Electrical Power and Energy Systems 143 (2022) 108516

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

Seeking patterns in rms voltage variations at the sub-10-minute scale from multiple locations via unsupervised learning and patterns' post-processing

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

Keywords:
Power quality
Fast voltage variations
Seeking patterns
Unsupervised learning
Kernel PCA (KPCA)
Time series clustering
Post-processing

ABSTRACT

This paper addresses the issue of seeking sub-10-min patterns in fast rms voltage variations from time-limited measurement data at multiple locations worldwide. This is a rarely considered time scale in studies that could be important for the incorrect operation of end-user equipment. Moreover, measurements from multiple locations could be significant from the view of seeking pattern methods. To learn more about this time scale, we propose an unsupervised learning method that employs a Kernel Principal Component Analysis (KPCA) with a Cosine kernel to extract principal features from 10-min time series of voltage variations with a 1-s resolution followed by a k-means clustering to group the features. The scheme is applied to measurements from 57 low-voltage locations in 19 countries from 2009 to 2018. Fifteen initial clusters/patterns are then extracted and converted to ten new (general) patterns using a clusters' merging strategy with highly similar patterns employed in a new post-processing approach useful for multiple locations. Utilizing data from multiple locations in multiple countries ensures a level of generality of the patterns. It also allows comparing the locations. Next to the ten general patterns, some typical patterns are extracted separately for every location. A statistical indices analysis confirms that a complete picture of sub-10-min oscillations needs both statistical indices (quantifying level and variations) and the proposed framework (quantifying patterns). The extracted patterns could be used as a reference for testing/putting requirements on the grid-connected equipment and quantifying the grid's hosting capacity for different types of new distributed generations connected to the grid. The framework is scalable and computationally cheap, making it appropriate for seeking typical patterns in the big data domain. Applying the framework to the much less understood phenomenon will result in providing general knowledge in the field of power quality.

1. Introduction

The voltage magnitude's (rms value) deviation from its nominal voltage varies over a range of time scales. Standards and regulations on voltage magnitude variations consider two distinctly different time scales: longer than several minutes and up to a few seconds. Slow voltage variations (also known as "supply voltage variations" or "voltage regulation") take place at time scales of minutes and longer. The IEC 61000–4-30 standard on power-quality monitoring prescribes that the rms voltage is calculated over a 10-min window [1,2]. Moreover, the overview of voltage-quality regulation in Europe [3] also shows that 10 min is the most common value. Fast voltage variations (also known as "voltage flicker", "voltage fluctuations", and "continuous rapid voltage changes") take place at time scales up to a few seconds. Voltage-quality indicators used in this time scale are short-term flicker severity (Pst) and long-term flicker severity (Plt), as defined in IEC 61000–4-15 [4] and IEEE 1453 [5].

However, there is a lack of performance indicators for voltage quality and knowledge about voltage magnitude variations with a time scale between a few seconds (in our study, 1 s) and several minutes (in our study, 10 min), referred to as sub-10-min values in [6], and power-quality monitoring programs seldom include it. However, this time scale should not be neglected because equipment may be susceptible to sub-10-min variations. Moreover, tripping of PV installations due to overvoltages is seen for the values belonging to the 10-min time scale. Many reported adverse consequences of fast voltage variations, next to light flicker, are also due to variations in this time scale [7–9]. Besides this, many different new types of equipment (generation or load) [10]

[*] Corresponding author.
E-mail address: Younes.mohammadi@ltu.se (Y. Mohammadi).

https://doi.org/10.1016/j.ijepes.2022.108516
Received 29 March 2022; Received in revised form 28 June 2022; Accepted 20 July 2022
Available online 1 August 2022

are the known sources of voltage variations in this time scale: PV power installations [9,11,12], wind power installations [13,14], EV charging [15] and electric heat pumps [16].

1.1. State of the art

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 [11,17]. However, voltage steps represent only one aspect of the sub-10-min variations. Later, some research has been done in regard to statistical indices and actual levels for sub-10-min values of rms voltage [9,18] and harmonic voltage [6] in a sub-10-min scale. Such statistics are defined, for example, as the 99th percentile of the 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 voltage variations, but they do not result in the 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 range of variations but also the patterns themselves.

Power-quality monitoring can result in large amounts of data, especially where it concerns measurements at multiple locations over a long period. 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 automatically identify such patterns. In general, two sets of methods can be implemented for training supervised and unsupervised learning [19,20]. The initial approaches, as supervised ones, needed a pre-labeled dataset. Artificial intelligent-based methods used expert classifiers like support vector machines [21–23], ensemble learnings [24,57] and neural networks [25,26]. Automatic extraction of input features has been done one step before the supervised classifiers in the literature [27,28]. Seeking patterns from signals, the 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 [27,28], the automatic extraction of principal features has a normally better role than the manually extracted ones (e.g., statistical indices) [1,29,30] to group a dataset.

There are many works previously done on time series clustering, e. g., clustering on the areas of big data in [31], clustering by utilizing various tools than k-means, and the Euclidean distance measurement criterion addressed in [32] as shape-based clustering and in [33] as fuzzy-based one by using Distance Time Wrapping (DTW) as the similarity measure criteria. However, a limited number of applications in power quality data measurement analysis have been found, such as a time series clustering methodology for knowledge extraction in energy consumption data in [34], a clustering method for the probabilistic evaluation of harmonic load flow in [35], and a k-means clustering for identification of distributed generation contribution in [36]. 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 to seek the daily patterns for measurement from one location [37], multiple locations [38], and the use of a post-processing method [39]. The refs. [37–39] are concerned with a rather well-understood phenomenon (daily variation in harmonic voltage), so 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 yet applied to seek patterns for rms/harmonic voltage fast variations in the sub-10-min scale, which is a not-yet-well discovered phenomenon and different from daily variational patterns. Moreover, no framework applicable for time-limited (about one day and a few hours) measurements from multiple locations has been designed.

1.2. Contribution and applicability

Refs. [6,9,18] shows that quantifying voltage magnitude variations in the sub-10-min time scale is not trivial. To learn more about this phenomenon/disturbance type, this paper aims to seek patterns in rms voltage variations at the time scale. Measurements from multiple locations worldwide (with possibly different behaviors) will be used to identify those patterns. Using data (which is short-limited time) from multiple locations in multiple countries ensures a level of the generality of the patterns. It also allows comparing locations. In this way, a complete picture of the time scale is obtained as a part of the long-term aim of power quality studies. Hence, a framework is proposed; it includes (a) unsupervised learning methodology (Kernel Principal Component Analysis (KPCA) with Cosine kernel) and (b) a pattern's post-processing approach necessary for the multiple location measurements (to avoid occurring highly similar patterns with only a difference in voltage magnitude). In this paper, it was decided to go for unsupervised instead of supervised learning to obtain information on the kind of patterns that could be expected in this time scale. Such patterns can be used as a reference when designing equipment connected to the grid to reduce the probability of interference (i.e., equipment not connecting the way they should behave). Next to that, the authors of this paper have a general interest in the kind of patterns that can be expected in this time scale. The typical patterns per location, as found from this study, increase the overall power quality knowledge. A very important finding of this paper is that a complete picture of fast voltage variations in the sub-10-min scale needs both statistical indices and an expert framework to extract the possible patterns.

The main contributions/novelties of this paper are:

(a) Considering fast voltage variations in a time window of 1 s-10 min. This unknown phenomenon is still largely unexplored, and there is very limited information/ knowledge on what it looks like.
(b) Proposing a comprehensive framework applicable for short-limited measurement in multiple locations, to seek possible patterns for the very first time in the "sub-10-min" rms voltage variations (solving this problem is a challenge by the manual analysis of the measurement from several locations).
(c) Obtaining 10 patterns (Fig. 9) generally for multiple locations (because of time limitations in measurements, expertly, all locations are considered together) and some new typical patterns (Fig. 18) per single location (taking an average of the samples with specific cluster number per location).
(d) The aim of the paper is pursuing a novel application of the well-known existing unsupervised methods (KPCA with different kernels followed by k-means) and adding a post-processing term to bring new knowledge to this field of study that has not yet been explored (a supplement for part (a)).
(e) The necessary post-processing approach for multiple locations and the approach to extract patterns for every separate (single) location; simple mathematic relations are used, and the framework is not made complex.
(f) A statistical analysis using power-quality indices is done on the obtained clusters and patterns to validate them. A power-quality look is also applied to the obtained different patterns. The results (Fig. 19) show the following real variations by the pattern-based variations for the locations.

The applications of the proposed framework are:

(a) Our proposed framework is applied to a much less understood phenomenon/disturbance type, which refers variations in rms voltage in a 10 min window with 1 s resolution, so the work will result in providing general knowledge beyond the specific case study. Moreover, each 10-min window may follow a pattern

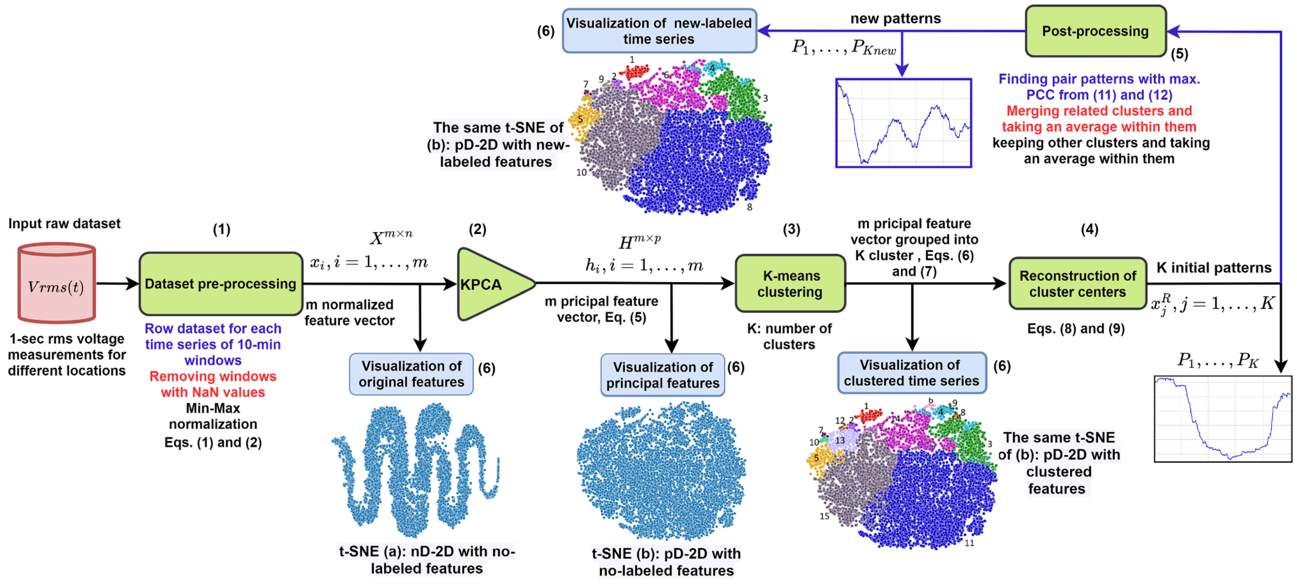

(6)
Visualization of new-labeled time series
new patterns
$P_1, \ldots, P_{Knew}$
Post-processing
(5)
Finding pair patterns with max. PCC from (11) and (12)
Merging related clusters and taking an average within them
keeping other clusters and taking an average within them
The same t-SNE of (b): pD-2D with new-labeled features
Input raw dataset
$Vrms(t)$
1-sec rms voltage measurements for different locations
(1)
Dataset pre-processing
Row dataset for each time series of 10-min windows
Removing windows with NaN values
Min-Max normalization
Eqs. (1) and (2)
$X^{m \times n}$
$x_i, i = 1, \ldots, m$
m normalized feature vector
(2)
KPCA
$H^{m \times p}$
$h_i, i = 1, \ldots, m$
m pricipal feature vector, Eq. (5)
(3)
K-means clustering
K: number of clusters
m pricipal feature vector grouped into K cluster , Eqs. (6) and (7)
(4)
Reconstruction of cluster centers
Eqs. (8) and (9)
K initial patterns
$x_j^R, j = 1, \ldots, K$
$P_1, \ldots, P_K$
Visualization of original features
(6)
t-SNE (a): nD-2D with no-labeled features
Visualization of principal features
(6)
t-SNE (b): pD-2D with no-labeled features
Visualization of clustered time series
(6)
The same t-SNE of (b): pD-2D with clustered features

Fig. 1. Process of the proposed framework for seeking patterns in sub-10-min rms voltage variations from measurements at multiple locations.

among ten obtained patterns. Meanwhile refs. [37–39] have studied the rather well-understood phenomena, i.e., daily variation in harmonic voltage (a 24-h window with 10 min resolutions). Furthermore, each 24-h window follows a pattern between only two concluded straightforward patterns. In our proposed framework, solving the problem of extracting patterns from the fast voltage variations (including lots of variations) is much harder than seeking patterns from slow voltage variations in [37–39].

(b) The manufacturers of grid-connected equipment could use the extracted patterns as a reference for every single location to see the obtained patterns and design equipment, including testing/protective/control requirements in relation to the location connected to special equipment like PVs, EVs, electric pumps and wind power installations.

(c) The results obtained from this study regarding the 10 patterns can develop future standards/classification methods by labeling patterns through the sources causing the patterns (if the information about the connected loads exists).

(d) The obtained patterns can also be used to quantify the grid's hosting capacity for different types of new equipment connected to the grid, such as PVs. A study in [9] showed that PVs mainly impacted the patterns of variations. Using the high-resolution patterns can help to calculate the capacity of PVs in the grid more precise, as compared to the values recorded only for one hour or more.

(e) By keeping track of the obtained patterns, as updated variations in a day, month and even long-term periods, trends can be identified easily for the network operators.

(f) Even if there are yet not many impacts of the time scale on equipment, it is not a good idea to wait for such impacts to occur. Hence, this work, by quantifying patterns next to the existing quantifying variations, will operate as a preventive work.

The method is applied to 10-min time series with a 1-s time resolution obtained from 57 different locations in 19 countries. The measurements, which were from 2009 to 2018, were performed at a wall outlet, 220 V or 230 V, 50-Hz low voltage networks. A statistical power-quality analysis of the obtained results shows that the proposed framework is effective in pattern extraction and confirms that a full representation of voltage variations at the sub-10-min scale needs both results of the statistical indices and extracted patterns.

1.3. *Paper organization*

Section 2 of this paper describes the proposed framework in six subsections. Section 3 presents the measurement dataset and shows some examples of variations on the sub-10-min time scale. The proposed framework results, correlation analysis of obtained final patterns, and a statistical analysis on the obtained clusters' samples and patterns are given in Section 4. Section 5 discusses the application of the proposed methodology using multiple location measurements for each location, the importance of the obtained patterns and the reason why the proposed framework has not been run separately for each location. Section 6 discusses the paper and suggests future works; finally, Section 7 concludes the paper.

2. The proposed framework

This section proposes a framework to seek patterns in 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; it is commonly used in power quality monitoring. The lower limit of the window (1 s) is not part of any standard; it is not commonly used either. The 1-s period is partly set by the available measurement data; also, it is partly set by the computation effort needed and by the fact that standards and regulations exist for time scales up to a few seconds. The process of the proposed framework consists of six modules as shown in Fig. 1: (a) pre-processing measurement dataset, (b) applying KPCA with different kernels on the feature vectors (normalized high-dimensional) x_i, which results in the vectors h_i with principal features, (c) using k-means clustering to group the principal features h_i, (d) reconstructing cluster centers using an inverse KPCA, (e) applying a new post-processing approach to the reconstructed cluster centers, and (f) visualizing the original features, size reduced principal features, clustered features and new-labeled clustered features in the 2D space using t-SNE.

2.1. *Pre-processing the measurement dataset*

The first part of the proposed framework 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 resolutions including n = 600 dimensions (600 × 1 s = 600 s/10 min).

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

m is the number of samples (time series as 10-min windows). Second, the windows including missing data (NaN values), are removed. Later on, from several possible approaches, a Min-Max normalization (2) is applied to matrix X,including × elements. In this way, each of the 600 features is considered an independent coordinate which means that those samples with very high (low) 1 s rms voltages will have values close to 1 (0). Each time series is scaled within [0,1] at the end of this operation and will have an equal contribution to the matrix X.

$$x_{[0,1]} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

where $x_{[0,1]}$ is the normalized value of each x element of X, and max (x) and min(x) are the minimum and maximum of each column of X, respectively.

2.2. Feature extraction using KPCA

Principal Component Analysis (PCA) is one of the most powerful algorithms in data pre-processing for dimensionality reduction in many applications. PCA is a linear algorithm that transforms the original data into a linear combination of the new uncorrelated features. The new features aid in bringing non-obvious patterns in the data to the forefront and can improve the performance of the ML methods [40,41]. However, in our work, a KPCA algorithm was chosen and employed to extract principal features after using and testing PCA on the dataset. KPCA is a nonlinear PCA that uses kernel methods to deal with the nonlinearity of the input data. KPCA captures more complex data patterns, which would not be present under typical linear PCA transformations. The main idea underlying KPCA is similar to the support vector machine, which (in our case) takes high-dimensional data sequences (x_i^{600D}) from $X^{m\times 600}$ and maps the data space $x_i (i = 1, 2, ..., m)$ to a higher dimension space $\Phi(x_i)$ in which the data is linearly separatable. KPCA then makes a non-linear function using the kernel matrix $\boldsymbol{K}$ (3). The kernel methods in KPCA compute the distance between each sample, which makes this method computationally expensive when compared to PCA [42]. The detailed mathematical analysis, proof and comparison of KPCA with different kernels can be found in [43–45], which is beyond the scope of this paper.

After checking several kernels as Linear, Polynomial, RBF, Sigmoid and Cosine (4), a Cosine kernel is chosen (d is the polynomial degree, $\gamma = 1/2\delta^2, and\theta \geq 0$). The reason for this selection has been based on amount of help the kernel provides to K-means clustering to find patterns with a wider/clearer range in voltage magnitude variations (i.e., patterns as a representative of the clusters will include the oscillations clearer). In the next step, PCA in this high-dimensional feature space is calculated to reduce the dimension linearly. Hence, KPCA, like PCA, does an eigen analysis and projects the feature vectors on the first p (in our case 10) 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).

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

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

In addition to the size feature reduction, this step may help to better initialize centroids for k-means clustering [46].

2.3. Clustering using k-means

Principal feature vectors (h_i^{pD}) concluded from KPCA are inputted to the k-means clustering block. The *k-means*++ initialization scheme [47] finds out K initial centroids (cluster centers)μ_j in an effective way. K-means clustering aims to group the vectors h_i into K clusters (in our case, initial $K = 15$). 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:

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

$$\omega_{ij} = \begin{cases} 1, ifj = \underset{j}{argmin} \|h_i - \mu_j\|^2 \\ 0, otherwise \end{cases}, \mu_j = \frac{\sum_{i=1}^{m} \omega_{ij} h_i}{\sum_{i=1}^{m} \omega_{ij}}, j = 1, 2, ..., K \tag{7}$$

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.

2.4. 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 (8). x_j^{rec} is the reconstructed data sequence (600 dimensions) for the feature vector μ_j (the jth cluster center with the dimension p). As normalized, these reconstructed data sequences represent data patterns for the individual clusters.

$$x_j^{rec} = f_{inv}(\mu_j), j = 1, 2, ..., K \tag{8}$$

To de-normalize (8) as x_j^R, each element of x_j^{rec} is de-normalized using (9) as follows:

$$x_j^R = x_j^{rec}(\max(x) - \min(x)) + \min(x) \tag{9}$$

Finally, $P_1, ..., P_K$ as the sub-10-min patterns are concluded.

2.5. A new post-processing approach to the reconstructed cluster centers (patterns)

The dataset used in this paper is related to a number of different locations worldwide with possibly different behaviors in voltage magnitude variations (depending on the type of the connected equipment). The variations deviate over the nominal voltages, 220 V and 230 V. However, the obtained patterns from the proposed framework at multiple locations may have similar variation shapes despite having different ranges of voltage magnitude. This is because k-means inherently have checked the Euclidian distance (6) between each low

$$[\Phi(x_i), \Phi(x_j)] = \underbrace{\Phi(x_i)\Phi(x_j)^T}_{\text{Linear}} | \underbrace{(1 + \Phi(x_i)\Phi(x_j)^T)^d}_{\text{Polynomial}} | \underbrace{exp\left(-\gamma\Phi(x_i) - \Phi(x_j)^2\right)}_{\text{RBF}} | \underbrace{tanh(\Phi(x_i)\Phi(x_j)^T + \theta)}_{\text{Sigmoid}} | \underbrace{\frac{\Phi(x_i)\Phi(x_j)^T}{\Phi(x_i)\Phi(x_j)}}_{\text{Cosine}} \tag{4}$$

Table 1
Measurement dataset from multiple locations per country and type of customer.

Country	No. of measurement locations			No. of 1-s rms voltage values	Measurement hours	No. of 10-min windows
	Other*	Hotel	Total			
Sweden	16	6	22	3,348,600	930.2	5581
China		6	6	601,800	167.2	1003
Bosnia and Herzegovina		1	1	100,200	27.8	167
Austria		3	3	301,200	83.7	502
Italy		2	2	200,400	55.7	334
Turkey		2	2	201,000	55.8	335
Hong Kong		2	2	200,400	55.7	334
India	1	2	3	288,000	80.0	480
Spain		1	1	100,200	27.8	167
Switzerland		2	2	200,400	55.7	334
Romania		1	1	100,200	27.8	167
Netherland		3	3	295,800	82.2	493
Singapore		1	1	100,200	27.8	167
Portugal		2	2	200,400	55.7	334
Scotland		1	1	100,200	27.8	167
United Kingdom		1	1	100,200	27.8	167
Ireland		2	2	200,400	55.7	334
Slovenia	1		1	100,200	27.8	167
Zambia	1		1	100,200	27.8	167
Total (19 countries)	19	38	57	6,840,000	1900	11,400

* Apartment, restaurant, and office, detached homes.

dimensional normalized sample (10D) and one centroid. This result may not be an issue for a single location with deviations over its defined nominal voltage. This is since the patterns with only different ranges of voltage magnitude can be considered distinct patterns.

However, for the used case study, regarding multiple locations, manual post-processing with a different distance measurement criterion on the K reconstructed high dimensional patterns (600D) $P_1, ..., P_K$, which could be representative of their own samples in each K cluster, is necessary to avoid the problem of having patterns with similar variation shapes. In this way, the Pearson Correlation Coefficients (PCCs) between the K initial patterns, which is an adjusted Cosine similarity between centered patterns, are first calculated (10). Then, the PCCs between patterns (P_j) and the (K-j) other ones (P_i) ($i > j$) are compared, and a maximum value (if any, with a $Thr. > +0.9$)[1] is chosen, and the related pair patterns (P_i, P_j) are extracted (11). Eq. (12) is then applied to the obtained pair patterns (P_i, P_j) from (11) to ensure that one pattern is selected only once with the others.

where $\bar{P}_i(\bar{P}_j)$ refers to the mean values of patterns $P_i(P_j)$. The next step is merging the pair clusters related to the pair patterns (with maximum PCC showing similar shapes of variations) obtained from (12). The rest of the clusters and the related patterns, with the cross PCCs between -1 and $+0.9$, are kept. The maximum used in (11) and (12) is based on the fact that the number of K initial clusters should not be reduced much, so that the basis of initial k-means clustering is kept.

The time series within the created new clusters (with new label numbers) are then averaged, and the new patterns $P_1, ..., P_{Knew}$ are obtained (in our case, $Knew = 10$). These new patterns are concluded from both Euclidian distance and centered Cosine similarity. Thus, they are highly different and separated in terms of both parameters, the range of voltage values and their variation shape.

$$\text{PCC}(P_i, P_j) = \frac{(P_i - \bar{P}_i).(P_j - \bar{P}_j)}{\|P_i - \bar{P}_i\|^2 \|P_j - \bar{P}_j\|^2}, P_i, P_j \in x_j^R, i, j = 1, 2, ..., K(i > j) \tag{10}$$

$$(P_i, P_j) = \arg\{ \max_{\substack{i>j \\ 1 \le i \le K \\ 1 \le j \le K}} \text{PCC}(P_i, P_j) > +0.9 | j = const.\} \tag{11}$$

$$(P_i, P_j) = \begin{cases} arg\{max\text{PCC}(P_i, P_j)\} & i = i^{'} \\ (P_i, P_j) & else \end{cases} \tag{12}$$

2.6. *Visualization of features by t-SNE*

To visualize the original feature vectors (x_i^{nD}) (no label) from $X^{m\times n}$, principal feature vectors (h_i^{pD}) (no label) from $H^{m\times p}$, the clustered principal vectors (clustered h_i) from the output of k-means, and the new-labeled features, a t-SNE method [48], is used. In a t-SNE, first, the similarity between two feature vectors, i and j, are modeled by p_{ij} and q_{ij} in the input (pD) and output (in our case, 2D) of t-SNE, respectively. The mapping is then obtained by minimizing the KL divergence between those two distributions:

$$\text{KL}(\text{P}||\text{Q}) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{13}$$

In our case, t-SNE is only for 2D visualization of feature vectors to see how the proposed methodology extracts initial patterns (K) and secondary patterns ($Knew$).

The best results from the unsupervised part of the proposed framework are obtained with the criteria of the good separation of clusters and identification of patterns with a wider/clearer range in voltage magnitude variations (i.e., patterns as a representative of the clusters will include the oscillations more clearly). Finally, the new post-processing approach will find the most district patterns.

3. Measurement dataset

Time series of the 1-s rms voltage were obtained from recorded measurements at 57 locations in 19 countries worldwide, as given in Table 1. The measurements were non-continuous from 2009 to 2018; they were performed at a wall outlet, 220 V or 230 V, 50-Hz low voltage networks. The Metrum PQsmart portable and Dranetz PX5 monitors were used for the measurements. All measurements were following IEC 61000–4-30 Class A. As can be seen in Table 1, most measurements were

[1] After checking the pair patterns and the related PCCs, this number was selected as a criterion showing highly similar patterns.

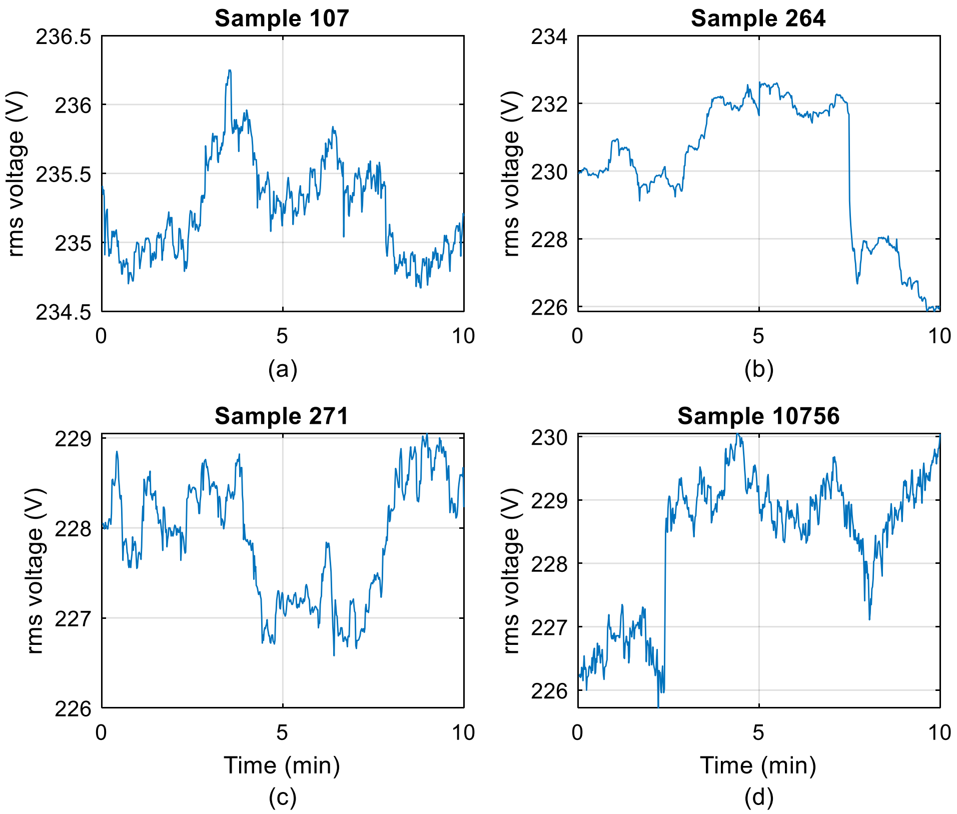


Fig. 2. Four examples of variations in rms voltage within a 10-minute window. (a) Location 1; (b) and (c) Location 2; (d) Location 54.

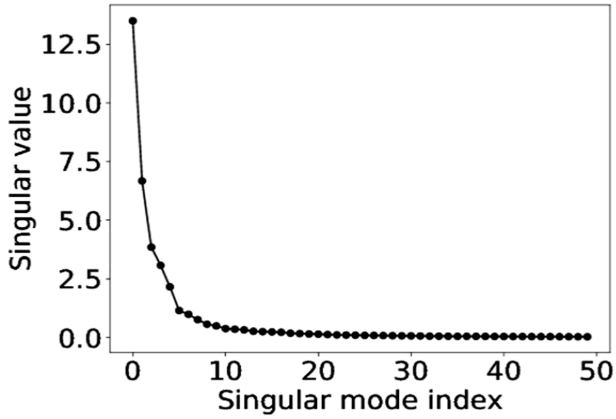


Fig. 3. Singular values corresponding to each singular mode index for KPCA with Cosine kernel.

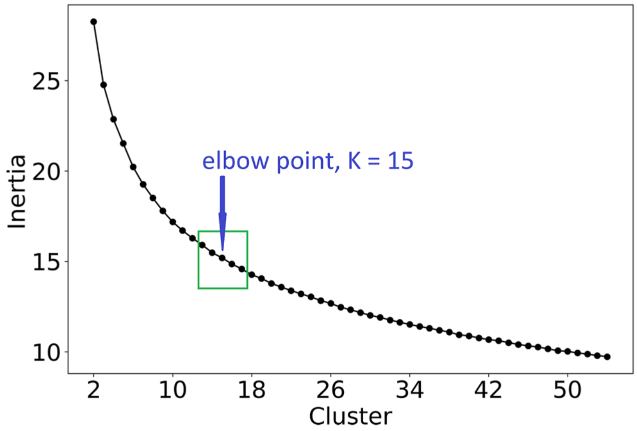


Fig. 4. The inertia (6) of k-means vs. a different number of clusters.

performed in hotel rooms, which was the easiest place to connect the monitors. 53 out of 57 locations had time-limited measurement intervals (about 28-h, 168 10-min windows). In total, 6,840,000 1-s rms voltage values were recorded during 1900 h, corresponding to a total dataset consisting of 11,400 10-min windows. By pre-processing the dataset, $m = 11237$ input feature vectors x_i (10-min sequences, containing 600 1-s samples each) could be obtained, concluding a time series matrix $X^{11237\times600}$.

As four different examples of the variations in the rms voltage within a 10-min window from the input dataset, Samples 107 (recorded in an apartment in Skellefteå, Sweden), 264 and 271 (a hotel in Shanghai, China), and 10,756 (an apartment in Ljubljana, Slovenia) are shown in Fig. 2. The variations can belong to a similar pattern or some different patterns. Therefore, it is worth discovering the underlying patterns from the dataset, including time-limited measurements, where one may find good interpretations of each location's physical reality. Moreover, this can obtain a comprehensive picture of voltage magnitude variations (oscillations) at the time scales below ten minutes.

4. Results and analysis

4.1. Results of the proposed methodology

The results of a combination of KPCA with different kernels (4) and k-means have been investigated (sub-Section 6.3.2). None of the used kernels as Linear, RBF, Polynomial, and Sigmoid could change the original data distribution in a way to help k-means clustering, like the Cosine kernel. Moreover, the Cosine kernel has concluded patterns with a wider/clearer range in voltage magnitude variations. Hence, KPCA

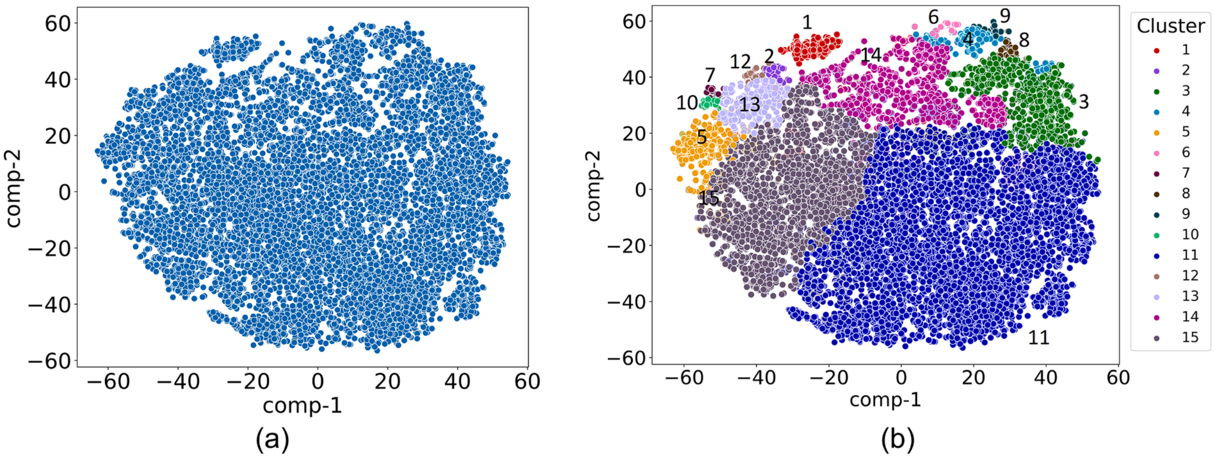


Fig. 5. Visualization of principal feature vectors (10D) by 2D t-SNE. (a) After KPCA with Cosine kernel, before k-means; (b) Clustered, after k-means with initial K = 15 without post-processing).

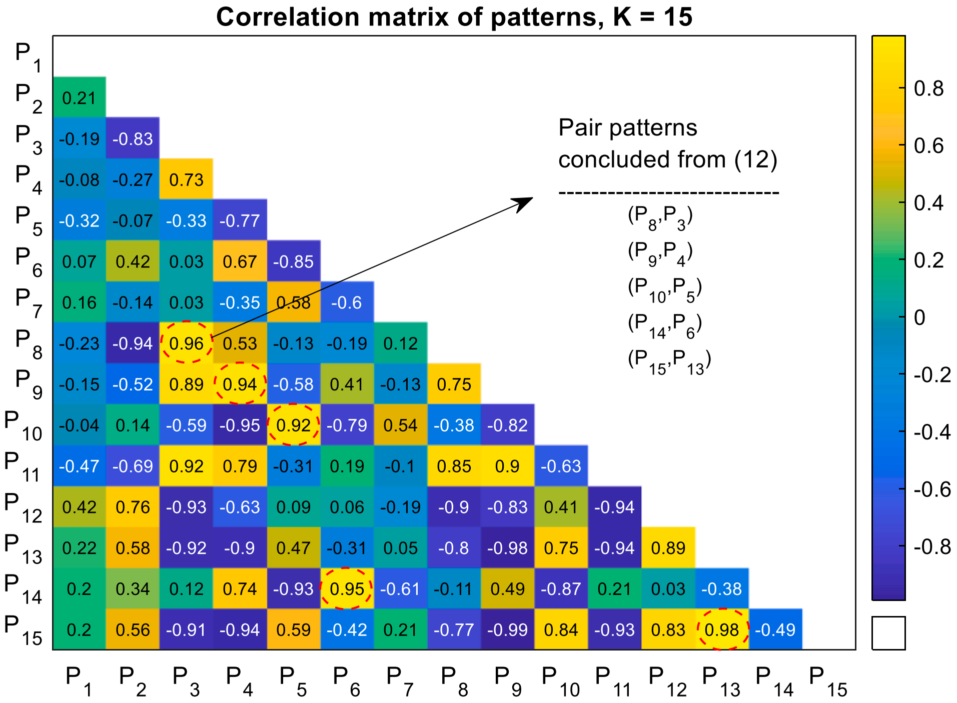


Fig. 6. Pearson correlation coefficients between the 15 initial patterns.

with Cosine kernel is chosen; after many runs of the function of KernelPCA by the “cosine” kernel on the input dataset, the singular values (2-norms of the principal components’ variables in the lower-dimensional space) corresponding to each singular mode index are shown in Fig. 3. The singular values are related to every one of the first 50 principal components. We have considered only the first p = 10 principal components (saving 88.9% information) during the training process of KPCA since almost the same results were obtained, as compared to 50 components, which could save 96.8% information. Although the choice of K in k-means is up to the user, in this work, firstly, this selection has been obtained using the elbow point analysis. Hence, according to Fig. 4, the interval to choose the optimal number of clusters is marked by a rectangular showing a number from 13 to 17 (selected according to the knee of the curve). Fig. 4 states different inertia (6) (within-cluster sum of squares criterion) versus a different number of clusters. After checking the results of all the six numbers of K, $K = 15$ was chosen as the elbow point [49] (see sub-Section 6.2). In this way, some good interpretations coupled with physical reality were achieved.

A function 2D t-SNE was then used to visualize the no-labeled principal feature vectors (10 dimensions) and the clustered principal features (10 dimensional), as shown in Fig. 5a and b, respectively. The parameters of t-SNE (input:10D, output:2D) were set as Barnes-Hut algorithm, Euclidean distance metrics and perplexity = 30. The best 2D embedding space for visualization was chosen by selecting the minimum loss values from running t-SNE 100 times. Fifteen clusters, colored in Fig. 5b, express 15 possible patterns whose few overlap, as seen among some clusters, could show a need to plot more dimensions (principal components more than three). However, each time series belongs to only one cluster (hard clustering).

The 15 cluster centers, which are 10D time series of rms voltages, are an average of all 10D time series belonging to the 15 clusters. All 15 cluster centers are reconstructed as 600D time series using an inverse function of KPCA. Then, the initial reconstructed data sequences

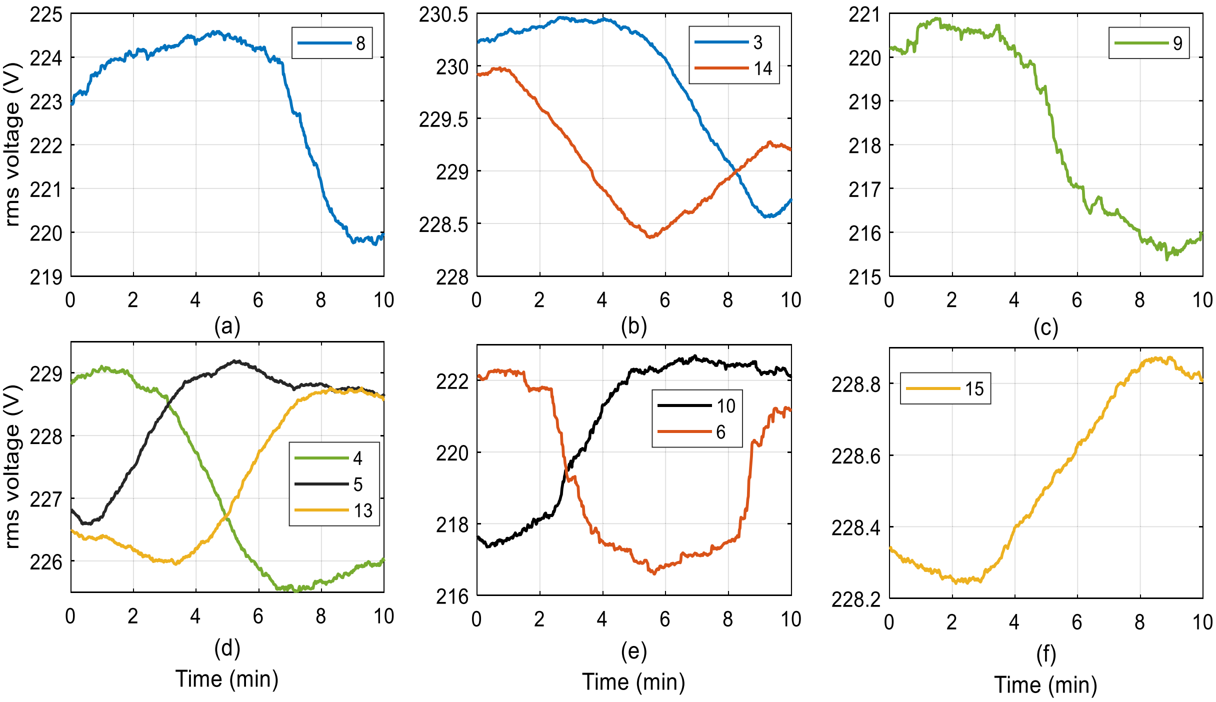


Fig. 7. Patterns with similar shape of variations, concluded from Fig. 6. (a) and (b) (P_8, P_3); (c) and (d) (P_9, P_4); (e) and (d) (P_{10}, P_5); (b) and (e) (P_{14}, P_6); (f) and (d) (P_{15}, P_{13}).

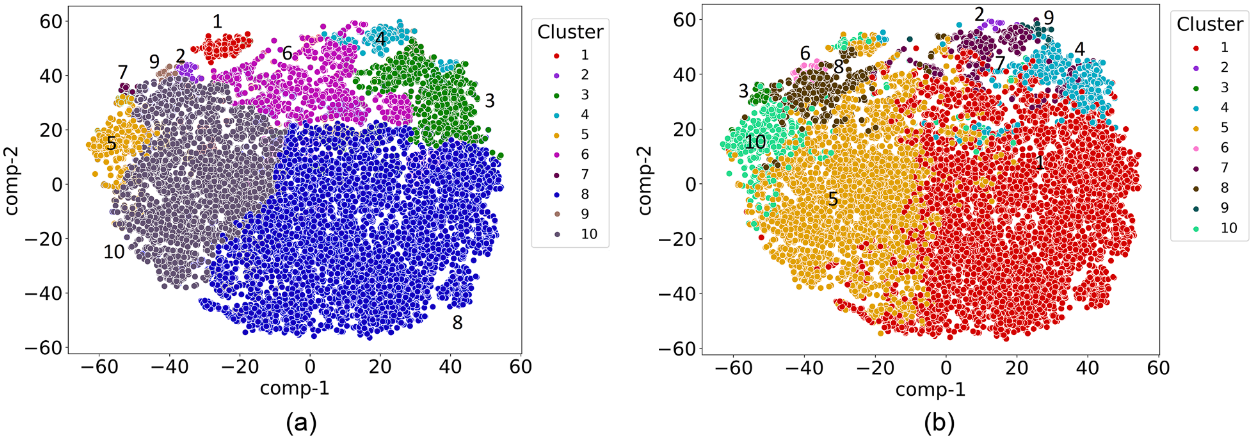


Fig. 8. Visualization of feature vectors (10D) by 2D t-SNE. (a) Clustered as new labeled with *Knew* = 10 after post-processing; (b) Clustered with initial $K = 10$ without post-processing).

(representative patterns for the 15 individual clusters) are concluded as $P_1, \ldots, P_{15}$. The correlation matrix between the obtained patterns (10) calculates the cross PCCs, as shown in Fig. 6. The pair patterns with a highly similar shape of voltage variations were calculated (using (11) and (12)) and indicated in this figure as (P_8, P_3), (P_9, P_4), (P_{10}, P_5), (P_{14}, P_6), and (P_{15}, P_{13}). Fig. 7 also shows a plot of the similar patterns, in which the similarity between each pair pattern can be observed. These 10 patterns are plotted in six subplots just to get less space in the paper. In this way, patterns 3 and 14 (Fig. 7b), 4, 5 and 13 (Fig. 7d), and 6, and 10 (Fig. 7e) are plotted in the same subplot because they have a similar magnitude range of voltage.

Comparing Fig. 8b with the clusters from the proposed framework with post-processing patterns (Fig. 8a) shows a better distinction of the clusters in the proposed framework.

See, for example, cluster 1 in Fig. 8a, a mix of clusters 5 and 10 as non-clear in Fig. 8b. Moreover, there is no clear separation in clusters 2, 4, 9 and 7 in Fig. 8b. Besides this, checking the similarities between patterns concluded from Fig. 8b shows the highly similar shape of variations in patterns 1, 4, 7 and 9, which means that four clusters have been separated for no reason.

According to Fig. 9, the following observations can be made:

(i) Cluster 8 with 5376 samples (Fig. 9g and blue circles in Fig. 8a) and cluster 10 with 3052 samples (Fig. 9i and gray circles in Fig. 8a) are the biggest, respectively.

(ii) The differences between the patterns are in the range of the rms voltage magnitude, the shape of variations/oscillation (growth pattern), and variations times. Cluster 9 (Fig. 9h) has a maximum range of variations.

(iii) The patterns realized in Fig. 9 are smoother than the real samples (Fig. 2) because of the intrinsic characteristic of averaging in k-means.

In order to show that the patterns are a good representation of all

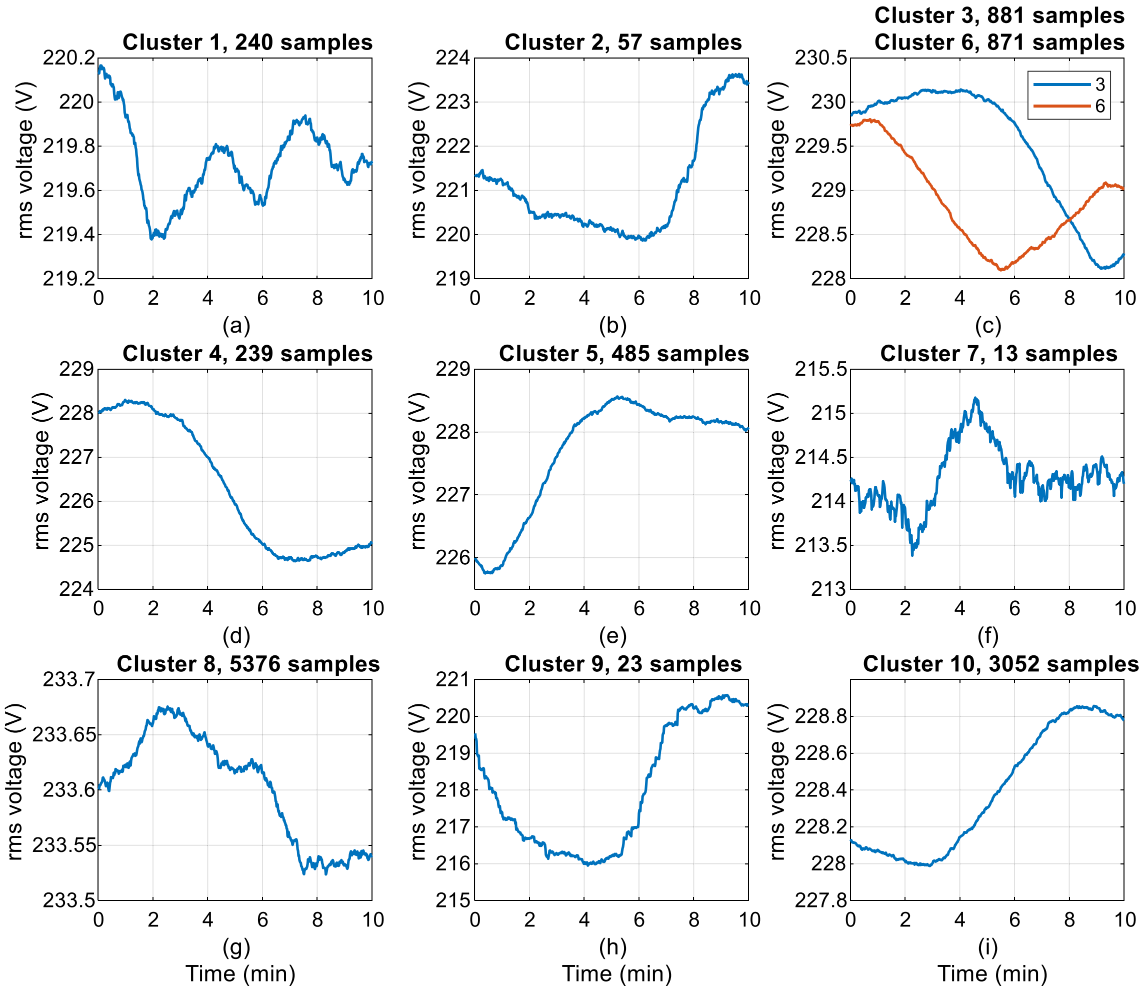


Fig. 9. Reconstructed patterns (cluster centers) including the number of samples belonging to each cluster. (a) Cluster 1; (b) Cluster 2; (c) Clusters 3 and 6; (d) – (i) Clusters 4 – 10, respectively.

samples, some samples, along with the patterns surrounded by a 99% confidential interval (CI)[2], are shown in Fig. 10 for clusters 1, 7, 9 and 10. The high and low values of the highlighted area for CI are calculated from $\bar{x} \mp 2.58\sigma/\sqrt{N}$, where each cluster is considered as a matrix with N row samples and 600 columns, $\bar{x}$ is a mean value of each column, and σ is the standard deviation for each column. A 600D sequence of $\bar{x}$ makes a pattern. As can be seen in Fig. 10, the samples in each cluster 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.

Another observation from Fig. 10 is that a number of samples show some steps in voltage variations, but the pattern does not. This observation is clearer for cluster 8 (Fig. 10c). A criterion is used in [12] as a single-window index for quantifying the number of steps in voltage variations. An analysis of the ten patterns and their samples is explained in the next sections to show: 1) the good separation of the ten clusters and 2) the necessity of the obtained patterns beside statistical indices introduced in literature [6,18] over a sub-10-min time scale.

4.2. Analysis of the correlation between ten-new obtained patterns

In order to show there is a good separation between ten-new obtained patterns, the Pearson Correlation Coefficients (PCCs) are calculated, as shown in Fig. 11, as a correlation matrix. The coefficients are less than + 0.9, which means a low similarity between the growth patterns obtained from the proposed framework. The only correlation more than + 0.9 is + 0.92 for (P_3, P_8), the same as the pair pattern (P_3, P_{11}) seen in Fig. 6. (P_3, P_{11}) was not merged into one because there was already a higher correlation of + 0.96 for (P_3, P_8), as seen in Fig. 6 (according to our post-processing strategy (11)). Moreover, the Euclidian distance between (P_3, P_8) in Fig. 11 is 101.51, and P_3 is totally below P_8 in terms of voltage magnitude, which shows another difference between these two patterns. Additionally, t-SNE in Fig. 8 shows the separation of the two related clusters. A maximum negative correlation of −0.94 is obtained for (P_4, P_{10}), (P_8, P_{10}) and (P_8, P_9), which shows that there is an inverse behavior between the pair patterns, as seen in Fig. 9. This inverse behavior can also be seen in Fig. 8a since each pair pattern is somewhere in the 2D t-SNE plot with an angle difference of about 160°.

4.3. Analysis of applied statistics

4.3.1. Single-window indices on the whole dataset

Previous research has introduced statistical indices quantifying the voltage levels [6,18]. By employing the fourteen single-window indices (Table A.1), obtained over all 10-min windows of the whole dataset, the general situation of the dataset looks as follows: Voltage typically is varied by 0.5 V-5 V within a 10-min window, where a range exceeding

[2] CI shows that the patterns fall into the highlighted part with 99% confidence. It also confirms that the most samples (cluster members) would be around the CI area.

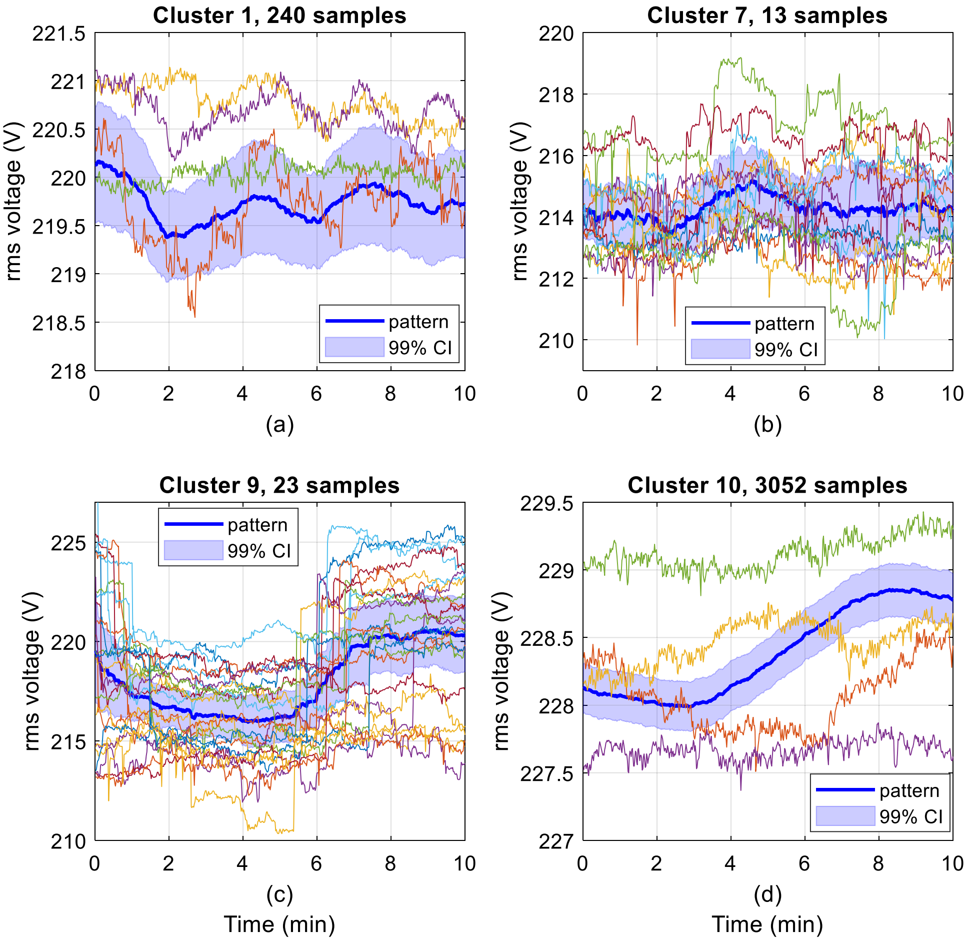
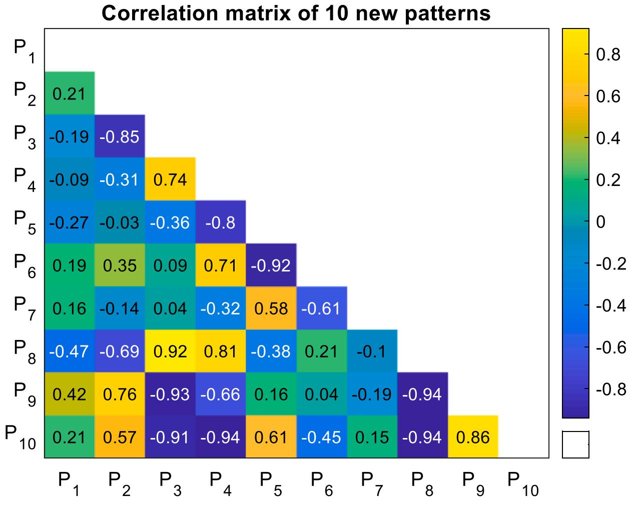


Fig. 10. Four clusters including patterns, 99% CI, and their samples. (a) Cluster 1 with four random samples; (b) Cluster 7 with all 13 samples (c) Cluster 9 with all 23 samples (d) Cluster 10 with four random samples.


Correlation matrix of 10 new patterns
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1										
P_2	0.21									
P_3	-0.19	-0.85								
P_4	-0.09	-0.31	0.74							
P_5	-0.27	-0.03	-0.36	-0.8						
P_6	0.19	0.35	0.09	0.71	-0.92					
P_7	0.16	-0.14	0.04	-0.32	0.58	-0.61				
P_8	-0.47	-0.69	0.92	0.81	-0.38	0.21	-0.1			
P_9	0.42	0.76	-0.93	-0.66	0.16	0.04	-0.19	-0.94		
P_{10}	0.21	0.57	-0.91	-0.94	0.61	-0.45	0.15	-0.94	0.86	

Fig. 11. The similarity/dissimilarity between ten new patterns.

1–2 V is common. The differences between higher-order statistics, compared to lower-order statistics for the indices quantifying the range in values (R100, R98, R90, R80), overdeviations (P100, P99, P95, PP0) and (VSV, Std.) indices, are higher. They quantify variations within a 10-min window. An opposite behavior is seen for underdeviation indices. The probability distribution functions for (R100, R98, R90, R80) and (P100, P99, P95, PP0) have a similar pattern (only a factor of two is the difference). VSV is slightly higher than Std., but somehow similar to a distribution function. Fig. 12 shows a cross-correlation between all indices mentioned. A low/high negative/positive value of the coefficient between two indices shows a strong correlation, and they vary together in the opposite/same direction. R98-R100 and P99-P100 have the highest positive 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 overdeviation indices), and P5 (from the underdeviation indices). As a conclusion derived from this section, the most suitable indices, along with Std. (which is somehow similar to VSV, with a 93% correlation), will be used for the 10-min windows within each cluster to show how well the 10-min time series are grouped into ten clusters, displaying some homogeneity between the time series within each cluster.

4.3.2. Selected single-window indices on the cluster's samples

The selected indices of R90, P95, P5 and Std. are employed (Section 4.3.1), and a probability distribution function (PDF) for each cluster, including their samples, is shown in Figs. 13-16, respectively. It is seen from all four indices that the probability distribution function for clusters 8 and 10 shows a softer curve because the clusters are the most dominant ones with the highest number of samples. There is also a clear separation of values for classical indices between different clusters in terms of the probability distribution. These results come from the well-separation of 10-min samples as grouped into ten clusters, displaying some homogeneity between the samples within each cluster.

The selected indices, which show an almost similar result for pair

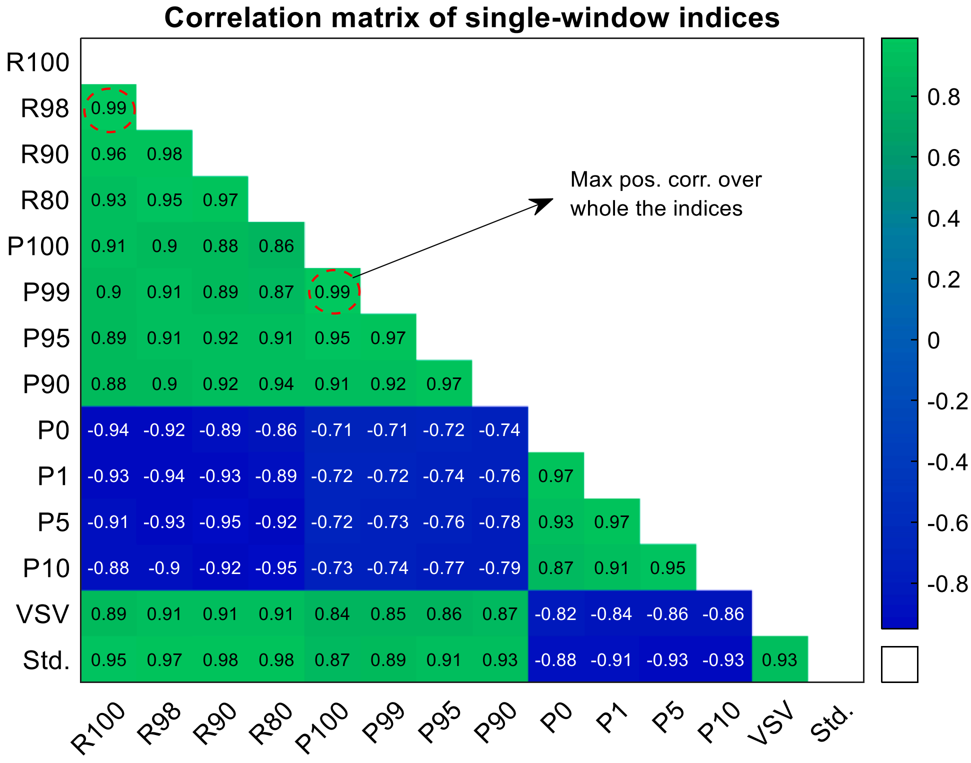


Fig. 12. Correlation coefficients between various single-window indices.

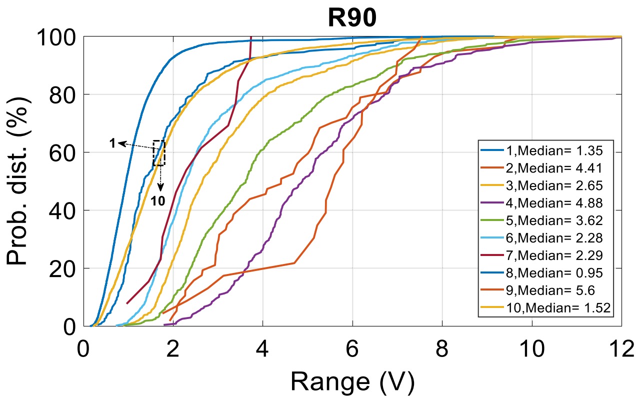


Fig. 13. PDF of R90 for the group of time series within each cluster.

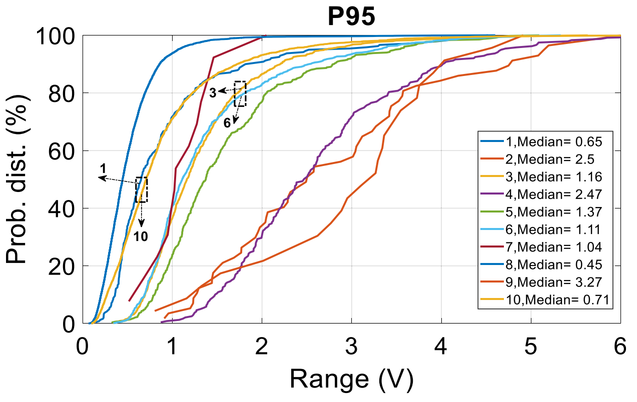


Fig. 14. PDF of P95 for the group of time series within each cluster.

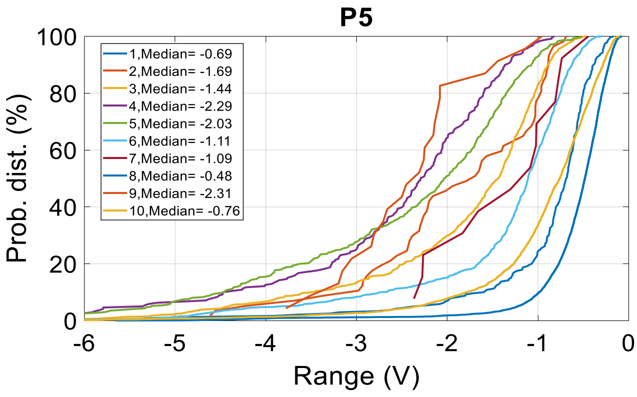


Fig. 15. PDF of P5 for the group of time series within each cluster.

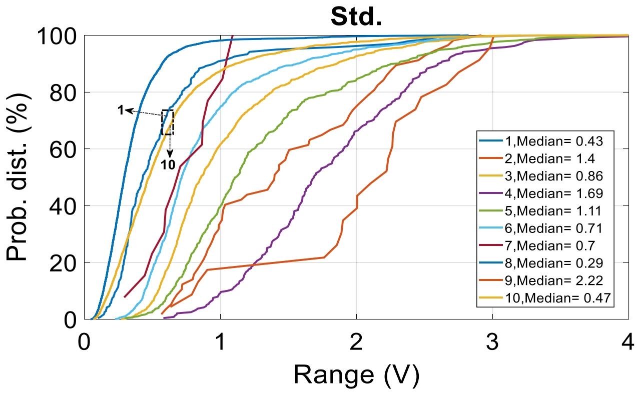


Fig. 16. PDF of Std. for the group of time series within each cluster.

Table 2
Selected single-window statistics for each pattern.

Pattern	R90		P95		P5		Std.	
no.	Range	(% V_n)	Range	(% V_n)	Range	(% V_n)	Range	(% V_n)
1	0.66	0.29	0.35	0.15	-0.31	-0.13	0.18	0.08
2	3.61	1.57	2.48	1.08	-1.13	-0.49	1.17	0.51
3	2.02	0.87	0.62	0.27	-1.37	-0.6	0.71	0.31
4	3.57	1.55	1.95	0.85	-1.61	-0.7	1.46	0.63
5	2.7	1.17	0.83	0.36	-1.87	-0.81	0.91	0.4
6	1.89	0.7	0.91	0.4	-0.71	-0.31	0.53	0.23
7	1.25	0.54	0.67	0.29	-0.58	-0.25	0.35	0.15
8	0.14	0.06	0.07	0.03	-0.07	-0.03	0.05	0.02
9	4.44	1.93	2.47	1.07	-1.97	-0.86	1.71	0.74
10	0.85	0.37	0.46	0.2	-0.39	-0.17	0.33	0.14

V_n: Nominal voltage; Colored highlights show nearby values of indices for different patterns.

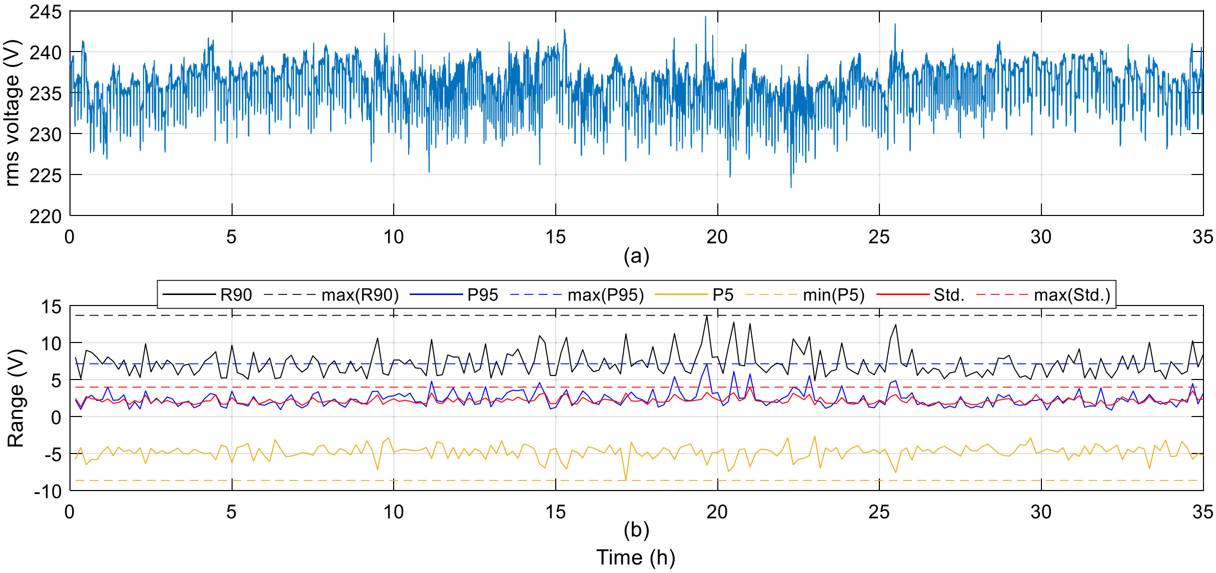


Fig. 17. A 35-h expression of measurements at location 47. (a) rms voltage (b) selected single-window indices.

clusters, are explained here:

- R90 and Std.: Clusters {1,10} (marked in Figs. 13 and 16)
- P95: Clusters {1,10} and clusters {3,6} (marked in Fig. 14)

However, for the cluster centers {1,10}, the shape of variations of rms voltage is different (Fig. 9a and i), and there is only a + 0.21% correlation between those cluster centers (Fig. 11). Also, the shape of variations for the cluster centers {3,6} is dissimilar (Fig. 9c), and a + 9% correlation is seen in Fig. 11.

4.3.3. *Selected single-window indices on the ten-new patterns*

In the following, the selected indices are applied to the ten patterns, 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:

- R90: {P2, P4}
- P95: {P2, P9} and {P3, P7}
- Std.: {P7, P10}

However, Fig. 9 (the patterns) and Fig. 11 (PCCs) show that the pair patterns have different patterns despite a similar range of variations seen in the statistical indices.

It can be concluded from Section 4.3 that the patterns obtained from the proposed framework are correctly separated, and their related time series have mostly similar behavior. Moreover, the statistical indices applied to cluster centers or samples may consider some clusters in the same category and cannot distinguish between clusters. The PCCs between cluster centers are a good measure to show the separation of the clusters. It can also be concluded that the statistical indices may not be enough to show a full picture of the sub-10 min real variations. Hence, beside the statistics, seeking 10-min window patterns from the proposed framework in this paper is essential.

5. Patterns for each single location

The post-processing part of the proposed framework makes it applicable for seeking sub-10-min patterns from multiple locations with time-limited measurements. Let's consider the location 47 out of the total 57 ones. This place is a detached house in Dalsland, southern Sweden, with a measurement period of 35 h (210 10-min windows). Fig. 17a shows the 1-s rms voltages, and Fig. 17b represents the selected single-window indices over the 35-h period. The maximum of each index is also shown as a dashed horizontal line. The lower values of the indices (Fig. 17b) indicate that the 1-s rms values are closer to the 10-min ones (Fig. 17a). Although the indices quantify voltage levels and the variations, the typical patterns need also to be found. Once again, given the necessity of our proposed framework, there can be two ways to assign the patterns for the samples of this location, as explained in sub-Sections

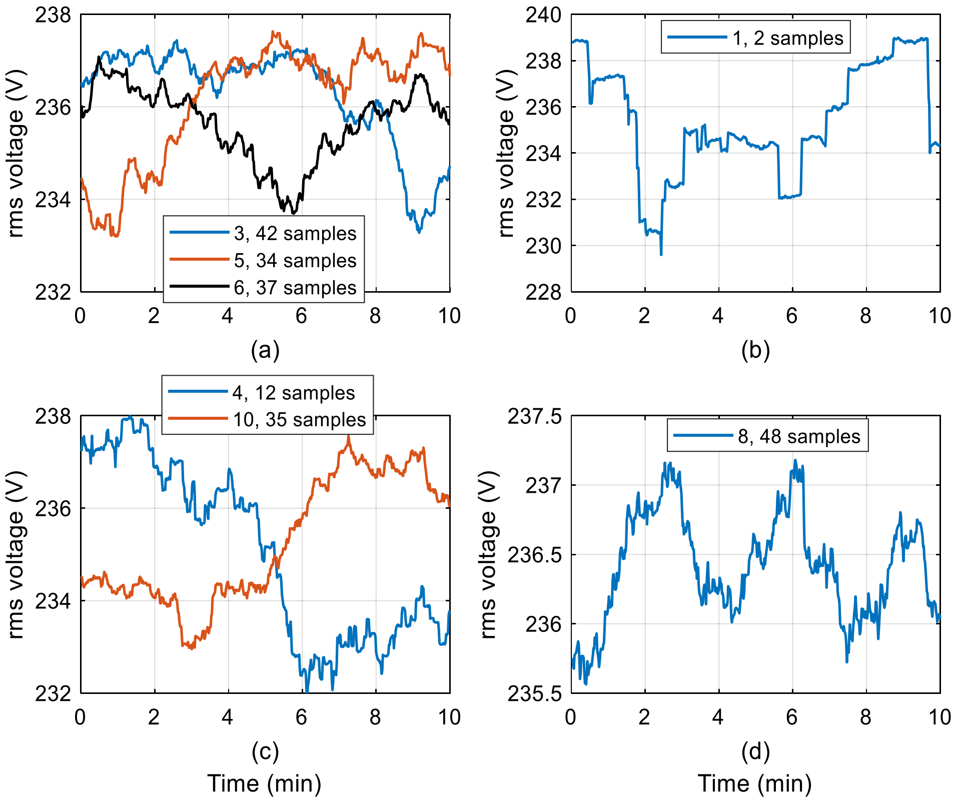


Fig. 18. Patterns extracted for location 47. (a) Patterns 3, 5 and 6; (b) Pattern 1; (c) Patterns 4 and 10; (d) Pattern 8.

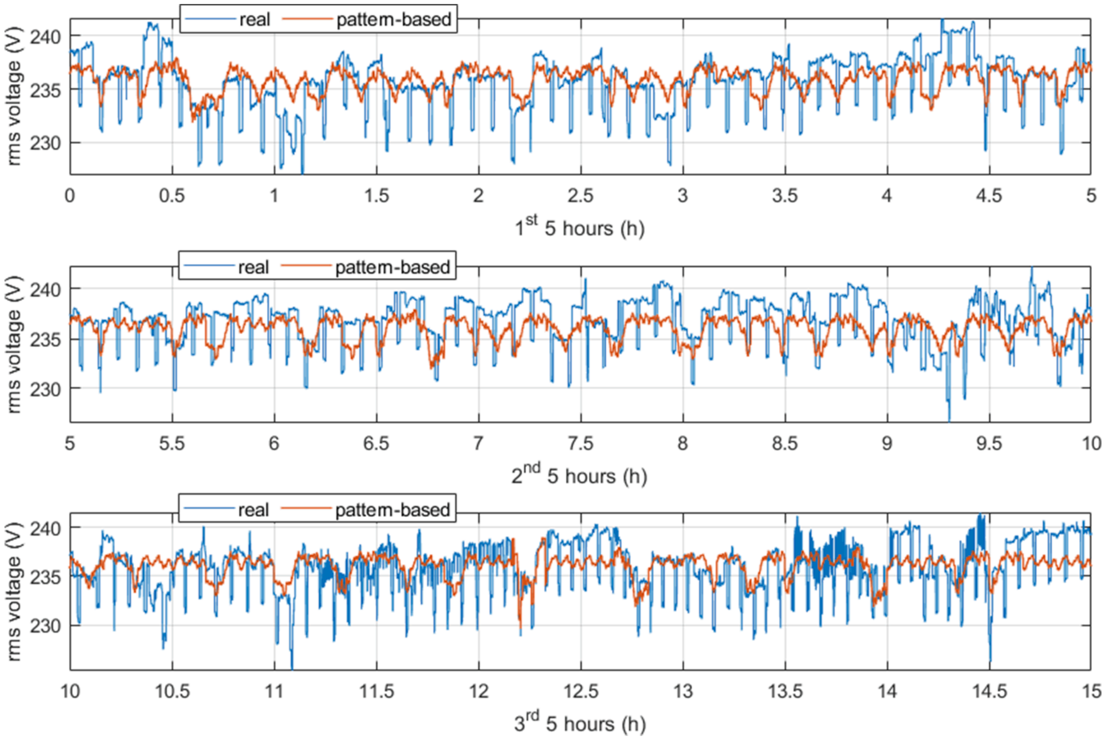


Fig. 19. Real and pattern-based rms voltage for 1st, 2nd, and 3rd 5 h of the 35 h measurements at location 47.

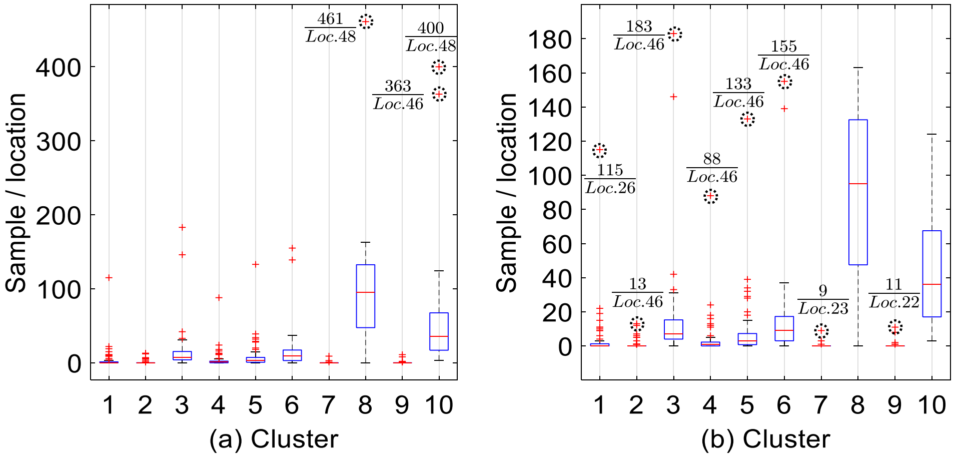


Fig. 20. Boxplot for samples of 10 clusters at 57 locations. (a) Full picture showing the importance of clusters 8 and 10, maximum values are indicated (b) Showing the contribution of clusters 1 – 7 and 9, maximum values (highest outliers are indicated).

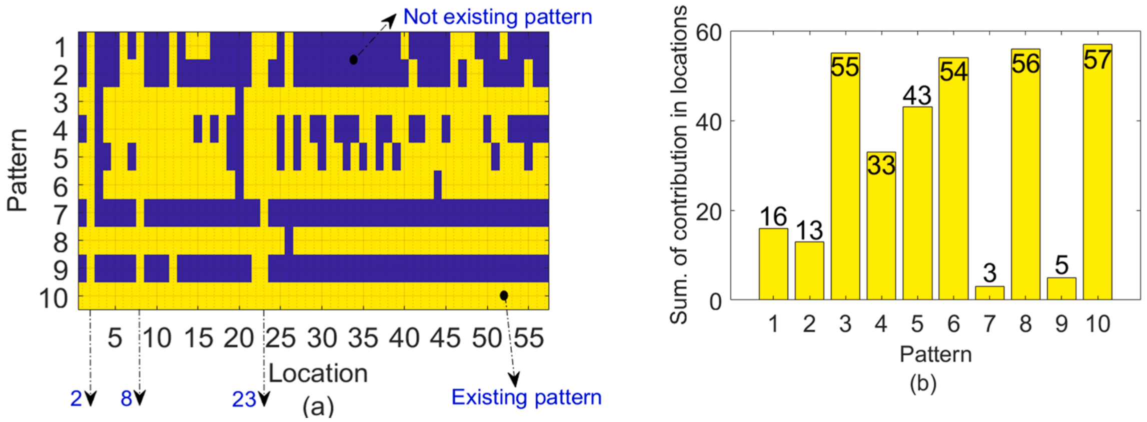


Fig. 21. Analysis of "pattern-location". (a) The pattern distribution into the location (b) The summation of the pattern's contribution to the location.

5.1 and 5.2.

5.1. Seeking patterns for each single location using the data from multiple locations – The first way

The first way is that after obtaining the labels (clusters 1–10) for each 10-min time series from the proposed framework, the patterns belonging to each separate measurement location are accessible. The clusters' labels and their number of samples per location 47 are as follows: Cluster 1 (2 samples), Cluster 2 (no samples), Cluster 3 (42 samples), Cluster 4 (12 samples), Cluster 5 (34 samples), Cluster 6 (37 samples), Cluster 7 (no samples), Cluster 8 (48 samples), Cluster 9 (no samples), and Cluster 10 (35 samples). Therefore, this location includes seven patterns (P_1, P_3, P_4, P_5, P_6, P_8, P_{10}) out of 10 existing ones. Other locations may include more/fewer patterns (Fig. 21a).

5.2. Seeking patterns for each single location using the data from multiple locations – The second way

Another way to obtain some patterns per location 47 is to take an average within each cluster obtained in sub-Section 5.1. In this way, more meaningful patterns are obtained for the location. Fig. 18 shows seven local patterns obtained specifically for the location in this manner. The extracted patterns were taken out from a number of time-limited samples for only location 47; therefore, voltage oscillations could be shown more clearly, when compared to the main patterns in Fig. 9.

Fig. 19 shows the first, second and third 5 h of the 35-h rms voltage measurement (Fig. 17a) for real and pattern-based values versus time. This figure shows that the voltage steps could not be detected by the different 5-hour-time series obtained from the seven patterns. However, the pattern-based voltages follow the real values.

Table 3
Information of three sample locations.

Location no.	Place	Date of measurement	Patterns obtained in the location
2	Hotel/Shanghai/China	2009–07-25	All 10 patterns, max. cont. (P_3, P_{10})
12	University campus/ New Delhi/India	2014–02-21	All except P_7, max. cont. (P_5, P_8)
53	University campus/ Gothenburg/Sweden	2018–03-21	P_3,P_5,P_6,P_8,P_{10} patterns, max. cont. (P_8, P_{10})

5.3. Triangle of cluster/pattern-sample-location

This section shows the importance of the clusters/patterns obtained from the proposed framework. The number of samples within each of the 10 clusters as grouped per each of the 57 locations is obtained and shown in Fig. 20. Clusters 8 and 10 are the two biggest clusters, and cluster 7 is the smallest one, as seen before in Fig. 9. Moreover, cluster 8 has a maximum sample number of 461/location 48. Then, cluster 10 includes values 400/location 48 and 363/location 46, as marked in Fig. 20a. Locations 46 and 48 are two detached houses at Kristinehamn

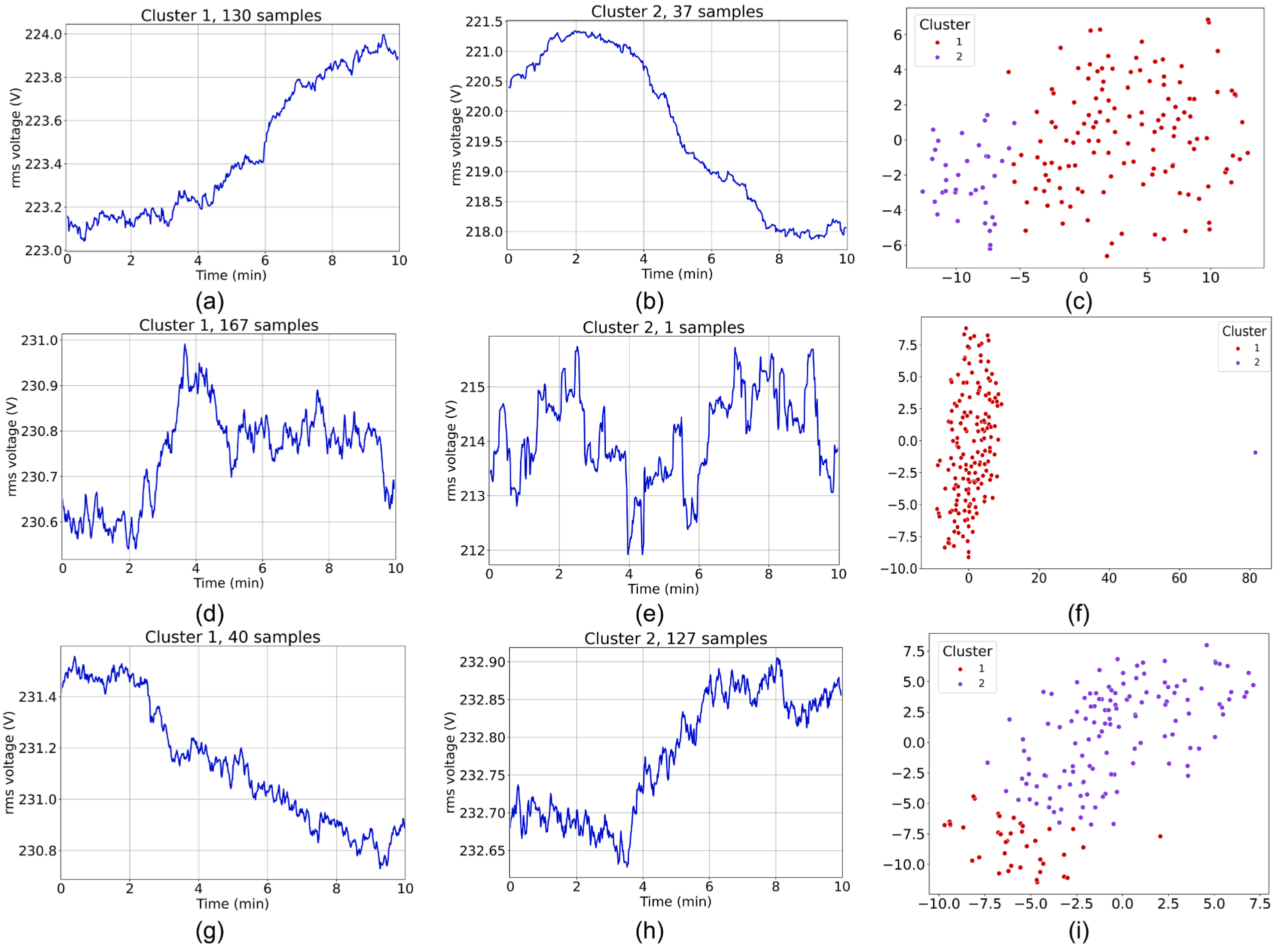


Fig. 22. Two patterns and visualization of feature vectors by t-SNE (10D-2D). (a) - (c) Location 2; (d) - (f) Location 12; (g) - (i) Location 53.

and Ludvika in southern Sweden with the longest measurement periods of about 182 (1091 10-min windows) and 202 h (1213 10-min windows) in the dataset (Table 1), respectively. These two locations include eight patterns (P_1 - P_6, $P_{8,}$ and P_{10}) out of 10 existing ones. Fig. 20b shows the contribution of other clusters. The highest outliers display a maximum number of samples per a defined location, while not considering location 48 in cluster 8, and 46 and 48 in cluster 10. Clusters 1, 7 and 9 have shown their own maximum contribution in the hotels placed at locations 26 (Vienna/Austria), 23 (Istanbul/Turkey) and 22 (Shanghai/China), respectively. The patterns in clusters 2 – 6 are seen more in location 46.

The distribution of 10 patterns in the 57 locations is shown in Fig. 21a. Pattern 10 is the only one seen in all locations (Fig. 21b). Patterns 3, 6, 8 and 10 are common, while patterns 4 and 5 are less common. This conclusion is also seen from the number of cluster memberships, as seen in Fig. 9. Other patterns are seen in a few locations. The locations including all the ten patterns are hotel rooms at locations 2 (Shanghai/China) and also, 8 and 23 (both in Istanbul/Turkey).

5.4. Seeking patterns for each single location using each location data separately

In this paper, clustering has been based on the whole locations in the dataset $X^{7356\times600}$. Sub-sections 5.1, 5.2 and 5.3 also show the proposed framework's applicability for each location. This section shows why the proposed framework has not been run for each location separately. The answer is that the measurement periods for each location are not good enough to seek the real patterns for each location separately. The proposed framework is run for each single location distinctly to confirm this. Table 3 gives the information regarding three sample locations for a period of about 27 h (167 10-min windows for locations 2 and 53, and 168 windows for the location 12). After many empirical tests, the number of clusters was chosen as $K = 2$ for all these locations. Fig. 22 shows the two reconstructed patterns and the related t-SNE (10D-2D). In this manner, clustering is based on each location separately. The patterns are different per location. However, they can still be seen in Fig. 9 (all locations were considered). For example, the patterns in Fig. 21a and b for location 2 are close to the patterns 10 and 3/4 in Fig. 9. This shows the effectiveness of our proposed framework and the feasibility of the obtained patterns for each location, which mixed all locations and was run once (for instance, Fig. 18 for location 47).

Nevertheless, the two obtained patterns per location are for a period of 27 h, which cannot represent the total behavior of the locations. The scheme (choosing two clusters for each location separately) cannot obtain real patterns, and some abnormal patterns may occur, as seen in Fig. 22e, in which cluster 2 has only one sample.

Hence, although the time-limited measurements used in this work are enough to quantify the variations in the sub-10-min scale [18], a recommendation for running the proposed framework at each location is recording measurements as weekly, monthly, seasonally or yearly per location. This can be different for each location with different behavior; this is because the patterns per location are sometimes repeated daily/week; hence, the daily/weekly measurements might be enough. Then, by choosing $K = 10$ (or maybe another value), the real patterns/locations can be obtained. Since each location is investigated separately, there may be no need for the post-processing part in Fig. 1, because the patterns with almost similar shapes of variations and different voltage magnitude ranges can be considered two distinct patterns at the location

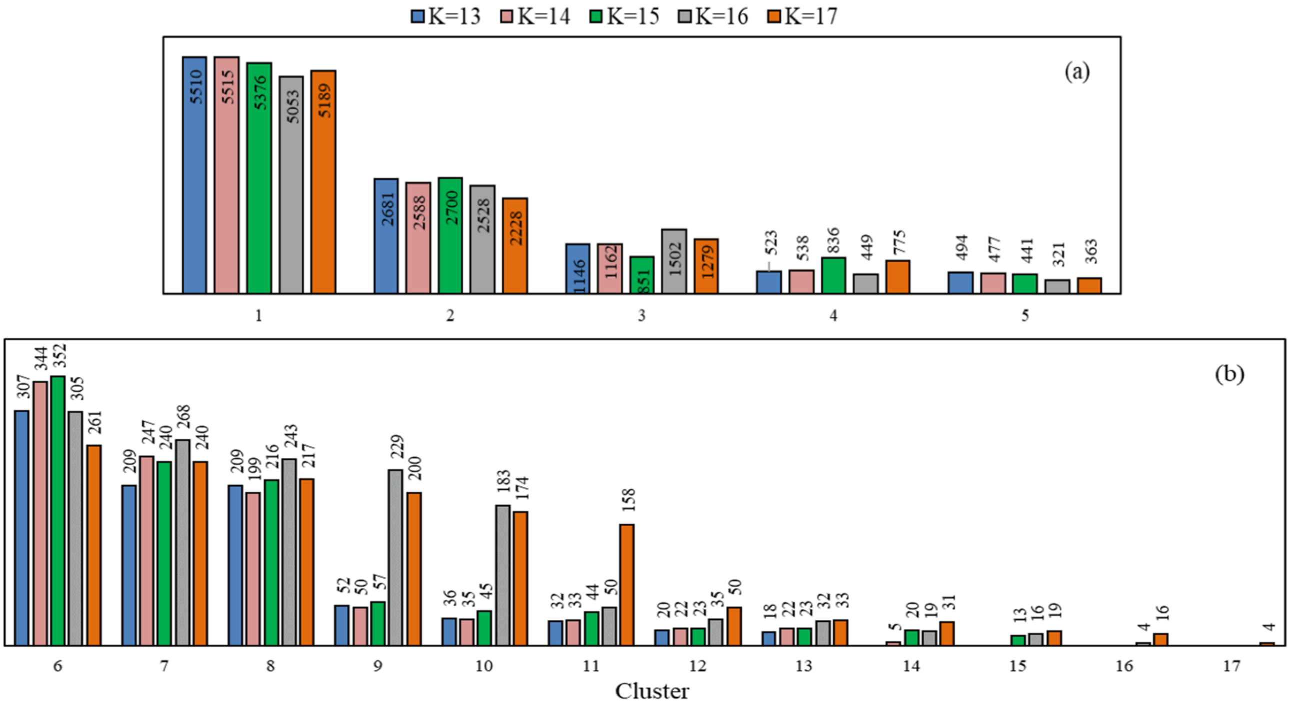


Fig 23. Number of samples for K: 13–17. (a) Clusters 1–5; (b) Clusters 6–17. The number of samples per cluster is sorted.

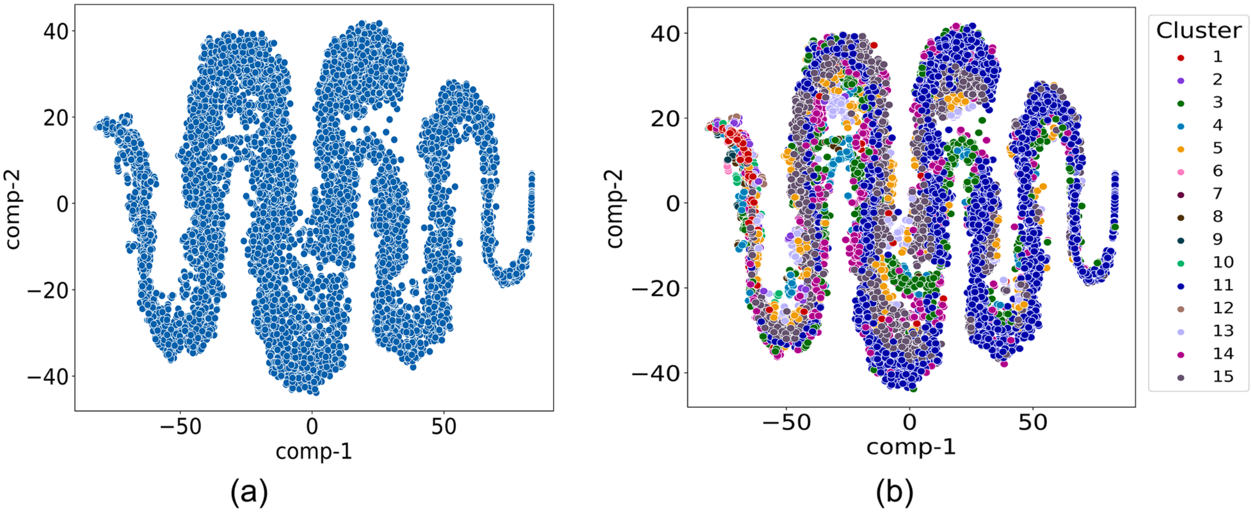


Fig. 24. Visualization of original feature vectors (600D) by 2D t-SNE without KPCA. (a) Before k-means; (b) Clustered, after k-means with initial K = 15 without post-processing.

[50]. For example, the voltage magnitude range for the location 47 is somewhat between 230 V and 240 V, as shown in Fig. 17a, 18 and 19. Hence, a different range out of the period in patterns with similar variation shapes must be considered an individual pattern.

6. Discussion and future works

6.1. Other applications of the proposed methodology

The work presented in this paper uses time series of 1-s rms voltages over a 10-min window. As one way to show the oscillations observed in the sub-10 min period, this work sought the sub-10 min patterns of rms voltage using an unsupervised learning method followed by a new post-processing approach. In that way, some important patterns in the large data of low voltage variational measurements from multiple locations were indicated. Compressing a huge amount of data from power-quality monitoring was also done in this study since the input data size was reduced from 600 to 10 by a factor of 60. The proposed scheme, beside the statistics, is the first step before studying the potential impacts of the sub-10-min variations on equipment. The proposed scheme is scalable and computationally cheap, which makes it appropriate for seeking the typical patterns in the big data domain.

In comparison with some previous research [51–55,58] that uses predefined scalar features in a labeled dataset to locate the source of voltage dips which are some kind of steps in voltage magnitude, one important application of our proposed method can be solving that two-classes problem in an unsupervised intelligent scheme by using measurements from multiple power-quality monitors. In this way, the input of KPCA is the rms voltages and currents (merged as one signal or into two separate signals) with a high time resolution over a selected window, including

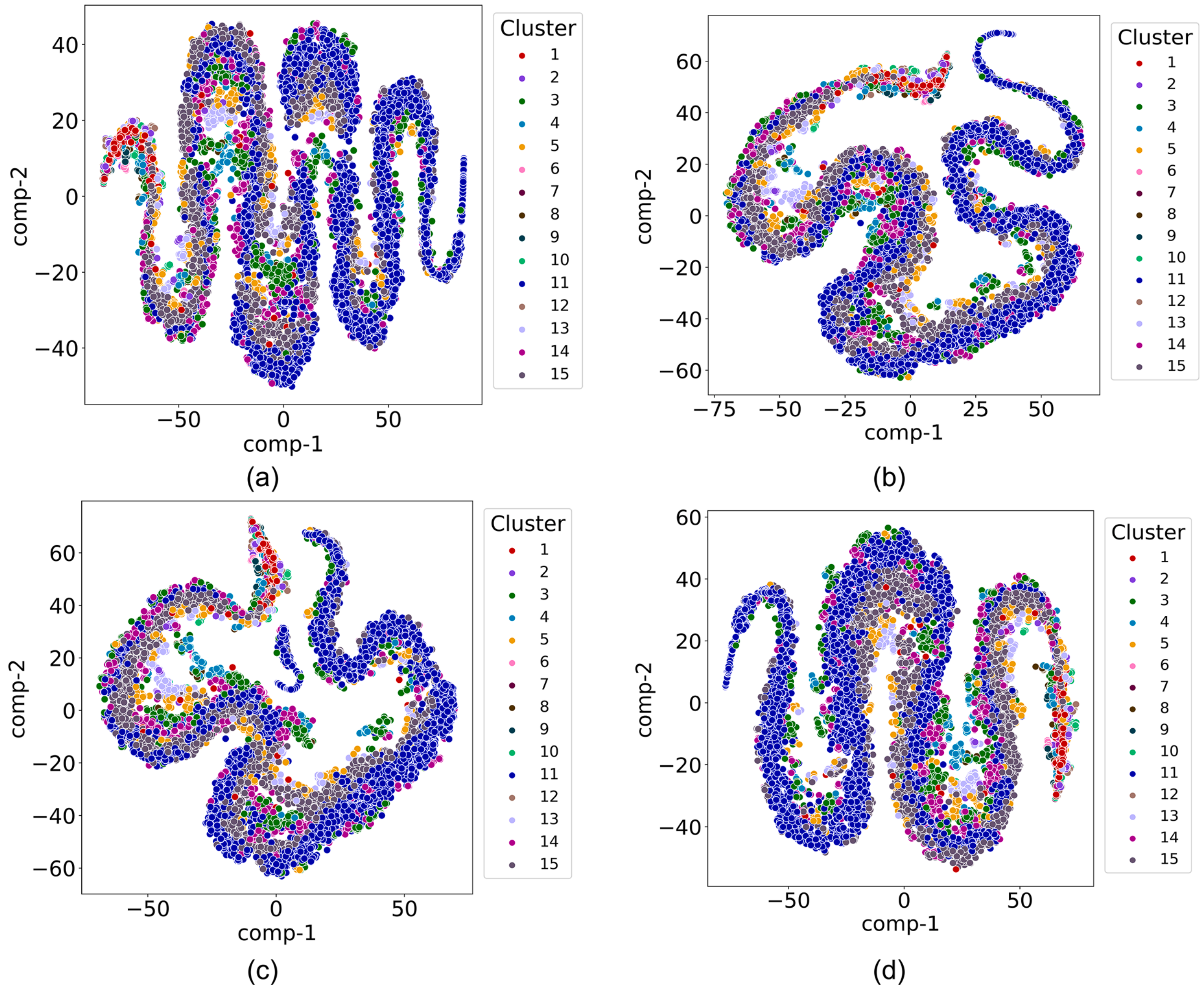


Fig. 25. Visualization of clustered principal feature vectors (10D) by 2D t-SNE using KPCA-different kernels, after k-means with initial K = 15 without post-processing. (a) Linear; (b) Polynomial; (c) RBF; (d) Sigmoid.

pre and after dip cycles. The output of k-means will show two groups, by which an expert can label the group's members as downstream and upstream as the source location of voltage dips.

6.2. The number of initial clusters in the proposed framework

The k-means clustering requires the user to select the number of clusters in advance. In our study, the interval to select an optimum K number was from 13 to 17 (Fig. 4). Fig. 23 compares the five numbers and their samples to group the $m = 11237$ samples. As can be seen, somewhat 3 to 8 main clusters are seen for all different ranges of K. Choosing $K = 14$, 16 and 17 splits the clusters for no reason. Also, their last cluster has only 5, 4 and 4 samples, respectively. Between $K = 13$ and 15, through multiple checking of patterns showing different values and variation shapes, $K = 15$ was chosen, so that some good interpretations coupled with physical reality were found.

6.3. Some other feature-size reduction tools, distance measurements and clustering methods

During the preparation of the proposed framework, different scenarios have been investigated as follows:

6.3.1. Clustering without KPCA

The results without using KPCA and only k-means were investigated. Fig. 24 shows a 2D visualization of the input data for 15 initial clusters. As can be seen, there is no good separation between clusters when compared to Fig. 5b (KPCA + k-means). A selection of three clusters

Table 4
A comparison of our proposed framework and ref. [37].

Method	Case study	Input samples	Window length	Time resolution	Measurement locations	Principal components	Clusters	Used workstation
Proposed framework	rms voltage	11,237	600 (10 min)	1 s	57	10	10	Intel-i7 8700 K-3.7 GHz × 12 CPU, 16 GB RAM, NVIDIA GeForce RTX 2080, Ubuntu 20.04.3 LTS-OS
[37]	Voltage harmonics*	365	144 (24 h)	10 min	1	16	2	Intel-i7 3.4 GHz × 12 CPU, 48 GB RAM, NVIDIA Titan Xp 12 GB GPU.

* V_2, V_3, V_4,….

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

Name	Time (s)	
	Proposed framework	[37]
Training KPCA/DAE and feature extraction (1000 runs)	12.5	142.88
Clustering (100 runs)	0.19	0.15
Reconstruction from cluster centers	0.23	0.29
t-SNE	40.25 (100 runs)	78.86 (50 runs)
Total	53.17	222.18

would be effective for Fig. 14 (only k-means), which cannot show the real patterns of our dataset with 11,237 input samples. Hence, using KPCA is essential in the proposed framework.

6.3.2. *Clustering by considering different kernels of KPCA*

The clustering results for different kernels of KPCA using optimal parameters are also shown in Fig. 25 (the results of the Cosine kernel have been already discussed in Fig. 5b). The Linear kernel (Fig. 25a) does not change the distribution of points in the original dataset much (Fig. 24a). There are also not so many changes while using the Sigmoid kernel, as shown in Fig. 25d. The Polynomial (Fig. 25b) and RBF (Fig. 25c) kernels have changed the original data distribution's shape almost similarly. However, none of the used kernels could change the original data distribution like the Cosine kernel shown in Fig. 5b, in which the dataset in the 2D plan is opened and distributed so that it can help the k-means to group the clusters more easily. Hence, only the Cosine kernel assisted k-means clustering in finding the initial centroids in a better data-distributed space. The Cosine kernel could map mapped the original data space x_i (11237 10-min windows with 600D obtained from the measurements) to the higher dimension space $\Phi(x_i)$ and apply a non-linear combination as cosines of the vectors (4). Then, an eigen analysis was done, and the feature vectors were projected on the first 10 dominant eigenvectors/principal components. Another observation of KPCA is that it can compress the Euclidean distance of intra-cluster pairs while preserving the Euclidean distance of inter-cluster pairs. That type of compressibility, due to the Cosine kernel in KPCA, could considerably help k-means. Beside all these reasons, the selection of Cosine kernel has been based on the concluded patterns with a wider/clearer range in voltage magnitude variations.

6.3.3. *Different feature size reduction tools and required time analysis*

As shown before, KPCA with Cosine kernel was chosen as a simple tool to reduce the feature size from 600 to 10. However, other feature size reduction tools like a deep autoencoder (DAE) can be considered in future works instead of KPCA to see how the clustering results would be for high-resolution time series (i.e., 600D). Nevertheless, using a DAE would not be as simple as the used KPCA. Table 4 shows a structured-based comparison of our proposed framework (KPCA + k-means + post-processing) and (DAE + k-means) [37]. As shown, the proposed framework/ref. [37] reached 10/2 clusters/patterns, while having a higher/fewer number of input samples, longer/shorter window lengths, higher/lower time resolutions, less/more principal components and somehow, the slower/faster workstations. The findings, as can be seen from Table 4, confirm the much simpler schema of our proposed methodology.

Although the proposed methodology is intended for off-line use and aimed initially at obtaining general knowledge about a new phenomenon (10-min rms voltage variations), Table 5 shows the detailed time required for running the proposed methodology and the method in [37]. The total running time of the proposed framework is about 24% of the total running of [37]. Note that the post-processing part of the proposed framework checks the patterns using a number of mathematics calculations/rules, emulating a human expert; hence, that takes only a few milliseconds.

6.3.4. *Different distant measurements in k-mean*

Replacing the Euclidian distance used in k-means by distance time warping DTW [33] can be another future work. This way, the time shift between the extracted patterns will not be considered a distance or different criteria. Concerning the ten obtained patterns, the minimum DTW and Euclidian distances are calculated for (P_3, P_6) (Fig. 9c) as 176 and 26, respectively. In contrast, centered Cosine distance concluded a ($1 - 0.09 = 0.01$), which is the second-largest distance between pair patterns. Being close patterns (P_3, P_6) by the DTW and Euclidian shows close voltage ranges and far shape of variations, despite an observed time shift between them.

6.4. *Sub-10-min variations from both views of power-quality and ML*

6.4.1. *ML view*

The ten clusters' samples (time series with 10-min windows) show some differences, depending on the intra-class variance. However, the overall pattern of the samples remains largely the same. Since each pattern is an average of its own samples, the fewer the number of samples within a cluster, the more similar those samples to the cluster center (pattern). Hence, all clusters (for example, clusters 7 (Fig. 10b) and 9 (Fig. 10c)), except 3, 6, 8 and 10, look very similar to their own samples.

Each pattern is representative of a number of 10-min windows, which can be considered "ten patterns" according to the actual recordings. According to Figs. 9 and 21a, patterns 3, 6, 8 and 10 are the common patterns, while patterns 4 and 5 are less common. Other patterns belong to clusters with fewer samples and are seen in few locations. Patterns 3 and 6 show a single triangle form superimposed on a very small variation in the rms voltage among the four common patterns. Pattern 8 shows small random variations, and pattern 10 indicates a positive ramp form of variations.

From the output of the proposed framework, the samples 107, 264, 271 and 10756, as shown in Fig. 2, belong to the patterns 8, 3, 6 and 5, respectively. The three real samples of 264, 271 and 10,756 are more following the patterns of 3, 6 and 5 than the sample 107 in cluster 8. Because cluster 8 is the biggest, even though the overall pattern of the 5376 samples is similar, averaging many samples within the cluster to generate the cluster center (pattern 8) may make some differences in oscillation values between the patterns and real samples. However, by considering the patterns for every single location and extracting new patterns per location (like the patterns per location 47, Fig. 18), "some typical patterns" per every single location can be determined. In that way, the patterns per location can be much more similar to the real recorded samples. For the locations with longer time measurements series, such as locations 46 (about 7.5 days) and 48 (about 8.5 days), finding the medoids [56] within each obtained cluster per location is recommended. This way, the most center sample within the cluster will be selected as the cluster pattern, which may show more real patterns than the averaging method discussed in Section 5.

6.4.2. *Power-quality view*

There are some reasons or physical phenomena behind the 10 general obtained patterns, depending on the upstream grid (how weak or strong it is and the loose connections) and the type of connected loads close to the power-quality monitors at the 57 different locations studied in this paper. Equipment that exhibit continuous, rapid load current variations (mainly in reactive component) can cause fast voltage variations at a sub-10-min scale. Examples of such loads are starting water boilers, microwaves under non-nominal power conditions, air conditioners, elevators and printers. Variations in the generation capacity of PVs and wind power installations, as well as the EV charging and starting electric heat pumps, and transformer tap changer operation (see steps in Fig. 18b), are other sources of fast voltage variations. In order to have a closer look into the patterns-sources in different locations, a reverse check from some of the patterns/locations is done to see which locations have similar patterns. This will somehow show that the places in the locations may have used similar loads or similar equipment exist nearby. For example, as can be seen in Fig. 21, locations 2, 28 and 23 have the similar patterns P_1 to P_{10}. These locations are hotels in Shanghai/China and Istanbul/Turkey, respectively. The hotels in locations 3 (Prague/Czech) and 20 (Skelleftea/Sweden) include the similar patterns P_8 and P_{10}. The hotels in locations 4 (Sarajevo/Bosnia and Herzegovina), 35 (Turin/Italy), 39 (Kiruna/Sweden) and 51 (Indian restaurant in Gutenberg/Sweden) have the similar patterns P_3, P_4, P_6, P_8 and P_{10}. The hotel in locations 6 (Milan/Italy) and the detached houses in 46 (Kristinehamn/Sweden) and 48 (Ludvika/Sweden) consist of similar patterns as all others, except P_7 and P_9.

The observations, thus, show that the mentioned hotels or the detached houses in different countries may have used similar loads, or similar equipment exist nearby. Moreover, comparing locations 3 and 20 with locations 6, 46 and 49 can show that first, the locations 6, 46, and 48 have more connected loads. Secondly, the patterns P_8 and P_{10} are seen in all five locations, thus showing that they may have similarly connected loads. Note that there is no detailed information about the connected loads in the 57 locations; some of the loads may have multiple patterns, and two different loads may have similar patterns. However, by having information about the connected loads, the results obtained from this study in regard to the 10 patterns can help to understand the sources behind the patterns. In this way, using the 10 patterns can develop future standards/classification methods (labeling patterns through the sources causing the patterns) and also methods for testing/putting requirements (according to the shape and voltage level of patterns) on the connected equipment.

Another observation is that there is a clear separation of clusters for the statistical power-quality indices, which showed the well choosing of the samples within clusters by the proposed framework. Moreover, the patterns are a good representative of the clusters. For example, patterns 3 and 6 have the value of R90 as 2.02 and 1.89, respectively (Table 2), which is somewhat close to the median value of the index for the clusters 3 (2.56) and 6 (2.28), as can be seen in Fig. 13.

Another point about the ten patterns is that they cannot detect multiple steps in rms voltage variations, but they show the single steps in patterns 2 and 9. Although the 10 min period selected for our study is according to standard IEC 61000–4-30, considering some other windows like 5 min may generate some patterns showing the multiple steps of variations.

6.5. Supplementary works

As mentioned in Section 5, another way to find more actual patterns for each location is weekly, monthly, seasonally or yearly measurements. On the other side, the authors of this paper tried to consider feature engineering. They inputted all the 14 statistical indices (Section 4.3.1) beside the 10-min windows as the input features to the proposed framework. However, the high correlation between some indices did not positively affect the clustering results. Despite this, using only the four selected indices as the input features added to each 10-min window for clustering may make more complete patterns (or maybe similar). Quantifying the current variations and then voltage variations, due to the presence of solar powers, electric vehicle charging, wind powers, electric heat pumps, and railway stations, is needed by installing a PQsmart monitor close to the equipment for future works. Measurements at higher voltage levels like the medium voltage at industrial installations and the impact of the variations on the connected equipment are also needed for extra investigations. Another future study will be seeking the patterns for variations in frequency and harmonic voltages recorded from multiple power-quality monitors in a sub-10 min period, which is an ongoing work pursued by the authors.

7. Conclusion

A comprehensive framework for short-term measurements was proposed to seek the sub-10-min patterns in rms voltage variations from the data at multiple locations. The framework used an unsupervised learning term as the kernel principal component analysis followed by principal feature clustering and a term as a suggested post-processing approach to the initial patterns. The proposed framework was applied to measurements from 57 low-voltage locations worldwide through the years 2009 to 2018. Fifteen initial clusters/patterns were converted into ten new clusters/patterns using the clusters' merging strategy with highly similar patterns used in the suggested post-processing approach. Since there was a time limitation in the measurements, all locations were considered together expertly. This ensured a level of the generality of the patterns and also allowed the comparison of locations. The ten extracted patterns, 2D embedded space (t-SNE) plots, and single-window statistical indices showed that the proposed framework effectively extracted patterns. A statistical analysis also confirmed that a complete picture of sub-10-min oscillations needed both single-window indices (quantifying levels and variations) and the proposed framework (quantifying patterns).

The results also showed the contribution of the ten general patterns in 5% to 100% of the 57 locations, of which four patterns were seen as the most common. Three hotel rooms in Shanghai/China (one) and Turkey/Istanbul (two) included all the ten patterns. The feasibility of the obtained patterns from multiple locations was also confirmed for the single locations (separately) as the typical patterns. However, running the proposed framework for every single location with short measurements led only to two patterns per location, which were not matched to the real samples and had some abnormalities when compared to the patterns extracted for each location from running the proposed framework for multiple locations with the presented post-processing.

It is also worth mentioning that the proposed framework can be applied to any kind of signals like sub-10-min harmonic voltage/current, power consumption and frequency. The proposed framework is scalable and computationally cheap, which makes it appropriate for seeking typical patterns in the big data domain. The necessary post-processing approach for multiple locations and the approach to extract patterns for each single location use simple mathematic relations and do not make the framework complex. The framework was applied on the low voltage measurements, but it could equally be applied at medium/high voltage levels. This paper could provid general knowledge beyond a specific case study on the much less understood phenomena (sub-10-min rms voltage fast variations). The patterns can be used as a reference for the manufacturers to design the equipment. The obtained patterns are a compromise between storing large amounts of raw data with high time resolution, resulting in different big-data challenges, and completely neglecting the time scale. The authors realize that the data used for this work is one of the initial reviews of the time scale. For example, the 35 h variations (real and pattern-based), as shown in Fig. 19, could be studied for a longer period and in more locations. Future work must study the potential impacts on the equipment after quantifying the rapid voltage variation levels and patterns.

CRediT authorship contribution statement

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

Data availability

Data will be made available on request.

Acknowledgments

Most of the measurements presented in this paper were performed by a PQsmart portable monitor provided by Metrum Sweden AB. Also, part of this work was sponsored by the Swedish Energy Agency.

Appendix A. Single-window existing statistical indices

Table A1 explains the single-window indices [6,18] used in this paper. The indices quantify the range of 1-s values within a 10-min window. 1-s very short variations are a difference between 1-s rms voltages and the 10-min rms one. A 10-min rms voltage is the rms of 1-s values within the 10-min window.

Table A1
Existing single-window 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 Standard deviations	VSV Std.	10-min sliding-window rms on 1-s very short variations 10-min non-sliding-window rms on 1-s very short variations

References

[1] Bollen MHJ, Gu IYH. Signal Processing of Power Quality Disturbances. 2005. https://doi.org/10.1002/0471931314.

[2] Bollen M, Milanović J, Cukalevski N. CIGRE/CIRED JWG C4.112 Power Quality Monitoring. Renew Energy Power Qual J 2014:1037–45. https://doi.org/10.24084/repqj12.011.

[3] CEER. Ceer Benchmarking Report on the Quality of Electricity and Gas Supply-2016: Gas-Technical Operational Quality 2016:138–201.

[4] IEEE Recommended Practice–Adoption of IEC 61000-4-15:2010, Electromagnetic compatibility (EMC)–Testing and measurement techniques–Flickermeter–Functional and design specifications - Redline. IEEE Std 1453-2011 - Redline 2011: 1–89.

[5] IEEE Guide for Voltage Sag Indices. IEEE Std 1564-2014 2014:1–59. https://doi.org/10.1109/IEEESTD.2014.6842577.

[6] Gil-de-Castro A, Bollen MHJ, Rönnberg SK. Variations in harmonic voltage at the sub-10-minute time scale. Electr Power Syst Res 2021;195. https://doi.org/10.1016/j.epsr.2021.107163.

[7] Schlabbach J, Blume D, Stephanblome T. Voltage Quality in Electrical Power Systems. The Institution of Engineering and Technology, Michael Faraday House, Six Hills Way, Stevenage SG1 2AY, UK: IET; 2001. https://doi.org/10.1049/PBPO036E.

[8] Guide to Quality of Electrical Supply for Industrial Installations. 1999.

[9] Lennerhag O, Bollen M, Ackeby S, Rönnberg S. Very short variations in voltage (timescale less than 10 minutes) due to variations in wind and solar power. Int Conf Exhib Electr Distrib 15/06/2015 - 18/06/2015 2015.

[10] Ravindran V, Sakar S, Rönnberg S, Bollen MHJ. Characterization of the impact of PV and EV induced voltage variations on LED lamps in a low voltage installation. Electr Power Syst Res 2020;185:106352. https://doi.org/10.1016/j.epsr.2020.106352.

[11] Lodetti S, Bruna Romero J, Melero J. Methods for the Evaluation of New Power Quality Parameters: a Review of Rapid Voltage Changes and Supraharmonics; 2019.

[12] Bletterie B, Pfajfar T. Impact of Photovoltaic generation on voltage variations-how stochastic is PV. CIRED 19th Int Conf Electr 2007:21–4.

[13] Widén J, Carpman N, Castellucci V, Lingfors D, Olauson J, Remouit F, et al. Variability assessment and forecasting of renewables: a review for solar, wind, wave and tidal resources. Renew Sustain Energy Rev 2015;44:356–75.

[14] Liu G, Zhou J, Jia B, He F, Yang Y, Sun N. Advance short-term wind energy quality assessment based on instantaneous standard deviation and variogram of wind speed by a hybrid method. Appl Energy 2019;238:643–67. https://doi.org/10.1016/j.apenergy.2019.01.105.

[15] Shukla RM, Sengupta S, Patra AN. 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;2018-Janua:632–8. https://doi.org/10.1109/CCWC.2018.8301710.

[16] Seljeseth H, Taxt H, Solvang T. Measurements of network impact from electric vehicles during slow and fast charging. IET Conf Publ 2013;2013:10–3. https://doi.org/10.1049/cp.2013.1197.

[17] Macii D, Petri D. Rapid voltage change detection: limits of the IEC standard approach and possible solutions. IEEE Trans Instrum Meas 2020;69(2):382–92. https://doi.org/10.1109/TIM.2019.2903617.

[18] Bollen M, de Castro AG, Rönnberg S. Characterization methods and typical levels of variations in rms voltage at the time scale between 1 second and 10 minutes. Electr Power Syst Res 2020;184:106322. https://doi.org/10.1016/j.epsr.2020.106322.

[19] Miraftabzadeh SM, Longo M, Foiadelli F, Pasetti M, Igual R. Advances in the application of machine learning techniques for power system analytics: a survey†. Energies 2021;14(16):4776. https://doi.org/10.3390/en14164776.

[20] Miraftabzadeh SM, Foiadelli F, Longo M, Pasetti M. 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. https://doi.org/10.1109/EEEIC.2019.8783340.

[21] Axelberg PGV, Gu I-H, Bollen MHJ. Support vector machine for classification of voltage disturbances. IEEE Trans Power Deliv 2007;22(3):1297–303. https://doi.org/10.1109/TPWRD.2007.900065.

[22] Mohammadi Y, Moradi MH, Chouhy LR. A novel method for voltage-sag source location using a robust machine learning approach. Electr Power Syst Res 2017;145:122–36. https://doi.org/10.1016/j.epsr.2016.12.028.

[23] De Yong D, Bhowmik S, Magnago F. An effective power quality classifier using wavelet transform and support vector machines. Expert Syst Appl 2015;42(15-16):6075–81. https://doi.org/10.1016/j.eswa.2015.04.002.

[24] Mohammadi Y, Salarpour A, Chouhy Leborgne R. 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 2021;124:106363. https://doi.org/10.1016/j.ijepes.2020.106363.

[25] Valtierra-Rodriguez M, de Jesus Romero-Troncoso R, Osornio-Rios RA, Garcia-Perez A. Detection and classification of single and combined power quality disturbances using neural networks. IEEE Trans Ind Electron 2014;61(5):2473–82. https://doi.org/10.1109/TIE.2013.2272276.

[26] Mishra S, Bhende CN, Panigrahi BK. Detection and classification of power quality disturbances using S-transform and probabilistic neural network. IEEE Trans Power Deliv 2008;23(1):280–7. https://doi.org/10.1109/TPWRD.2007.911125.

[27] Cai K, Cao W, Aarniovuori L, Pang H, Lin Y, Li G. Classification of power quality disturbances using Wigner-Ville distribution and deep convolutional neural networks. IEEE Access 2019;7:119099–109. https://doi.org/10.1109/ACCESS.2019.2937193.

[28] Qiu W, Tang Q, Liu J, Yao W. An automatic identification framework for complex power quality disturbances based on multifusion convolutional neural network. IEEE Trans Ind Informatics 2020;16(5):3233–41. https://doi.org/10.1109/TII.2019.2920689.

[29] Räsänen T, Kolehmainen M. Feature-Based Clustering for Electricity Use Time Series Data. vol. 5495. 2009. https://doi.org/10.1007/978-3-642-04921-7_41.

[30] Fulcher B. Feature-based time-series analysis 2017.

[31] Gong C, Su Z-G, Wang P-H, You Y. Distributed evidential clustering toward time series with big data issue. Expert Syst Appl 2022;191:116279. https://doi.org/10.1016/j.eswa.2021.116279.

[32] Wen L, Zhou K, Yang S. A shape-based clustering method for pattern recognition of residential electricity consumption. J Clean Prod 2019;212:475–88. https://doi.org/10.1016/j.jclepro.2018.12.067.

[33] Izakian H, Pedrycz W, Jamal I. Fuzzy clustering of time series data using dynamic time warping distance. Eng Appl Artif Intell 2015;39:235–44. https://doi.org/10.1016/j.engappai.2014.12.015.

[34] Ruiz LGB, Pegalajar MC, Arcucci R, Molina-Solana M. A time-series clustering methodology for knowledge extraction in energy consumption data. Expert Syst Appl 2020;160:113731. https://doi.org/10.1016/j.eswa.2020.113731.

[35] Galvani S, Rezaeian Marjani S, Morsali J, Ahmadi Jirdehi M. A new approach for probabilistic harmonic load flow in distribution systems based on data clustering. Electr Power Syst Res 2019;176:105977. https://doi.org/10.1016/j.epsr.2019.105977.

[36] Jasiński M, Sikorski T, Borkowski K. 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 2019;166:52–60. https://doi.org/10.1016/j.epsr.2018.09.020.

[37] Ge C, de Oliveira RA, Gu I-H, Bollen MHJ. Deep feature clustering for seeking patterns in daily harmonic variations. IEEE Trans Instrum Meas 2021;70:1–10. https://doi.org/10.1109/TIM.2020.3016408.

[38] Ge C, Oliveira RAD, Gu IYH, Bollen MHJ. Unsupervised deep learning and analysis of harmonic variation patterns using big data from multiple locations. Electr Power Syst Res 2021;194:107042. https://doi.org/10.1016/j.epsr.2021.107042.

[39] de Oliveira RA, Ravindran V, Ronnberg SK, Bollen MHJ. Deep learning method with manual post-processing for identification of spectral patterns of waveform distortion in PV installations. IEEE Trans Smart Grid 2021;12(6):5444–56. https://doi.org/10.1109/TSG.2021.3107908.

[40] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst 1987;2:37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

[41] Halko N, Martinsson PG, Tropp JA. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev 2011;53(2):217–88. https://doi.org/10.1137/090771806.

[42] Feng M. Project 1 Report: Dimensionality Reduction n.d.:1–11.

[43] Schölkopf B, Smola A, Müller K-R. Kernel principal component analysis BT - Artificial Neural Networks — ICANN'97. In: Gerstner W, Germond A, Hasler M, Nicoud J-D, editors., Berlin, Heidelberg: Springer Berlin Heidelberg; 1997, p. 583–8.

[44] Martinsson P-G, Rokhlin V, Tygert M. A randomized algorithm for the decomposition of matrices. Appl Comput Harmon Anal 2011;30(1):47–68. https://doi.org/10.1016/j.acha.2010.02.003.

[45] Cao LJ, Chua KS, Chong WK, Lee HP, Gu QM. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing 2003;55:321–36. https://doi.org/10.1016/S0925-2312(03)00433-8.

[46] Sakthi M, Selvadoss TA. An effective determination of initial centroids in K-means clustering using kernel PCA 2011;2:955–9.

[47] Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding. In: Proc Annu ACM-SIAM Symp Discret Algorithms 2007;07-09-Janu:1027–35.

[48] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[49] Saputra DM, Saputra D, Oswari LD. Effect of distance metrics in determining K-Value in K-Means clustering using elbow and silhouette method 2020;172:341–6. https://doi.org/10.2991/aisr.k.200424.051.

[50] Mohammadi Y, Miraftabzadeh SM, Bollen MHJ, Longo M. An unsupervised learning schema for seeking patterns in rms voltage variations at the sub-10-minute time scale. Sustain Energy, Grids Networks 2022;31:100773. https://doi.org/10.1016/j.segan.2022.100773.

[51] Mohammadi Y, Leborgne RC. 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 2020;182. https://doi.org/10.1016/j.epsr.2020.106222.

[52] Mohammadi Y, Moradi MH, Chouhy LR. Employing instantaneous positive sequence symmetrical components for voltage sag source relative location. Electr Power Syst Res 2017;151:186–96. https://doi.org/10.1016/j.epsr.2017.05.030.

[53] Mohammadi Y, Leborgne RC. 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 2020;117:105664. https://doi.org/10.1016/j.ijepes.2019.105664.

[54] Mohammadi Y, Leborgne RC. Modified methods for voltage-sag source detection using transient periods 2022;207. https://doi.org/10.1016/j.epsr.2022.107857.

[55] Moradi MH, Mohammadi Y, Hoseyni TM. A novel method to locate the voltage sag source: a case study in the Brazilian power network (Mato Grosso). Prz Elektrotechniczny 2012;88.

[56] Lucasius CB, Dane AD, Kateman G. On k-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasiblity and comparison. Anal Chim Acta 1993;282:647–69. https://doi.org/10.1016/0003-2670(93)80130-D.

[57] Mohammad Y, Seyed Mahdi M, Bollen MHJ, Longo M. Voltage-sag source detection: developing supervised methods and proposing a new unsupervised learning. Sustain Energy Grids Netw 2022. https://doi.org/10.1016/j.segan.2022.100855. In press.

[58] Mohammadi Y, Bollen MHJ. Voltage sag source location methods' performance during transient and steady-state periods. 2022 20th International Conference on Harmonics & Quality of Power (ICHQP) 2022:1–6. https://doi.org/10.1109/ICHQP53011.2022.9808649. In this issue.

Update

International Journal of Electrical Power and Energy Systems

Volume 154, Issue , December 2023, Page

DOI: https://doi.org/10.1016/j.ijepes.2022.108582

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

Corrigendum

Corrigendum to “Seeking patterns in rms voltage variations at the sub-10-minute scale from multiple locations via unsupervised learning and patterns’ post-processing” [Int. J. Electr. Power Energy Syst. 143 (2022) 108516]

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

The authors regret, to inform you that some parts of the text have been unintentionally included from our previous draft during the preparation of the 2nd revision of the paper; hence, 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 delete some unnecessary sentences, modify some sentences, cite some missing references from the already existing references, and don’t 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 paper	Corrections
P1, C1&2, S1, Paragraph 1:	Is rephrased by: “There are two time-scales on the voltage deviations according to standards: more than some minutes and up to a few seconds. Slow variations are related to scales as minutes and more. A 10-min window was used to calculate the rms voltages as defined in IEC 61000–4-30 [1, 2], and used as the most common value shown in [3]. Fast variations take place at scales up to a few seconds as defined in IEC 61000–4-5 [4] and IEEE 1564 [5].”
P1, C2, S1, Paragraph 2, L7-10, “~~Moreover, tripping of…this time scale [7–9]…~~”:	Is replaced by: “Moreover, some stated undesirable outcomes of fast voltage variations, besides tripping PVs and light flicker, come from variations in this time scale [7–9]”.
P2, C1, S1.1, Paragraph 1, L1&2, …“~~individual rapid voltage changes (voltage steps) as~~”…:	Is replaced by “voltage steps, as only one of the sub-10-min variations’ characteristics)”
P2, C1, S1.1, Paragraph 1, L4&5, … “~~However, voltage steps… variations~~”…:	Is deleted.
P2, C1, S1.1, Paragraph 1, L11&12, … “~~but do not result…time window. A~~”…:	Is replaced by “, while a”
P2, C1, S1.1, Paragraph 2:	Is rephrased as: “Machine learning methods as supervised and unsupervised learning [19,20], could extract such patterns. The supervised methods used classifiers like support vector machines [21–23], ensemble learnings [24,57], and neural networks [25,26]. Automatic feature extraction was done in the input of the supervised classifiers [27,28] which showed a better role than the manually extracted ones [1,29,30].”
P2, C1, S1.1, Paragraph 3:	Is rephrased as: “Time series clustering in unsupervised problems was done to extract patterns from signals. The works on big data in [31], clustering by different methods than k-means, and the Euclidean distance as shape-based [32], and fuzzy-based by using Distance Time Wrapping (DTW) distance [33]. However, a few applications in power quality studies have been done; clustering to extract knowledge in energy consumption data [34], data clustering to evaluate harmonic load flow [35], and a k-means method to find out the contribution of distributed generations [36]. A deep autoencoder along with a k-means clustering extracted daily voltage harmonic patterns, 10-min measurement resolution, from one location [37] and ten locations [38]. Since, [37, 38] are

(continued on next column)

(continued)

(continued on next page)

DOI of original article: https://doi.org/10.1016/j.ijepes.2022.108516.

* Corresponding author.
E-mail address: Younes.mohammadi@ltu.se (Y. Mohammadi).

https://doi.org/10.1016/j.ijepes.2022.108582

Available online 29 September 2022

(*continued*)

Location in the PDF of initial paper	Corrections
	concerned with a well-understood phenomenon (daily variation in harmonic voltage), it might be said that their method did not create new general knowledge. The same authors of [37, 38] used a post-processing method using both harmonic and inter-harmonic data [39]. Among the few unsupervised learning schemas applied to power quality data, none of them have been yet applied to seek patterns for rms voltage fast variations (for multiple locations), and for harmonic voltage variations (for one/multiple locations) in 10-min scales which is a different phenomenon from daily variational patterns. Moreover, no framework for time-limited (about one day and a few hours) measurements from multiple locations is designed.”.
P2, C2, S1.2, The applications of the proposed framework are: Part (a):	“Similar to our previous work [50]” is added to the beginning of part (a) and [37-39] are replaced by [37,38].
P3, C2, S2, Paragraph 1, L2-8, …“The upper …a few second.”…:	Is replaced by “The selection of 10-min and 1-s values are explained in [6, 18].”.
P3, C2, S2.1, beginning of paragraph 1:	“Similar to our previous work [50]” is added.
P4, C1, S2.1, Paragraph 1, L5: x is misprinted as ×:	× is replaced by “x”.
P4, C2, S2.3, Paragraph 1, L5-7, …“~~Each feature vector … vectors are assigned~~”…:	Is replaced by “The k-means as an unsupervised learning is widely used in literature like [37, 38, 47]”.
P4, C2, S2.3, Paragraph 1, L9, …“~~following steps~~”…and last paragraph, “~~where the…*j*th cluster~~”:	are deleted.
P4, C2, S2.4, Paragraph 1, L1-3, …“~~To further analyze the …are fed~~”:	Is replaced as: “The features vectors from centroids are inputted”
P4, C2, S2.4, end of the section:	“Since transformation back to the original sub-space by any reconstruction method is associated with a reconstruction error of modelling, this study ignored (8) and (9) as follows: by having the labels within clusters in the output of k-means, an average was made on the samples in the original input space so that no reconstruction error was included on the *K* initial centroids.” is added.
P5, C2, S2.6, Paragraph 1, L4-8, …“~~In a *t*-SNE…(13)~~”:	Is deleted.
P5, C2, S3, Paragraph 1, L6, after …“for the measurements”:	“The dataset is based on the dataset presented in [18]” is added.
P7, C1, S4.1, Paragraph 1, L7, after …“chosen;”:	“like our previous work [50], ” is added.
P7, C2, S4.1, Paragraph 2, L1-7, …“A function…t-SNE 100 times”:	Is replaced by …“A 2D t-SNE with the same setting used in [50] is employed and the visualization results are shown in Fig. 5.”
P9, C1, S4.1, Last paragraph, L7, …“beside”…:	Is replaced by “besides”.
P9, C2, S4.3.1, L4, after…“as follows”…:	“(a summary of work done by our 3rd co-author in [18])” is added.
P10, C2, S4.3.1, Paragraph 1, L18-22, after …“By taking…indices).”:	Is replaced by …“The most proper indices are selected as R90, P95, and P5 [50].”
P12, C1, S4.3.3, L11, Beginning of line:	“As a confirmation of findings in [50], ” is added.
P12, C2, S5, Paragraph 1, L1: A misprinted sign as “\” at the beginning line during the proof stage:	Is deleted.
P16, C1, S6.1, Paragraph 1:	Is replaced by: “In comparison with our previous work [50], this work extracts the sub-10-min patterns using a learning schema attached to a post-processing approach. So, some general and typical low voltage patterns from multiple locations were indicated. The schema is scalable, computationally cheap, and is the first step to start potential impacts of the variations on equipment, next to [6, 50].”
P17, C2, S6.3.1, L1-4, “The results…(only k-means),”…:	Is replaced by: “The clustering results with using only k-means were also investigated and shown in Fig. 25 in a 2D plot for 15 initial clusters. No clear separation is seen and seems choosing 3 clusters is enough,”
P18, C1, S6.3.3, Paragraph 1:	Is replaced by “As shown earlier and in [50], KPCA with the Cosine kernel was chosen as a simple tool to reduce the feature size from 600 to 10. However, other tools such as a deep autoencoder (DAE) might be useful instead of KPCA for high-resolution time series (i.e., 600D). Nevertheless, using a DAE would not be as simple as the used KPCA. Table 4 shows a simple structured-based comparison of our framework (KPCA + k-means + post-processing) and (DAE + k-means) [37]. The findings, as can be seen in Table 4, confirm the simpler schema of our schema.”
P18, C1, S6.3.3, Paragraph 2, L1-4:	Is replaced by “A simple running time comparison between our proposed schema and [37] is shown in Table 5 (although they are aimed to be used as off-line).”
P20, C1, S Acknowledgments:	“The authors would like to acknowledge Aurora Gil de Castro [6, 18], for statistic studies in the sub-10-min time scale and Chenjie Ge [37], for applying machine learning in other scales.” is added at the beginning of the paragraph.
P21, Appendix A, Paragraph 1:	Is shortened as “Table A1 summaries the existing single-window indices introduced in [6, 18] and used in this paper.”

(*continued on next column*)

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