 different periods of measurement, each period about 168 10-min windows, corresponded a total dataset consisting of 7380 10-min windows. By pre-processing the dataset and considering each window as an input sample, $m = 7356$ input feature vectors x_i (10-min sequences, containing 600 1-s samples each) is obtained and concludes matrix $X^{7356\times600}$ (removing only $7380-7356 = 24$ windows, i.e., 0.32% of all windows). Four different examples of the variations in rms voltage within a 10-min window from the input dataset, samples 9, 530, 2515, and 3030, are shown in Fig. 2. The variations can belong to a similar pattern or into some different patterns. Therefore, it is worth discovering underlying patterns from the dataset, where one may further find good interpretations coupled with physical reality. Moreover, this can obtain a picture of voltage magnitude variations at the time scales below ten minutes.

5. Results and analysis

5.1. Analysis of discussed single-window indices on the whole dataset

The probability distribution functions of the fourteen indices (Table 1), obtained over all 10-min windows of the whole dataset,

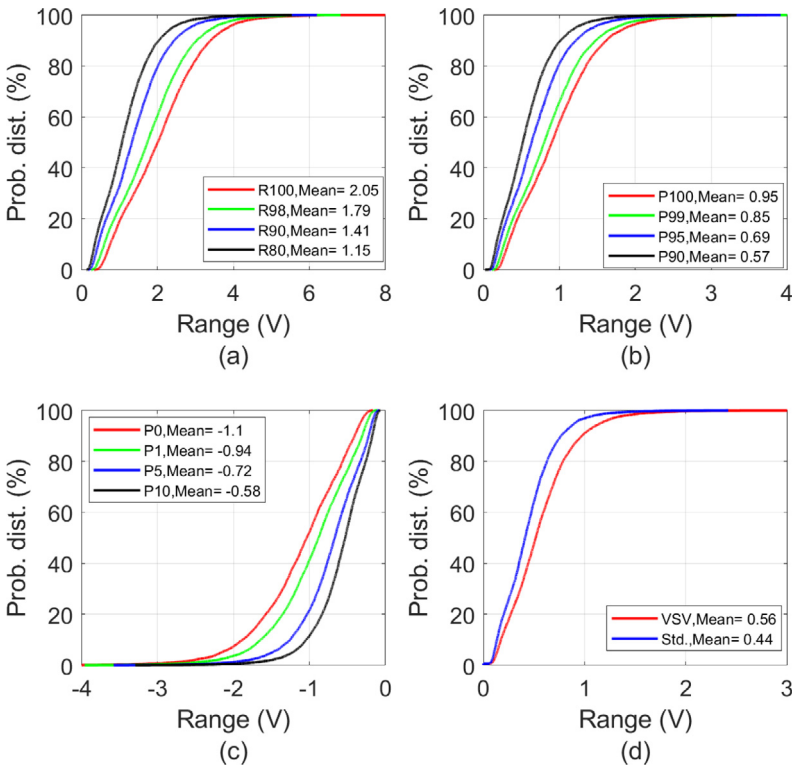


Fig. 3. A probability distribution for single-window indices over the whole dataset. (a) Indices quantifying the range in value (b) Overdeviation indices (c) Underdeviation indices (d) 10-min VSVs and standard deviations.

are shown in Fig. 3 to see how the general situation of the dataset looks like in the statistical view. The average value of the indices within the windows is almost equal to the values corresponding to the median. From Fig. 3a, voltage typically has varied by 0.5 V–3 V within a 10-min window. A range exceeding 1–2 V is common. The differences between higher-order and lower-order statistics for the R100, R98, R90, and R80 indices are higher and show a kind of variations within a 10-min window. For example, for the 80th percentile, the difference between R80 and R100 is 1.23 V and, for the median (50th percentile), the difference is 1 V. The distribution functions for Fig. 3a and b have a similar pattern (only a factor of two is the difference). For the 80th percentile, there is a difference of 0.56 V between P90 and P100 for overdeviation indices, but there is a 0.42 V difference for the median. The difference for the median between P0 and P10 (for underdeviations) is 0.5 V, as seen in Fig. 3c, which is higher than the value for overdeviations. As shown in Fig. 3d, VSV is slightly higher than Std., and there is only a 0.12 V difference for the median.

Fig. 4 shows a correlation matrix between all the fourteen indices mentioned above. A high value of the coefficients between two indices shows a strong correlation, and they vary together. The pairs R98-R100, P99-P100, and R80-Std. have the highest correlations (99%). By taking an average over the correlation between each index and other ones, the most suitable indices (strongest correlation) are calculated as R90 (from the range indices), P95 (from the overdevition indices), and P5 (from the underdeviation indices).

The authors of this paper tried to consider a feature engineering and inputted all the 14 statistical indices besides the 10-min windows as input features to the proposed unsupervised learning schema. However, the high correlation between some indices did not positively affect the clustering results. As a conclusion from Section 5.1, the most suitable indices along with Std. (which is somehow similar to VSV) are used, in the next section, for the

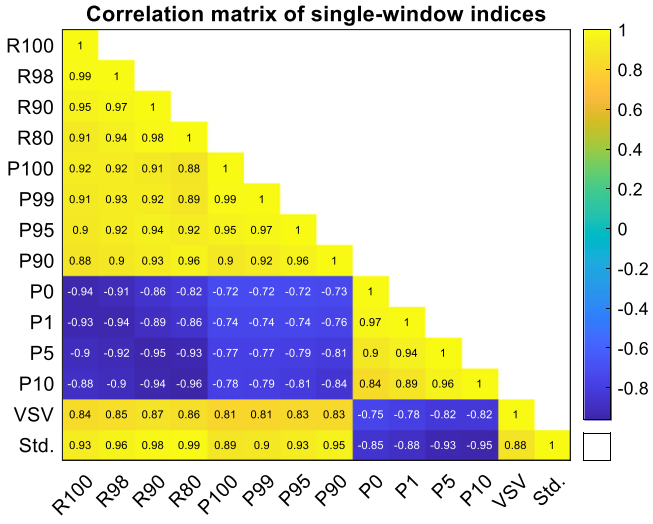


Fig. 4. Correlation coefficients between various single-window indices.

windows belonging to each cluster to show how the concluded patterns are proper and to confirm almost similar 1-s rms voltage variations within the windows belonging to each cluster.

5.2. Results of proposed unsupervised learning schema

After many runs of the function of *KernelPCA by "cosine"* kernel on the input dataset, the singular values corresponding to each singular mode index, related to every one of the first 50 principal components, are shown in Fig. 5. The singular values are equal to the 2-norms of the variables of principal components in the lower-dimensional space. In this study, we have considered only

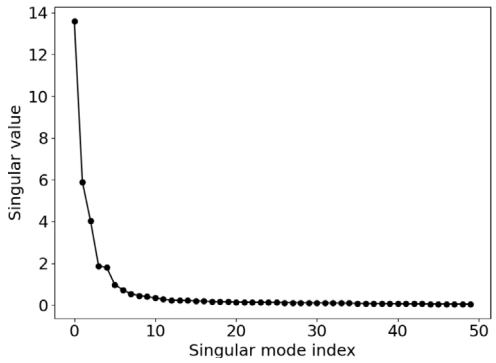


Fig. 5. Singular values corresponding to each singular mode index for KPCA.

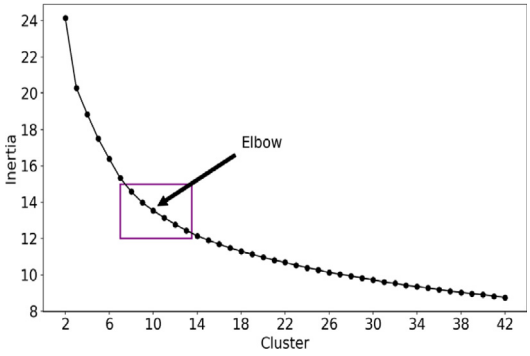


Fig. 6. The inertia of k-means vs. a different number of clusters.

the first $p = 10$ principal components (saving 82.38% information) during the training process of KPCA since almost the same results were obtained in comparison to 50 components, which could save 94.8% information.

Fig. 6 shows different inertia (within-cluster sum of squares criterion, a measure of how internally coherent clusters are, low values are better) versus a different number of clusters. The interval to choose the optimal number of clusters is marked by a rectangular, which is a number from 8 to 13 selected according to the knee of the curve. After checking the results of all the six numbers of K, $K = 10$ was chosen as the elbow point [51]. In this way, some good interpretations coupled with physical reality were found.

A function 2D t-SNE was used to visualize the principal feature vectors (10 dimensions) and the clustered principal features (10 dimensional), as shown in Figs. 7a and 7b, respectively. The parameters of the t-SNE (input:10D, output:2D) were set as Barnes–Hut algorithm, Euclidean distance metrics, perplexity = 30. The best 2D embedding space for visualization was selected by selecting the minimum loss values from running t-SNE 100 times. Ten clusters, colored in Fig. 7b, show 10 different possible patterns, whose overlap between some clusters shows a need for more dimensions (principal components more than three) to be plotted. However, each time series belongs to only one cluster (hard clustering). Later on, Fig. 9 will show a better visualization of the clusters.

The ten centroids (cluster centers), which are 10D time series of rms voltages, are an average of all the 10D time series belonging to each of 10 clusters. All ten centroids are reconstructed as 600D time series using an inverse function of KPCA. The reconstructed data sequences indicate representative patterns for the ten individual clusters, as shown in Fig. 8. Cluster 5 (Fig. 8e) includes 3287 samples, the maximum number of 10-min time series belonging to clusters (the dark yellow circles in Fig. 7b).

Cluster 8 (Fig. 8f) consists of only one sample. In order to check the behavior of this cluster, a plot of the two first principal components of KPCA has shown in Fig. 9. As seen in this figure, this type of plot shows a clearer visualization of clusters than t-SNE (Fig. 7b) and that cluster 8 is different from other clusters and may be considered an outlier. Hence, one may consider the number of obtained clusters as 9 and not 10. The observed different behavior for cluster 8 is related to two things: first, the cluster has a minimum 1-s value of rms voltage from 0 min–8 min between all input 10-min time series and the latter is regarding a different variation pattern. As seen in Fig. 8, the differences between the patterns are in the range of rms voltage (value/magnitude), the shape of variations (profile/growth pattern), and variations times. Cluster 4 (regardless of cluster 8) has a maximum range of variations. Clusters 6 and 9 seem to be similar, but the vertical axis range is different, and looking at Fig. 7b shows a separation of those clusters.

Moreover, the patterns realized in Fig. 8 are smoother than the real samples (Fig. 2) because there is an intrinsic characteristic of averaging in k-means clustering. In order to show that the patterns are a good representation of all samples, four normalized samples that belong to each of clusters 2, 4, 6, 7, 9, and 10 are shown in Fig. 10. The sample numbers are according to the row numbers of the input dataset. As seen in Fig. 10, the samples in each cluster can show some differences depending on the intra-class variance (associated with within-class spread). However, the overall patterns of the samples would remain largely the same.

5.3. Analysis of the reconstructed cluster centers and cluster samples

An analysis of the ten reconstructed cluster centers (patterns) and clusters' samples is explained in this section to show: (1) Well separation of the ten clusters; (2) The necessity of obtained patterns beside statistical indices over a sub-10-min time scale.

5.3.1. The similarity between the ten-reconstructed cluster centers

In order to show similarity/dissimilarity between the ten-reconstructed cluster centers, a correlation matrix is calculated and shown in Fig. 11. This similarity measure estimates the cosine of the angle between two centered (adjusted) cluster centers, therefore only checking the growth patterns and not magnitude differences. Actually, we used a Cosine similarity measure, unlike the Euclidian distance used in k-means. As marked in this figure, the maximum positive correlation is between cluster centers 3 and 4 because they have an almost similar growth pattern. However, Fig. 8c (cluster 3) and Fig. 8d (cluster 4) show a clear difference in the voltage ranges.

The Euclidian distance between them (real values, not normalized) is 61.82, and cluster 4 is totally below cluster 3 in terms of voltage magnitude. Another seen difference is the variation shape after $t = 7$ min. Moreover, t-SNE in Fig. 7b and 9 also show the separation of the two clusters. A maximum negative correlation is obtained for clusters 4 and 9. As seen in Fig. 8d, there is a very inverse behavior, which can also be seen in Fig. 11 since clusters 4 and 9 are placed in the opposite directions of principal component 1 in the t-SNE plot. The other correlation coefficients confirm distinguishing between the obtained cluster centers from the proposed schema.

5.3.2. Selected single-window indices for the clusters' samples

In order to show how well the 10-min time series are grouped into ten clusters and to display a homogeneity between the time series placed in each cluster, a statistical analysis is done. In this regard, the selected indices of R90, P95, P5, and Std. are employed (Section 4.1), and a probability distribution for

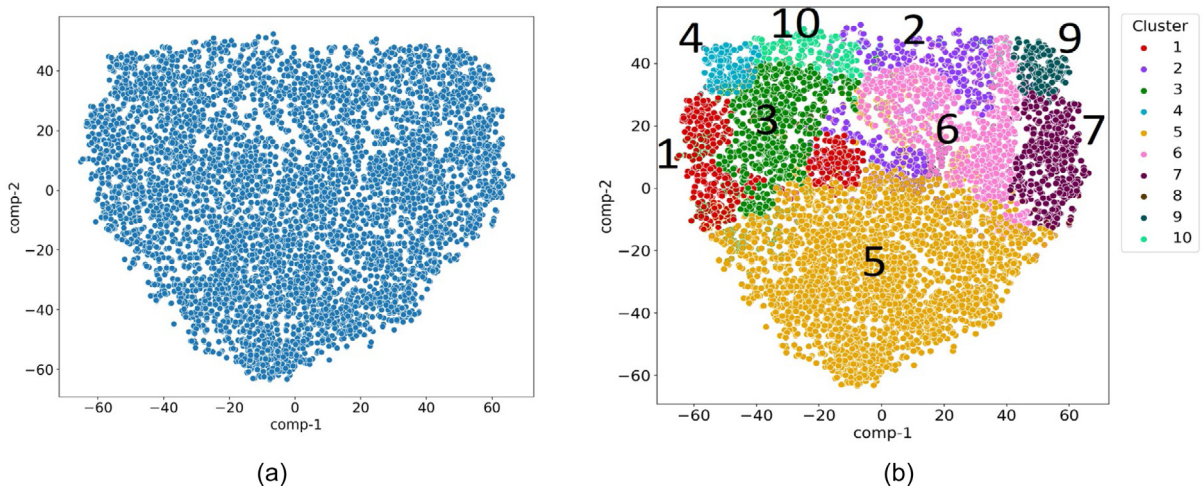


Fig. 7. Visualization of feature vectors (10D) by 2D t-SNE. (a) Principal (before k-means); (b) Clustered principal (after k-means), cluster 8 includes only one sample.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

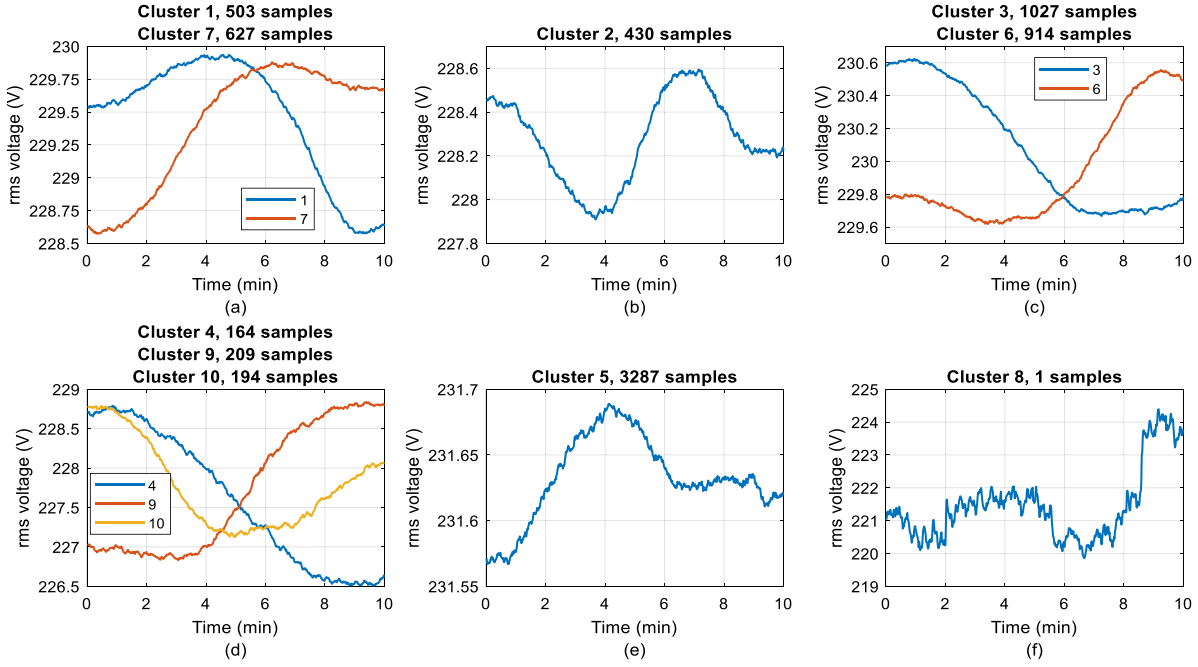


Fig. 8. Reconstructed patterns (cluster centers) including the number of samples belonging to each cluster. (a) Cluster 1 and 7; (b) Cluster 2; (c) Cluster 3 and 6; (d) Cluster 4, 9 and 10; (e) Cluster 5; (f) Cluster 8.

each cluster, including their samples, is shown in Figs. 12–15, respectively.

It is seen from all four indices that the probability distribution function for clusters 5 and 3 shows a softer curve, which is because the clusters are the most dominant ones with the highest number of samples. There is also a clear separation of the values for the indices between different clusters in terms of the probability distribution. This is a result of the well separation of the samples as grouped into ten clusters and displaying a homogeneity between the samples within each cluster. The indices, which show an almost similar result between a pair cluster, are explained here:

- R90: Clusters {1,7} (marked in Fig. 12)
- P95 and Std.: Clusters {1,7} and clusters {3,6} (marked in Figs. 13 and 15)
- P5: Clusters {1,7} and clusters {2,6} (marked in Fig. 14)

However, for cluster centers {1,7}, the shape of variations of rms voltage is completely different (Fig. 8a), and there is only a −30% correlation between those cluster centers (Fig. 11). Also, the shape of variations for cluster centers {3,6} is totally dissimilar (Fig. 8c), and a −67% correlation is seen in Fig. 11. Regarding cluster centers {2,6}, the shape and range of variations are different (Fig. 8b and c), and a small +29% correlation is obtained between them (Fig. 11).

5.3.3. Selected single-window indices for the ten-reconstructed cluster centers

In the following, the selected indices are applied to the reconstructed cluster centers, and the results in terms of range and a percentage of nominal voltage are given in Table 2. The indices showing an almost similar result between a pair of clusters are as follows:

Table 2
Selected single-window statistics for each cluster center.

Cluster center no.	R90		P95		P5		Std.	
	Range	(% V_n)	Range	(% V_n)	Range	(% V_n)	Range	(% V_n)
1	1.33	0.58	0.45	0.2	-0.88	-0.38	0.46	0.2
2	0.63	0.27	0.3	0.13	-0.33	-0.14	0.2	0.09
3	0.93	0.4	0.54	0.23	-0.39	-0.17	0.37	0.16
4	2.21	0.96	1.14	0.5	-1.07	-0.47	0.84	0.37
5	0.11	0.05	0.05	0.02	-0.06	-0.03	0.03	0.01
6	0.9	0.39	0.6	0.26	-0.3	-0.13	0.32	0.14
7	1.26	0.55	0.45	0.2	-0.81	-0.35	0.46	0.2
8	3.72	1.62	2.47	1.07	-1.25	-0.54	1.1	0.48
9	1.94	0.84	1.09	0.47	-0.85	-0.37	0.79	0.34
10	1.59	0.69	0.99	0.43	-0.6	-0.26	0.54	0.23

V_n: Nominal voltage; Colored highlights show close values of indices for different cluster centers.

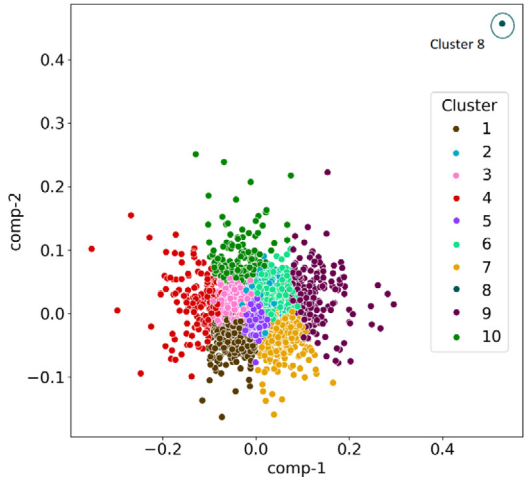


Fig. 9. Visualization of clustered principal feature vectors (10D) by only the first two principal components of KPCA. Note the place of cluster 8.

- R90, P95, and Std.: Clusters {1,7} and clusters {3,6}
- P5: Clusters {1,7,9} and clusters {2,6}

Concerning clusters {1,9}, there is a different range and variation shape (Fig. 9a and d) with a −82% correlation. For clusters {7,9}, the range and time of variations are dissimilar (Fig. 10a and d) even with a +75% correlation.

It is concluded from Section 4.3 that the patterns obtained from the proposed unsupervised schema are correctly separated, and their related time series have mostly similar behavior. Moreover, the statistics, the single-window indices, applied on cluster centers or their samples may wrongly consider some clusters in the same category and cannot distinguish between clusters. The correlation coefficients between cluster centers are a good measure to show the separation of the clusters. However, in this paper, clusters {3,4} are considered separately even if there is a high similarity in the variation shape between them, and that cluster 3 consists of only 164 samples compared with 1027 ones placed in cluster 4. This is because the range of variations of cluster 4 is totally below cluster 3.

As an overall conclusion, firstly, the variations in rms voltage at a time scale between 1-s and 10-min should not be neglected. Secondly, the statistical single-window indices may not be enough to show a full picture of the sub-10 min real variations. Hence, besides the statistics, extracting 10-min window patterns from a proper unsupervised framework (proposed schema in this paper) is essential.

6. Discussion and future works

6.1. Application of the unsupervised learning schema

The work presented in this paper uses time series of 1-s rms voltages over a 10-min window. According to the results of this paper and literature [16], oscillations are happening in the sub-10 min period. As one way to show the oscillations, this work seeks the sub-10 min patterns of rms voltage using a proposed unsupervised method. In that way, some important patterns in the large data of low voltage variational measurements were found. Compressing a huge amount of data from power-quality monitoring is also done in this study since the input data size is reduced from 600 to 10 by a factor of 60. The proposed schema, besides the statistics, is the first step before studying the potential impacts of the sub-10-min variations on equipment and quantifying the grid's hosting capacity for different types of new equipment connected to the grid. The proposed schema is scalable and computationally cheap, which makes it appropriate for extracting typical patterns in the big data domain.

Compared with some literature [52–56] that uses pre-defined scalar features in a labeled dataset to locate the source of voltage dips, one important application of our proposed schema can be solving that two-classes problem in an unsupervised intelligent schema. In this way, the input of KPCA is the rms waveforms of voltages and currents (merged as one signal or into two separate signals) with a high time resolution over a selected window, including pre and after dip cycles. The output of k-means will show two groups, with which one expert can label the group members as downstream and upstream as the source location of voltage dips.

6.2. The number of clusters

The k-means clustering method requires the user to select the number of clusters in advance. In our study, the interval to select an optimum number of K is a number between 8–13 (Fig. 6), and through multiple tries, $K = 10$ is chosen so that some good interpretations coupled with physical reality are found. Table 3 compares two numbers of K, 10, and 13 and their samples. As seen in this table, there is not much difference between the

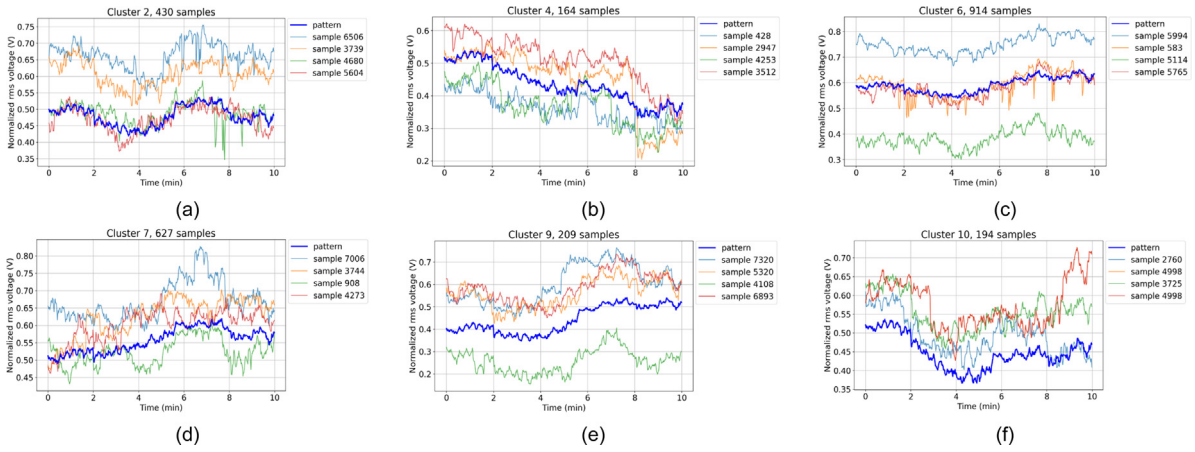


Fig. 10. Several reconstructed normalized samples in 6 clusters. (a) Cluster 2; (b) Cluster 4; (c) Cluster 6; (d) Cluster 7; (e) Cluster 9; (f) Cluster 10.

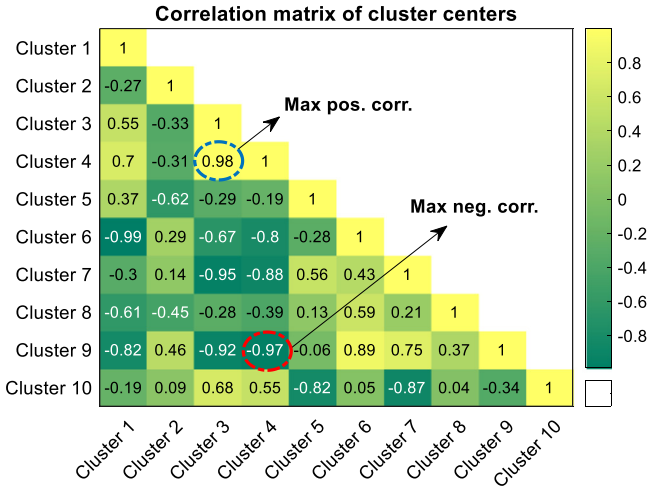


Fig. 11. The similarity between centered cluster centers.

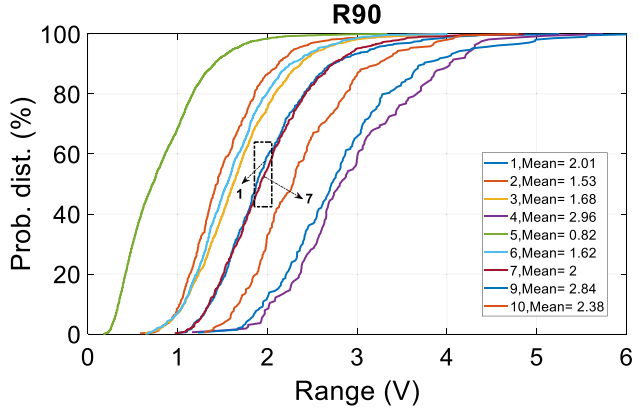


Fig. 12. Probability distribution of R90 for the time series of each cluster.

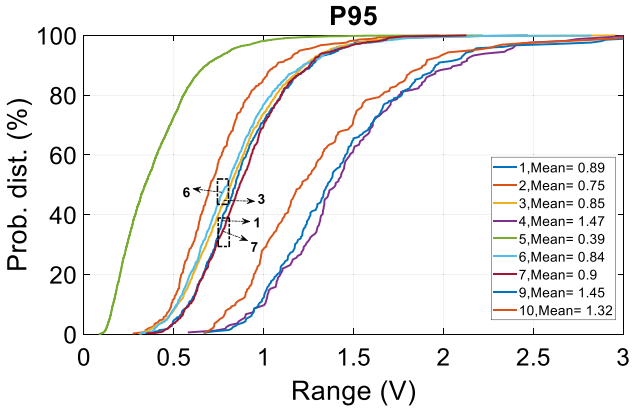


Fig. 13. Probability distribution of P95 for the time series of each cluster.

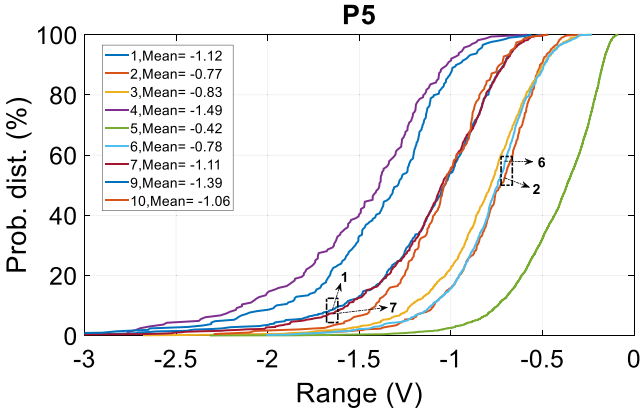


Fig. 14. Probability distribution of P5 for the time series of each cluster.

results, and by choosing $K = 13$, we might just be splitting the clusters for no good reason. Also, for both numbers of K, the cluster with 1 sample is seen as an outlier. One may consider only 4 clusters; however, choosing $K = 10$ (from Fig. 6), the possible sub-patterns showing different values and variation shapes under some clusters are extracted, where feature vectors can display large within-cluster variance.

A future implementation of the scheme could allow the user to select the number of clusters so that an interpretation of the results is possible.

6.3. Sub-10-min variations from the view of the unsupervised learning schema

The samples (10-min windows) within the clusters differ depending on the intra-class variance. However, the overall pattern of samples remains largely the same. Since each pattern is an average of its own samples, the fewer the number of samples

Table 3
Number of samples and a percentage of the total number of samples for two different numbers of K.

K		Cluster no.												
		1[a]	2	3	4	5	6	7	8	9	10	11	12	13
10	No. of samples	3287	1027	914	627	503	430	209	194	164	1			
	(% of total samples)	44.68	13.96	12.43	8.52	6.84	5.85	2.84	2.64	2.23	0.01			
13	No. of samples	2760	1174	905	519	386	379	364	356	248	127	74	63	1
	(% of total samples)	37.52	15.96	12.30	7.06	5.25	5.15	4.95	4.84	3.37	1.73	1.01	0.86	0.01

[a]The number of samples/clusters is sorted to show a better comparison between two numbers of K.

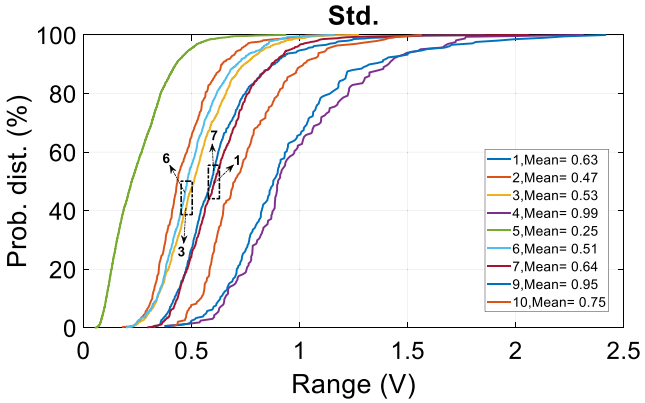


Fig. 15. Probability distribution of Std. for the time series of each cluster.

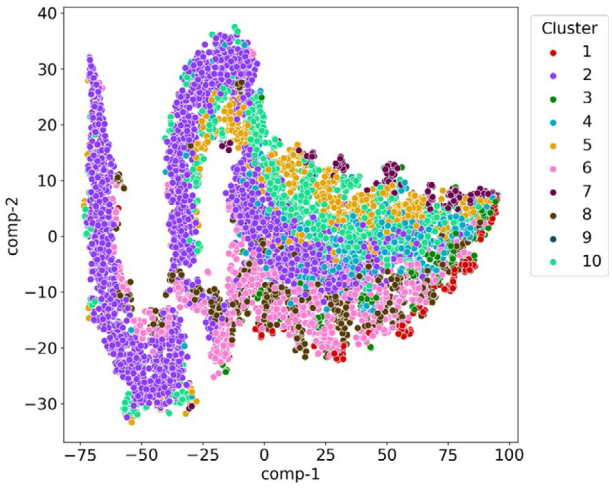


Fig. 16. Visualization of feature vectors (600D) by 2D t-SNE without using KPCA.

within a cluster, the more similar those samples are to the reconstructed cluster center (pattern). Hence, all clusters except 3, 5, and 6 look very similar to their own samples. Pattern 5 is very common (small random variations in rms voltage), and patterns 3 and 6 are less common (a negative/positive ramp or a single curvy step superimposed on a very small random variation).

Each of the ten patterns is a representative of a number of 10-min windows. According to the actual recordings, they can be considered "ten typical patterns". However, to say this with more confidence, there is still a need for more 10-min windows over months, seasons, or one year. Moreover, the number of clusters should also be selected carefully. From the output of the unsupervised schema, the samples 9, 530, 2515, and 3030 shown in Fig. 2 belong to the patterns 3, 3, 6, and 1, respectively. The real samples are somehow following the related patterns. Nevertheless, although the samples' overall patterns within clusters 3, 6, and 1 are similar, the averaging of many samples within the clusters generates smooth cluster centers, making some differences in oscillation values between the patterns and real samples.

6.4. Other feature-size reduction tools, distance measurement, and clustering methods

The KPCA with "Cosine" kernel, as a simple tool, was employed in the proposed unsupervised schema to reduce the feature size from 600 to 10. Meantime, other tools like linear PCA and sparse PCA were also investigated. Nevertheless, they were not successful in separating and unfolding input data as a tool to help k-means clustering. Moreover, the results without using KPCA and only k-means were checked. Fig. 16 shows a 2D visualization of input data for 10 clusters. As seen in this figure, there is no good separation between clusters compared to Fig. 7b (KPCA+k-means). A selection of 2 clusters would be effective for Fig. 16 (only k-means), which cannot show the real patterns of our dataset with 7356 input samples.

Other feature size reduction tools like a deep autoencoder (DAE) can be considered in future works instead of the KPCA to see how the clustering results would be for high-resolution time series. However, using a DAE is not as simple as the used KPCA. Table 4 shows a structured-based comparison between our proposed schema (KPCA+k-means) and (DAE+k-means) [42]. As shown in this table, the proposed schema/ref. [42] reached 10/2 clusters/patterns while having a higher/fewer number of input samples, longer/shorter window length, higher/lower time resolution, less/more principal components, and somehow a slower/faster workstation. The findings from Table 4 confirm the much simpler schema of the proposed method.

Although the proposed method is intended for off-line use and aimed initially at obtaining general knowledge about a new phenomenon, Table 5 shows the detailed time required for running the proposed schema and the method in [42]. The total running time of the proposed schema is about 20% of the total running of the [42].

On the other hand, replacing the Euclidian distance used in k-means clustering by DTW [38] can be another future work. Using the k-medoid method [57] instead of k-mean can be another way to obtain patterns. In this way, the most center of each cluster is selected as the cluster pattern.

6.5. Other variations

The proposed unsupervised learning schema was illustrated by applying it to rms voltage variations over a 10-min period. Since the schema is structured in such a way that its application is possible for any kind of signal, voltage, and current, future studies can seek the patterns for variations in harmonic voltage/current in a sub-10 min period.

6.6. Supplementary works

The measurements used in this study were related to 44 different periods in one location. In order to find more actual patterns, measurements from a number of locations around the

Table 4
A comparison of our proposed schema and ref. [42].

Method	Case study	No. of input samples	Window length	Time resolution	No. of principal components	K in k-means	Used workstation
Proposed schema	rms voltage	7356	600 (10 min)	1 s	10	10	Intel-i7 8700 K–3.7 GHz×2 CPU, 16 GB RAM, NVIDIA GeForce RTX 2080, Ubuntu 20.04.3 LTS-OS
[42]	Voltage harmonics[a]	365	144 (24 h)	10 min	16	2	Intel-i7 3.4 GHz×12 CPU, 48 GB RAM, NVIDIA Titan Xp 12GB GPU.

[a] V_2, V_3, V_4, ….

Table 5
Time required for different parts of the proposed schema to seek patterns.

Name	Time (s)	
	Proposed schema	[42]
Training KPCA/DAE and feature extraction	10	142.88
Clustering (100 runs)	0.17	0.15
Reconstruction from cluster centers	0.2	0.29
t-SNE	35 (100 runs)	78.86 (50 runs)
Total	45.37	222.18

world are recommended, which is an ongoing work of the authors of this paper. Using the four selected statistical indices in this paper as input features added to each 10-min window may make more complete patterns, while adding all the fourteen indices could not help in clustering results. Quantifying the variations by installing the PQsmart monitor close to solar power, electric vehicle charging, wind power, electric heat pumps, etc., which can easily change the variation patterns, is needed for future work. Measurements at higher voltage levels like the medium voltage at industrial installations and the impact of variations on the connected equipment are also needed for extra investigations. A possible future direction is to link actual events in the grid or at the load side with characteristics of the sub-10-min variations. However, this will require a bigger data collection.

7. Conclusion

This paper presented an unsupervised learning schema to find patterns in rms voltage variations at the time scale between 1 s and 10 min. First, an analysis of fourteen statistical single-window indices was made on 44 different periods of rms voltage measurements obtained from a location south of Sweden. It was shown that the voltage typically varies between (0.5–3) V over a 10-min window, in which a range exceeding 2 V is common. Calculated correlation coefficients between the fourteen indices showed a high correlation, indicating high similarity. Also, the most suitable ones were obtained: R90, P95, P5, and very short variation or standard deviation. Later, in order to show a full picture of sub-10-min variations, an unsupervised learning method was proposed to seek the sub-10-min patterns. The method consisted of a kernel principal component analysis, followed by principal feature clustering using a k-means algorithm. Ten typical patterns were obtained by reconstructing the 10-min time series from each clustering center. The results showed that the proposed scalable schema effectively extracted patterns, and a correlation matrix between the reconstructed cluster centers confirmed a good separation between them. It is worth mentioning that the unsupervised schema can be applied to any kind of signal.

Applying the most suitable statistics separately to the obtained cluster centers and their belonging samples showed that the existing statistical indices cannot be enough to show a full picture of the sub-10 min actual variations. Henceforth, besides the statistics, extracting 10-min window patterns from our proposed schema would be vital. According to the actual recordings, the patterns can be considered “ten typical patterns”. However, a trade-off between the number of clusters and the number of 10-min windows in a big dataset would be a vital task for future works. Future work must study potential impacts on the equipment after quantifying the variations, levels, and patterns.

CRediT authorship contribution statement

Younes Mohammadi: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Seyed Mahdi Miraftabzadeh:** Methodology, Software, Investigation. **Math H.J. Bollen:** Conceptualization, Investigation, Writing – review & editing. **Michela Longo:** Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The measurements presented in this paper were performed by a PQsmart portable monitor provided by Metrum Sweden AB.

References

[1] M.H.J. Bollen, IYH. Gu, Signal Processing of Power Quality Disturbances, 2005, http://dx.doi.org/10.1002/0471931314.
[2] M. Bollen, J. Milanović, N. Cukalevski, CIGRE/CIRED JWG C4.112 - power quality monitoring, Renew. Energy Power Qual. J. (2014) 1037–1045, http://dx.doi.org/10.24084/repqj12.011.
[3] CEER, Ceer Benchmarking Report on the Quality of Electricity and Gas Supply-2016: Gas-Technical Operational Quality, 2016, pp. 138–201.
[4] J. Schlabbach, D. Blume, T. Stephanblome, Voltage Quality in Electrical Power Systems, IET, The Institution of Engineering and Technology, Michael Faraday House, Six Hills Way, Stevenage SG1 2AY, UK, 2001, http://dx.doi.org/10.1049/PBPO036E.
[5] Guide to Quality of Electrical Supply for Industrial Installations, 1999.
[6] O. Lennerhag, M. Bollen, S. Ackeby, S. Rönnberg, Very short variations in voltage (timescale less than 10 min) due to variations in wind and solar power, in: Int Conf Exhib Electr Distrib 15/06/2015-18/06/2015, 2015.
[7] M.H.J. Bollen, I.Y.H. Gu, Characterization of voltage variations in the very-short time-scale, IEEE Trans. Power Deliv. 20 (2005) 1198–1199, http://dx.doi.org/10.1109/TPWRD.2005.844253.
[8] S. Lodetti, J. Bruna Romero, J. Melero, Methods for the evaluation of new power quality parameters: a review of rapid voltage changes and supraharmonics, 2019.
[9] B. Bletterie, T. Pfajfar, Impact of photovoltaic generation on voltage variations-how stochastic is PV, in: CIRED 19th Int Conf Electr …, 2007, pp. 21–24.

[10] J. Widén, N. Carpman, V. Castellucci, D. Lingfors, J. Olauson, F. Remouit, et al., Variability assessment and forecasting of renewables: A review for solar, wind, wave and tidal resources, Renew. Sustain. Energy Rev. 44 (2015) 356–375, http://dx.doi.org/10.1016/j.rser.2014.12.019.
[11] G. Liu, J. Zhou, B. Jia, F. He, Y. Yang, N. Sun, Advance short-term wind energy quality assessment based on instantaneous standard deviation and variogram of wind speed by a hybrid method, Appl. Energy 238 (2019) 643–667, http://dx.doi.org/10.1016/j.apenergy.2019.01.105.
[12] R.M. Shukla, S. Sengupta, A.N. Patra, Smart plug-in electric vehicle charging to reduce electric load variation at a parking place, in: 2018 IEEE 8th Annu Comput Commun Work Conf CCWC 2018, 2018-Janua, 2018, pp. 632–638, http://dx.doi.org/10.1109/CCWC.2018.8301710.
[13] H. Seljeseth, H. Taxt, T. Solvang, Measurements of network impact from electric vehicles during slow and fast charging, IET Conf. Publ. 2013 (2013) 10–13, http://dx.doi.org/10.1049/cp.2013.1197.
[14] J. Nömm, S.K. Rönnberg, M.H.J. Bollen, An analysis of voltage quality in a nanogrid during islanded operation, Energies 12 (2019) http://dx.doi.org/10.3390/en12040614.
[15] D. Macii, D. Petri, Rapid voltage change detection: Limits of the IEC standard approach and possible solutions, IEEE Trans. Instrum. Meas. 69 (2020) 382–392, http://dx.doi.org/10.1109/TIM.2019.2903617.
[16] M. Bollen, A.G. de Castro, S. Rönnberg, Characterization methods and typical levels of variations in rms voltage at the time scale between 1 second and 10 minutes, Electr. Power Syst. Res. 184 (2020) 106322, http://dx.doi.org/10.1016/j.epsr.2020.106322.
[17] A. Gil-de Castro, M.H.J. Bollen, S.K. Rönnberg, Variations in harmonic voltage at the sub-10-minute time scale, Electr. Power Syst. Res. 195 (2021) 107163, http://dx.doi.org/10.1016/j.epsr.2021.107163.
[18] S.M. Miraftabzadeh, M. Longo, F. Foiadelli, M. Pasetti, R. Igual, Advances in the application of machine learning techniques for power system analytics: A survey †, Energies 14 (2021) http://dx.doi.org/10.3390/en14164776.
[19] S.M. Miraftabzadeh, F. Foiadelli, M. Longo, M. Pasetti, A survey of machine learning applications for power system analytics, in: Proc - 2019 IEEE Int Conf Environ Electr Eng 2019 IEEE Ind Commer Power Syst Eur EEEIC/I CPS Eur 2019, 2019, http://dx.doi.org/10.1109/EEEIC.2019.8783340.
[20] E. Styvaktakis, M.H.J. Bollen, IYH. Gu, Expert system for classification and analysis of power system events, IEEE Trans. Power Deliv. 17 (2002) 423–428, http://dx.doi.org/10.1109/61.997911.
[21] P.K. Dash, S. Mishra, M.A. Salama, A.C. Liew, Classification of power system disturbances using a fuzzy expert system and a Fourier linear combiner, IEEE Trans. Power Deliv. 15 (2000) 472–477, http://dx.doi.org/10.1109/61.852971.
[22] M.B.I. Reaz, F. Choong, M.S. Sulaiman, F. Mohd-Yasin, M. Kamada, Expert system for power quality disturbance classifier, IEEE Trans. Power Deliv. 22 (2007) 1979–1988.
[23] P.G.V. Axelberg, I.Y. Gu, M.H.J. Bollen, Support vector machine for classification of voltage disturbances, IEEE Trans. Power Deliv. 22 (2007) 1297–1303, http://dx.doi.org/10.1109/TPWRD.2007.900065.
[24] Y. Mohammadi, M.H. Moradi, R. Chouhy Leborgne, A novel method for voltage-sag source location using a robust machine learning approach, Electr. Power Syst. Res. 145 (2017) 122–136, http://dx.doi.org/10.1016/j.epsr.2016.12.028.
[25] D. De Yong, S. Bhowmik, F. Magnago, An effective power quality classifier using wavelet transform and support vector machines, Expert Syst. Appl. 42 (2015) 6075–6081, http://dx.doi.org/10.1016/j.eswa.2015.04.002.
[26] Y. Mohammadi, A. Salarpour, R.. Chouhy Leborgne, Comprehensive strategy for classification of voltage sags source location using optimal feature selection applied to support vector machine and ensemble techniques, Int. J. Electr. Power Energy Syst. 124 (2021) 106363, http://dx.doi.org/10.1016/j.ijepes.2020.106363.
[27] M. Valtierra-Rodriguez, R. de J. Romero-Troncoso, R.A. Osornio-Rios, A. Garcia-Perez, Detection and classification of single and combined power quality disturbances using neural networks, IEEE Trans. Ind. Electron. 61 (2014) 2473–2482, http://dx.doi.org/10.1109/TIE.2013.2272276.
[28] S. Mishra, C.N. Bhende, B.K. Panigrahi, Detection and classification of power quality disturbances using S-transform and probabilistic neural network, IEEE Trans. Power Deliv. 23 (2008) 280–287, http://dx.doi.org/10.1109/TPWRD.2007.911125.
[29] K. Cai, W. Cao, L. Aarniovuori, H. Pang, Y. Lin, G. Li, Classification of power quality disturbances using wigner-ville distribution and deep convolutional neural networks, IEEE Access 7 (2019) 119099–119109, http://dx.doi.org/10.1109/ACCESS.2019.2937193.
[30] W. Qiu, Q. Tang, J. Liu, W. Yao, An automatic identification framework for complex power quality disturbances based on multifusion convolutional neural network, IEEE Trans. Ind. Inform. 16 (2020) 3233–3241, http://dx.doi.org/10.1109/TII.2019.2920689.
[31] T. Räsänen, M. Kolehmainen, Feature-Based Clustering for Electricity Use Time Series Data, Vol. 5495, 2009, http://dx.doi.org/10.1007/978-3-642-04921-7_41.
[32] B. Fulcher, Feature-based time-series analysis, 2017.
[33] G. Chaoyu, Z. Su, P. Wang, Y. You, Distributed evidential clustering toward time series with big data issue, Expert Syst. Appl. (2021) http://dx.doi.org/10.1016/j.eswa.2021.116279.
[34] X. Wang, A. Wirth, L. Wang, Structure-based statistical features and multivariate time series clustering, in: Seventh IEEE Int. Conf. Data Min., ICDM 2007, 2007, pp. 351–360, http://dx.doi.org/10.1109/ICDM.2007.103.
[35] H. Li, Multivariate time series clustering based on common principal component analysis, Neurocomputing 349 (2019) 239–247, http://dx.doi.org/10.1016/j.neucom.2019.03.060.
[36] J. Li, H. Izakian, W. Pedrycz, I. Jamal, Clustering-based anomaly detection in multivariate time series data, Appl. Soft Comput. 100 (2021) 106919, http://dx.doi.org/10.1016/j.asoc.2020.106919.
[37] L. Wen, K. Zhou, S. Yang, A shape-based clustering method for pattern recognition of residential electricity consumption, J. Clean. Prod. 212 (2019) 475–488, http://dx.doi.org/10.1016/j.jclepro.2018.12.067.
[38] H. Izakian, W. Pedrycz, I. Jamal, Fuzzy clustering of time series data using dynamic time warping distance, Eng. Appl. Artif. Intell. 39 (2015) 235–244, http://dx.doi.org/10.1016/j.engappai.2014.12.015.
[39] L.G.B. Ruiz, M.C. Pegalajar, R. Arcucci, M. Molina-Solana, A time-series clustering methodology for knowledge extraction in energy consumption data, Expert Syst. Appl. 160 (2020) 113731, http://dx.doi.org/10.1016/j.eswa.2020.113731.
[40] S. Galvani, S. Rezaeian Marjani, J. Morsali, M. Ahmadi Jirdehi, A new approach for probabilistic harmonic load flow in distribution systems based on data clustering, Electr. Power Syst. Res. 176 (2019) 105977, http://dx.doi.org/10.1016/j.epsr.2019.105977.
[41] M. Jasiński, T. Sikorski, K. Borkowski, Clustering as a tool to support the assessment of power quality in electrical power networks with distributed generation in the mining industry, Electr. Power Syst. Res. 166 (2019) 52–60, http://dx.doi.org/10.1016/j.epsr.2018.09.020.
[42] C. Ge, R.A. de Oliveira, I.Y.-H. Gu, M.H.J. Bollen, Deep feature clustering for seeking patterns in daily harmonic variations, IEEE Trans. Instrum. Meas. 70 (2021) 1–10, http://dx.doi.org/10.1109/TIM.2020.3016408.
[43] C. Ge, R.A.D. Oliveira, I.Y.H. Gu, M.H.J. Bollen, Unsupervised deep learning and analysis of harmonic variation patterns using big data from multiple locations, Electr. Power Syst. Res. 194 (2021) 107042, http://dx.doi.org/10.1016/j.epsr.2021.107042.
[44] R.A. De Oliveira, V. Ravindran, S.K. Ronnberg, M.H.J. Bollen, Deep learning method with manual post-processing for identification of spectral patterns of waveform distortion in PV installations, IEEE Trans. Smart Grid 12 (2021) 5444–5456, http://dx.doi.org/10.1109/TSG.2021.3107908.
[45] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab Syst. 2 (1987) 37–52, http://dx.doi.org/10.1016/0169-7439(87)80084-9.
[46] K. Van Deun, L. Thorrez, M. Coccia, D. Hasdemir, J.A. Westerhuis, A.K. Smilde, et al., Weighted sparse principal component analysis, Chemom. Intell. Lab Syst. 195 (2019) 103875, http://dx.doi.org/10.1016/j.chemolab.2019.103875.
[47] M. Feng, Project 1 Report : Dimensionality Reduction, pp. 1–11.
[48] M. Sakthi, T.A. Selvadoss, An effective determination of initial centroids in K-means clustering using kernel PCA, Int. J.Comput. Sci. Inf. Tech. 2 (2011) 955–959.
[49] D. Arthur, S. Vassilvitskii, K-means++: The advantages of careful seeding, in: Proc Annu ACM-SIAM Symp Discret Algorithms, 2007, pp. 1027–1035, 07-09-Janu.
[50] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[51] D.M. SAPUTRA, D. SAPUTRA, L.D. OSWARI, Effect of distance metrics in determining K-value in K-means clustering using elbow and silhouette method, 172, 2020, pp. 341–346, http://dx.doi.org/10.2991/aisr.k.200424.051.
[52] Y. Mohammadi, R.C. Leborgne, A new approach for voltage sag source relative location in active distribution systems with the presence of inverter-based distributed generations, Electr. Power Syst. Res. 182 (2020) http://dx.doi.org/10.1016/j.epsr.2020.106222.
[53] Y. Mohammadi, M.H. Moradi, R. Chouhy Leborgne, Employing instantaneous positive sequence symmetrical components for voltage sag source relative location, Electr. Power Syst. Res. 151 (2017) 186–196, http://dx.doi.org/10.1016/j.epsr.2017.05.030.
[54] Y. Mohammadi, R.C. Leborgne, Improved DR and CBM methods for finding relative location of voltage sag source at the PCC of distributed energy resources, Int. J. Electr. Power Energy Syst. 117 (2020) 105664, http://dx.doi.org/10.1016/j.ijepes.2019.105664.
[55] Y. Mohammadi, R.C. Leborgne, B. Polajžer, Modified methods for voltage-sag source detection using transient periods, Elect. Power Syst. Res. 207 (2022) http://dx.doi.org/10.1016/j.epsr.2022.107857.
[56] M.H. Moradi, Y. Mohammadi, M. Hoseyni Tayyebi, A novel method to locate the voltage sag source: A case study in the Brazilian power network (Mato Grosso), Prz Elektrotechniczny 88 (2012).
[57] C.B. Lucasius, A.D. Dane, G. Kateman, On k-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasiblity and comparison, Anal. Chim. Acta 282 (1993) 647–669, http://dx.doi.org/10.1016/0003-2670(93)80130-D.

Update

Sustainable Energy, Grids and Networks

Volume 32, Issue , December 2022, Page

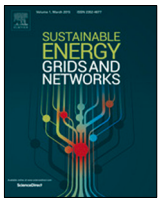
DOI: https://doi.org/10.1016/j.segan.2022.100918

Sustainable Energy, Grids and Networks 32 (2022) 100918

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

journal homepage: www.elsevier.com/locate/segan



Corrigendum to "An unsupervised learning schema for seeking patterns in rms voltage variations at the sub-10-minute time scale" [Sustain. Energy Grids Netw. 31 (2022) 1–12/100773]

Younes Mohammadi [a,*], Seyed Mahdi Miraftabzadeh [b], Math H.J. Bollen [a], Michela Longo [b]

[a] Department of Engineering Sciences and Mathematics, Luleå University of Technology, Skellefteå Campus, Forskargatan 1, 93187 Skellefteå, Sweden
[b] Department of Energy, Politecnico di Milano, Via Lambruschini 4, 20156 Milano, Italy

ARTICLE INFO

Article history:
Available online 2 September 2022

The authors regret <to inform you that the current status of 1st (Younes Mohammadi) and 2nd (Seyed Mahdi Miraftabzadeh) author as "Postdoctoral researcher" has been wrongly added to their names during the proofreading stage; Hence the updated list of authors are as below:

Younes Mohammadi,[a,*], Seyed Mahdi Miraftabzadeh,[b], Math H.J. Bollen, [a], Michela Longo[b]
[a] *Department of Engineering Sciences and Mathematics, Luleå University of Technology, Skellefteå campus, Forskargatan 1, 93187 Skellefteå, Sweden*
[b] *Department of Energy, Politecnico di Milano, Via Lambruschini 4, 20156 Milano, Italy*

Moreover, the following corrections, are made to the main version of the paper to make the text clearer and simpler for the readers and show the originality of the text purer. The corrections modify some sentences, cite some missing references from the already existing references, and do not affect the main parts of the paper; idea, contribution, and results. Abbreviations used are as: Page (P), Column (C), Section (S), and Line (L). References refer to the initial version of the paper.

Location in the PDF of initial version	Corrections
P1, C1, S1, L1-5, …"Voltage magnitude…network operators."…:	Is replaced by: "Deviation of rms voltage magnitude differs over different time scales. Collecting measurement data of power quality is significant for network operators (TSOs/DSOs) in both power-system performance and power-quality studies"
P1, C1, S1, L8&9, …". The overview…value used"…:	Is replaced by: ", and used as most common shown in [4]"
P1, C2, S1, L1-4, "Moreover, tripping…time scale [4-7]"…:	Is replaced by: "Moreover, several reported undesirable outcomes of fast voltage variations, further tripping PVs and light flicker, come from variations in this time scale [4-7]".
P1, C2, S1, L0&11, …"individual rapid voltage changes (voltage steps) as"…:	Is replaced by "voltage steps, as only one of the sub-10-min variations' components)"
P1, C2, S1, L13&14, …"However, voltage steps…variations"…:	Is deleted.

DOI of original article: https://doi.org/10.1016/j.segan.2022.100773.
* Corresponding author.
E-mail address: Younes.mohammadi@ltu.se (Y. Mohammadi).

https://doi.org/10.1016/j.segan.2022.100918

Location in the PDF of initial version	Corrections
P2, C1, S1, L1&2, …"but do not result…time window. A"…:	Is replaced by ", while a"
P2, C1, S1, L5-7, …"Power-quality…long period"…and L11-14, …"Automatic analysis methods…consuming human intervention"…:	Is deleted.
P2, C1, S1, L14-15, "Recent…patterns."	Is replaced as "Employing machine learning methods can automatically extract such patterns."
P2, C1, S1, L36-42, …"However,…generation distribution."…	Should be read as "However, few research have applied the unsupervised methods for power-quality data clustering. Examples are as clustering energy consumption [39], data clustering for harmonic load flow [40], and finding the distributed generators contribution using k-means [41]."
P2, C1, S1, L45-47, …"They are…new general knowledge."…	Is replaced by "[42,43] concern the daily harmonic voltage variations as a well-understood phenomenon.". As we believe that the ref. [44] has resulted in general knowledge of inter-harmonics as a less understood phenomenon, we have separated [44] from [42] and [43].
P2, C2, S1, L8-15, …"The upper…a few seconds."…:	Is replaced by "10 min is defined in IEC61000-4-30 as a commonly used value and 1 s is according to the available time resolution of the used monitor [16,17]."
P2, C2, S1, L35, …"[42-44]"…:	Is replaced by "[42,43]".
P2, Caption of Table 1:	Is replaced by "A summary of single-window existing statistics used in this research based on [16,17]"
P3, C1, S3, L8, after …"(e) Visualising the":	"original," is added. So, part (e) is now as "(e) Visualising the original size reduced…
P3, C1, S3(a), Paragraph 1:	Is replaced by "In this stage, the 10-min windows including, $n = 600$ dimensions, are prepared as each row x_j of input matrix $X^{m\times n}(1)$."
P3, C1, S3(b), Last paragraph:	Is replaced by "Step (b) may also support k-means (step (c)) for more effective initialization of centroids [48]. KPCA was optimised by the function *KernelPCA* with "*cosine*" in *Phyton*."
P4, C1, S3 (c), L5-8, … "Each feature … are assigned"…:	Is deleted.
P4, C1, S3 (c), Last paragraph:	Is replaced by "The k-means, an unsupervised clustering method, is widely used in the literature [39,42,43]."
P4, C1, S3 (d):	Is replaced as: "The feature vectors from centroids are inputted to the inverse KPCA function to reconstruct the illustrative cluster centers $x_j^{rec} = f_{inv}(\mu_j)$ (x_j has 600D and μ_j has pD)." Since transformation back to the original sub-space by any reconstruction method associated with a reconstruction error of modeling, this study ignored this step as follows: by having the labels within clusters in the output of k-means, an average was taken on the samples (with the same number) in the input space so that no reconstruction error was involved on the K centroids (and all other samples)."
P4, C1, S3 (e):	Is replaced by: *(e) Visualization of original, size-reduced principal and clustered feature vectors* For visualization of original x_i^{mD}, size-reduced principal h_i^{PD}, and clustered h_i, a 2D t-SNE tool [50] is used. Like the ref. [42], using this method is just for 2D visualization of the mentioned feature vectors to see how well the KPCA and k-means contributed to the proposed schema.

Location in the PDF of initial version	Corrections
P4, C2, S4, L7, after …"for the measurements":	"The used dataset is part of the dataset introduced in ref. [16]" is added.
P6, C2, S5.3, L3, …"(2) The necessity of obtained patterns beside statistical indices over a sub- 10-min time scale."	Is replaced by "(2) The importance of extracted patterns in addition to the statistical indices introduced in [16,17] for a sub-10-min time scale."
P7, C1, S5.3.2, Paragraph 2:	Is replaced by "The most dominant clusters, 5 and 3, show softer curves. The samples are well-divided into 10 clusters so that there is a clear separation between each of the clusters over each of the four indices. Some indices show similar results between a pair cluster:"
P7, C2, S5.3.3, Paragraph 1:	Is replaced by "Table 2 gives the result of chosen indices on the ten reconstructed patterns. The almost similar results between each two patterns are as follows:"
P8, C2, S6.1, Paragraph 2:	Is replaced by "Locating the source of voltage dips can be done by our proposed schema by automatic feature extraction compared with [52–56] that use a pre-defined set of features in a labeled dataset. The problem would be a two-class unsupervised learning (downstream/upstream class). The high-resolution voltages and current rms signals (including pre and after-dip) within a selected window would be inputted to KPCA. Then, a k-means concludes two classes, with which one expert can label each class member."
P10, C2, S6.4, L7, …" slower/faster"…:	Is replaced by: "faster/slower".

>.

The authors would like to apologize for any inconvenience caused to the journal and its Editors.