Electrical Power and Energy Systems 125 (2021) 106471

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

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

Monitoring of low voltage grids with multilayer principal component analysis

L. Souto*, J. Meléndez, S. Herraiz

Control Engineering and Intelligent Systems Group of the University of Girona, Girona 17003, Spain

ARTICLE INFO

Keywords:
Monitoring
Power distribution
Power system measurements
Principal component analysis
Statistical learning
Wide area measurements

ABSTRACT

This article presents a monitoring strategy based on multilayer principal component analysis (PCA) to detect and diagnose power system disturbances in large amounts of data collected by intelligent electronic devices in low voltage smart grids. The PCA models are built on multiple sliding windows, sized (in terms of length and sampling time) according to the type of phenomena to detect. Abnormalities are detected with use of two complementary statistical indexes, then diagnosed by computing the individual contributions of each monitored variable to the constraint violation of those statistics. As a result, its implementation enables an automatic analysis of multiple phenomena of interest in parallel over time using distinct electrical quantities. Furthermore, the method is demonstrated within the RESOLVD project with data from the OpenLV project containing measurements of active and reactive power gathered at different low voltage distribution substations.

1. Introduction

Power systems operation shall comply with a series of principles aimed at ensuring a satisfactory level of quality of supply and efficient allocation of infrastructure and resources while respecting constraints related to operational security [1,2]. However, this is becoming an issue of great complexity due to the increasing demand for electricity along with the integration of distributed renewable generation and new energy appliances, particularly at low voltage (LV) distribution level.

In this scenario, the application of digital technology for real-time observability of LV networks is still impractical due to their inner complexity – radial topology, heterogeneous lines, high spatial density of customers, and unbalanced phases. Even when phasor measurement units (PMU) and/or smart meters (SM) are deployed, either as dedicated devices or as built-in capabilities of intelligent electronic devices (IED), it is not straightforward to search for relevant information about abnormal operating conditions [3,4] in huge amounts of data with a very high temporal resolution (in the case of PMUs) and spatial density (in the case of SMs). This is particularly challenging when events of very different durations are considered, as a multiple time-scale resolution is necessary to properly identify distinct power system phenomena, as illustrated in Fig. 1. Therefore, a strategy capable of detecting and diagnosing generic power system disturbances and abnormal behavioral patterns in a standard, coherent, coordinate way at multiple timescales is necessary as a first step to ensure reliable operation of LV smart grids.

Current solutions for LV grid monitoring include networks of SMs installed at customer level and/or at the secondary substation together with a supervisory control and data acquisition (SCADA) system for power quality monitoring and state estimation [5–8]; distributed sensors for fault detection and location [9]; condition monitoring in underground LV cables [10]; devices with high sampling frequency for identification of power quality disturbances [11]; and an advanced monitoring system based on Geographic Information System (GIS) [12]. Among the solutions relying on data processing for feature extraction and classification, [11] calculates electrical characteristics of voltage and current signals and compares them with predefined thresholds. In [8], a deep learning strategy trained with synthetic power quality disturbances is embedded in SMs such that only information about detected disturbances is sent to the utilities, which reduces the flow (and enables exchange) of data between utilities and customers.

In comparison to qualitative knowledge-based methods, such as [8], statistical knowledge-based methods are computationally less expensive for training and testing. As a matter of a fact, the usage of statistical indicators for anomaly detection simplifies both training and testing and comes with a confidence level that ensures good performance of the method [13]. In this context, the usage of dimensionality reduction techniques is promising to overcome the limitations

* Corresponding author.
E-mail address: laiz.souto@gmail.com (L. Souto).
URL: http://exit.udg.edu (L. Souto).

https://doi.org/10.1016/j.ijepes.2020.106471
Received 10 January 2020; Received in revised form 15 June 2020; Accepted 21 August 2020

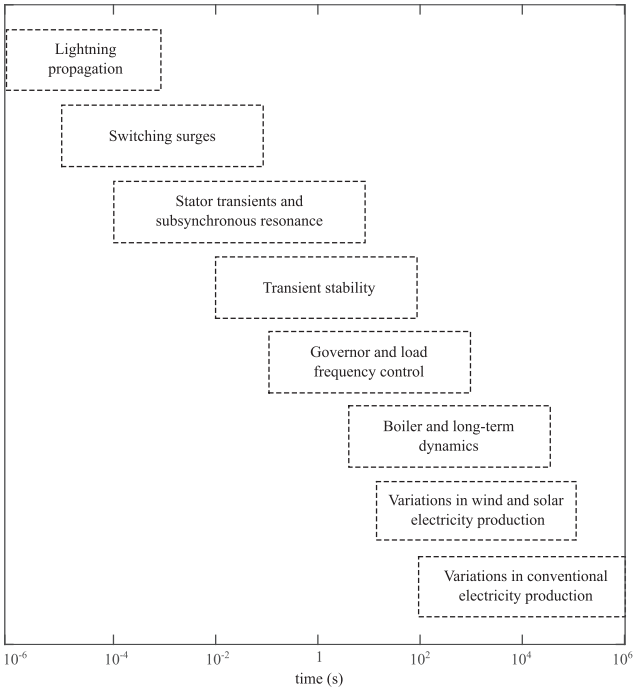

Fig. 1. Time ranges of distinct phenomena in power systems (adapted from [2]).

concerning exchange of data that still make real-time observability of LV networks unfeasible while allowing for an automatic identification of patterns and outliers in short-, medium-, and long-term operation. Among them, principal component analysis (PCA) is the most popular and particularly convenient to build a statistical model of the data for detection of outliers, as it represents the correlation structure by a few linear combinations of the original variables that depict the main trends in the data set. In recent years, some PCA-based strategies to identify different types of abnormal behaviors in power system operation have been presented in the literature. Notably, [14] used two statistical tests to identify the variables involved in generic power system disturbances and their magnitudes. Later, [15] focused on islanding detection with use of the same statistics, which was extended in [16] to distinguish islanding events from system-wide disturbances in power systems with high penetration of distributed generation and further in [17] to detect and classify islanding, loss-of-load, and loss-of generation. In [17], a moving window approach was applied, as in [18], to allow for continuous monitoring with improved situational awareness.

Notably, dimensionality reduction techniques are relevant for electric power systems with a large number of electrical quantities gathered over time and centralized monitoring and control, as the number of variables to be observed may be reduced dramatically. In this case, the same number of measured quantities is required for modeling and practical purposes, but not for monitoring, as the original data are projected onto a lower dimensional subspace. To this extent, modeling refers to the preparation phase, whereas monitoring refers to the actual usage with live data. As a result, dimensionality reduction may simplify the monitoring tasks so that relevant events of interest can be identified over different lengths of time. This is usually the case of LV smart grids, but applications of dimensionality reduction techniques in this domain have not been explored yet.

Fitting into this context, this work extends previous methods devoted to detection and diagnosis of multiple power system disturbances over time with an automatic multilayer PCA methodology focused on the monitoring of LV smart grids. The methodology consists of an adaptation of traditional PCA deployed at multiple timescale resolutions for an effective monitoring of LV smart grids with dimensionality reduction. Additionally, the methodology is demonstrated in a real LV power distribution network with data from the OpenLV project [19] recorded at Marshfield Village, South Gloucestershire, United Kingdom, where single-phase measurements of active and reactive power were gathered every minute by 4 IEDs with integrated PMU and SM capabilities from July 17, 2018 to June 30, 2019.

2. Methodology

This section presents the multilayer PCA strategy applied for modeling and monitoring the data over different timescales. It is divided in three main steps presented in Sections 2.1,2.2,2.3: problem setting, PCA-based modeling and evaluation, and sliding-window PCA.

The overall PCA formulation is based on [13], but focused on the study of power system disturbances. Each problem of interest requests different electrical quantities and sampling rates and is arranged hierarchically in accordance to its timescale. Input data may include voltage, current, impedance, power, or energy measurements acquired by any monitoring infrastructures or calculated from combinations of them.

The PCA models are built over a sliding time window to enable periodic updates reflecting changes in the generation and consumption patterns. The choice of an adequate length of time considers the trade off between the expected duration of the events and the elapsed time between consecutive events of interest.

The PCA models have the ability of separating correlated information (main features) from uncorrelated information (noise) into two orthogonal hyperplanes representing linear combinations of the observed variables. For a set of variables representing the same electrical quantity in LV grids, these linear relations are assumed to suffice to represent the steady-state operation of the system and enable detection of abnormal operating conditions at different layers.

Event detection relies on two complementary statistics: Hotelling's T^2, which measures the Mahalanobis distance of the projected data to the center of the model, and the square prediction error (SPE), which measures the square distance of the observation to the projection subspace. Once detected, abnormalities are isolated and diagnosed by analyzing the individual contributions of each measured variable to the T^2 and/or SPE statistics and selecting the greatest contributors to the violation of constraints posed by the statistical thresholds.

2.1. Problem setting

The goal of this step is to prepare the data in a matrix structure suitable for PCA.

2.1.1. Premises

Given $m_{raw} \in$ * original variables (electrical quantities) gathered by one or multiple IEDs over time at the sampling rate $f_{raw} \in$ *+ in Hertz, define $\ell \in$ * hierarchical levels (layers) in such a way that the events of interest can be correctly identified and characterized over different lengths of time (e.g., second, minute, hour, etc.). In accordance to the System Average Interruption Duration Index (SAIDI) values in the European Union [20], this analysis considers that abnormal phenomena should last no longer than 5% of the total duration of the analysis. Each layer is characterized by a length of time $\tau_k \in$ *+ (total time which the monitoring lasts), an observation period $\sigma_k \in$ *+ (if applicable), an observation time duration $\theta_k \in$ *+ (with $\theta_k \leqslant \sigma_k$), and contains $n_k \in$ * observations defined by $p_k \in$ * samples of $m_k \in$ * variables gathered at the sampling rate f_k. Thereby, a single observation in the kth layer is a $p_k \times m_k$-dimensional array gathered in discrete time domain, hereby denoted by $\mathbf{x_k}(i)$, $i = \{1, \cdots, n_k\}$, forming an $n_k \times (p_k \times m_k)$ observation matrix $\mathbf{X}_k$ such that (1) and (2) hold. Note that a particular case occurs when $\theta_k = \sigma_k = f_k^{-1}$, as the observations $\mathbf{x_k}(i)$ last a single sample and $\mathbf{X}_k$ is thereby an $n_k \times m_k$ observation matrix. Additionally, consider that $n_{k,\sigma}, n_{k,\partial} \in$ * observations are

gathered over σ_k and θ_k respectively.

$$n_k = \frac{\tau_k}{\sigma_k} \tag{1}$$

$$p_k = \theta_k f_k \tag{2}$$

It is noteworthy that the definition of θ_k allows for investigation of repetitive patterns lasting θ_k over time (e.g., daily and weekly energy consumption profiles). Thus, τ_k, σ_k, θ_k, f_k, and m_k are defined as design parameters for $k = 1, \cdots, \ell$ depending on the data organization required. In addition, assume that the layers are concatenated hierarchically over time such that $\tau_1 \leqslant \tau_2 \leqslant \cdots \leqslant \tau_\ell$, with $\tau_{k+1} = j \times \tau_k$, $k = \{1, \cdots, \ell - 1\}$, for some $j \in$ * Also, consider that screening is required every τ_k intervals, as long as abnormal behaviors are detected at the layers of shortest duration.

2.1.2. Data organization

The data gathered by the IEDs might require a previous preprocessing step to suit to the analysis. To this extent, three different techniques can be applied and combined as a previous step to the PCA modeling and monitoring, depending on the relation between f_{raw} and τ_k, σ_k, and θ_k at the k^{th} layer: (1) time windowing, (2) filtering and re-sampling, and (3) multiway re-arrangement. For illustration, Fig. 2 draws a comparison between them over the same τ_k, with different values of $\sigma_k = f_k^{-1}$ and θ_k chosen arbitrarily for each procedure.

2.1.2.1. Time windowing. This procedure, illustrated in Fig. 2, consists of defining adequate time settings τ_k, σ_k, and θ_k for each layer $k = \{1, \cdots, \ell\}$. This choice is arbitrary and depends on the phenomena under evaluation, as shown in Section 3, but made automatically at the beginning of the analysis according to relevant timescales for humans. In this context, the following situations may occur: continuous observations ($\sigma_k = \theta_k$) and gapped observations ($\sigma_k > \theta_k$) over τ_k.

2.1.2.2. Filtering and re-sampling. This procedure is recommended when re-sampling is required for the analysis every σ_k time instants to reduce the number of observations over τ_k, which is achieved by smoothing and re-sampling the original observations, with low-pass filtering required to avoid aliasing. It is exemplified in Fig. 2 (2), where the dark blue dots represent the re-sampled observations over the original discrete signal in light blue every $\sigma_k = f_k^{-1}$ time instants over τ_k.

2.1.2.3. Multiway re-arrangement. This procedure is recommended when repetitive patterns lasting θ_k are expected to occur (e.g., daily, weekly) during specific periodic intervals defined by σ_k, as it enables to exploit possible correlations between time instants within θ_k. In this case, $p_k > 1$ samples acquired over θ_k (between $(i-1)\sigma_k$ and $\theta_k + (i-1)\sigma_k$) are concatenated to form a single observation $\mathbf{x_k}(i)$, $i = \{1, \cdots, n_k\}$. As a result, $\mathbf{x_k}(i)$ is a $p_k \times m_k$ matrix with as many rows as samples over θ_k and as many columns as variables, as illustrated in Fig. 2 (3).

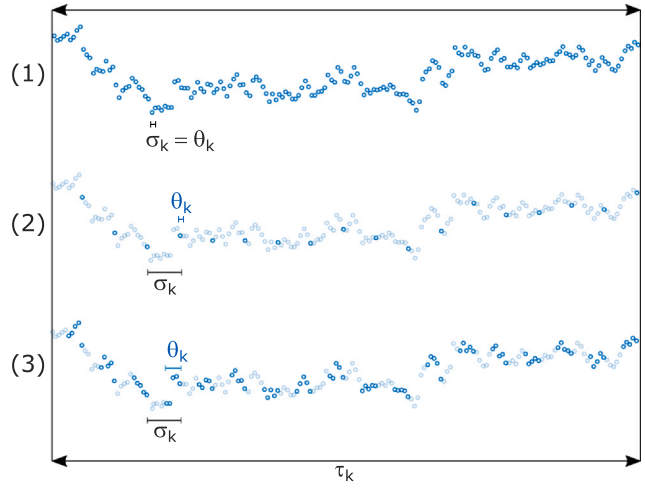

Fig. 2. Comparison between distinct data organization procedures over τ_k: (1) time windowing with original measurements: $x_k(i)$ single observations, $i = \{1, \cdots, n_k\}$, (2) re-sampling: $x_k(i\sigma_k)$ single observations, $i = \{1, \cdots, n_k\}$, and (3) multiway re-arrangement: a single observation $x_k(i)$ obtained between $(i-1)\sigma_k$ and $\theta_k + (i-1)\sigma_k$, $i = \{1, \cdots, n_k\}$.

2.2. PCA-based modeling and monitoring

The PCA algorithm presented in this article is executed at layer level over time, as long as new data are available, and can be divided in the following processes: building of the statistical model, detection of abnormalities in the projection subspace and residual subspace, and isolation and diagnosis of abnormalities in the projection subspace and residual subspace, as explained in Sections 2.2.1,2.2.2,2.2.3. It is noteworthy that the statistical models are trained and tested with the same measurement sets to improve the situational awareness of the model, insofar as the operating conditions of the grid are time-varying [17].

2.2.1. Building the PCA model

This section describes the PCA algorithm applied to train and test the statistical model of the data at each layer $k = 1, \cdots, \ell$. It is noteworthy that the statistical models are trained and tested with the same measurement sets to improve the situational awareness of the model, insofar as the operating conditions of the grid are time-varying and might not be represented appropriately in a static reference model [17,18].

First, let $\mathbf{X}_k$ be an $n_k \times (p_k \times m_k)$ observation matrix (assumed to be centered and scaled in the PCA algorithm) of all selected IEDs at the k^{th} layer with n_k observations and $(p_k \times m_k)$ sampled variables referred to distinct electrical quantities gathered at f_k such that (1) and (2) hold over τ_k.

Next, two matrices $\mathbf{V}_k$ and $\mathbf{\Lambda}_k$ $(p_k \times m_k) \times (p_k \times m_k)$ are obtained by computing the covariance matrix $\mathbf{S}_k$ from $\mathbf{X}_k$ in (3) and applying eigenvalue decomposition. Columns in $\mathbf{V}_k$ contain the principal components, which represent orthonormal vectors whose directions express the major variability of the data and the relative weights (or loadings) of the original variables, whereas $\mathbf{\Lambda}_k$ is a diagonal matrix whose elements $\lambda_{i,k}$, $i = 1, \cdots, (p_k \times m_k)$ express variability in the direction of each principal component.

$$\mathbf{S}_k = \frac{1}{n_k - 1}\mathbf{X}_k^T\mathbf{X}_k = \mathbf{V}_k\mathbf{\Lambda}_k\mathbf{V}_k^T \tag{3}$$

Dimensionality reduction in the number of variables can be performed by retaining the r_k principal components ($r_k < (p_k \times m_k)$) with the r_k largest eigenvalues. As a result, the $(p_k \times m_k) \times (p_k \times m_k)$ matrix $\mathbf{V}_k$ becomes an $(p_k \times m_k) \times r_k$ matrix $\mathbf{P}_k$ which defines a projection space of lower dimension represented by the r_k most important components. In this article, the Variance Reconstruction Error (VRE) criterion (further described in [21]) is applied to define an appropriate value of r_k, as this ensures the best reconstruction of the variable.

Transformation of $\mathbf{X}_k$ into the principal components representation space can be realized without loss of information by multiplying it by $\mathbf{V}_k$. However, using $\mathbf{P}_k$ instead of $\mathbf{V}_k$, $\mathbf{X}_k$ is projected onto a space of lower dimension in which some information contained in the original data is lost, as shown in (4), where $\mathbf{T}_k$ is the transformation score matrix.

$$\mathbf{T}_k = \mathbf{X}_k\mathbf{P}_k \tag{4}$$

Since $\mathbf{V}_k$ is a unitary matrix, there holds that $\mathbf{V}_k\mathbf{V}_k^T = \mathbf{I}$ and the inverse operation is carried out with the transpose matrix (^{T} operator), but $\mathbf{P}_k\mathbf{P}_k^T \neq \mathbf{I}$, since $\mathbf{P}_k$ is not a unitary matrix. As a result, transformation of $\mathbf{T}_k$ into $\mathbf{X}_k$ with $\mathbf{P}_k$ does not produce equivalent results and is

represented by $\widehat{\mathbf{X}}_k$ in (5).

$$\widehat{\mathbf{X}}_k = \mathbf{T}_k \mathbf{P}_k^T \tag{5}$$

The difference between $\mathbf{X}_k$ and $\widehat{\mathbf{X}}_k$ is the residual matrix $\widetilde{\mathbf{X}}_k$ which resumes the information contained in the $m_k - r_k$ components from the residual space for each observation. Thereby, the complete PCA model at the k^{th} layer can be described as in (6).

$$\mathbf{X}_k = \widehat{\mathbf{X}}_k + \widetilde{\mathbf{X}}_k \tag{6}$$

2.2.2. Detection of abnormalities

Detection of abnormal operating conditions in the projection subspace is realized with T^2 statistics. First, it computes a weighted distance of the projected data to the center of the model with (7), where $t_{i,k}$ denotes the score component of a single observation $\mathbf{x_k}$ of $\mathbf{X}_k$ calculated with the i^{th} principal component, $i = \{1, \cdots, r_k\}$.

$$T_x^2 = \sum_{i=1}^{r_k} \frac{t_{i,k}^2}{\lambda_{i,k}} \tag{7}$$

Then, the statistical threshold T_{thresh}^2 is calculated analytically with (8):

$$T_{thresh}^2 = \frac{r_k(n_k^2 - 1)}{n_k(n_k - r_k)} F_\alpha(r_k, n_k - r_k) \tag{8}$$

where α is the confidence level obtained from the χ^2 distribution for r_k degrees of freedom and $F_\alpha(r_k, n_k - r_k)$ is the critical point of the Fischer-Snedecor distribution for r_k and $n_k - r_k$ degrees of freedom. Finally, any projection $\hat{\mathbf{x}}_k$ that surpasses T_{thresh}^2 at the k^{th} layer is tagged as faulty (or abnormal) according to the T^2 statistics.

In turn, detection of abnormal operating conditions in the residual subspace is realized with SPE statistics. First, it computes the variation out of the projection space defined by the r_k principal components through the error component $\tilde{\mathbf{x}}_k$ of $\widetilde{\mathbf{X}}_k$ with (9).

$$SPE_x = (\mathbf{x}_k - \hat{\mathbf{x}}_k)(\mathbf{x}_k - \hat{\mathbf{x}}_k)^T = \|\tilde{\mathbf{x}}_k\|^2 \tag{9}$$

Then, the statistical threshold SPE_{thresh} is calculated analytically with (10):

$$Q_{lim} = z_1 \left[1 + \frac{h_0 c_\alpha \sqrt{2z_2}}{z_1} + \frac{h_0 z_2 (h_0 - 1)}{z_1^2} \right]^{\frac{1}{h_0}} \tag{10}$$

with

$$z_j = \sum_{i=r_k+1}^{(p_k \times m_k)} \lambda_i^j, \; j = \{1, 2, 3\} \text{ and } h_0 = 1 - \frac{2z_1 z_3}{3z_2^2} \tag{11}$$

where c_α is the normal deviation for $(1 - \alpha)$. Finally, any observation $\mathbf{x}_k$ whose residual $\tilde{\mathbf{x}}_k$ surpasses SPE_{thresh} is tagged as faulty (or abnormal) according to the SPE statistics.

2.2.3. Isolation and diagnosis of abnormalities

Isolation of abnormal operating conditions is performed with contribution analysis of T^2 statistics in the projection subspace and SPE statistics in the residual subspace. Thus, it is necessary to compute the influence of each variable $x_{j,k}, j = \{1, \cdots, (p_k \times m_k)\}$ of $\mathbf{x}_k$ in the calculated values of T_x^2 and SPE_x that exceed T_{lim}^2 and SPE_{thresh}, respectively, to identify those responsible for the abnormal behavior, as well as the individual thresholds of each variable with T^2 and SPE statistics.

Considering each score $t_{i,k}$ in (7) as the contribution of the original variables weighted by the corresponding components of the i^{th} principal component, the total contribution of $x_{j,k}$ to T_x^2, $contr_{T_x^2}(x_{j,k})$, is given by the sum of its individual contributions to $t_{i,k}$, $i = \{1, \cdots, r_k\}$ in (12), whereas the individual thresholds $thre_{T_x^2}(x_{j,k})$ are given by the average plus three times the standard deviation of the calculated values of $contr_{T_x^2}(x_{j,k})$ over τ_k.

$$contr_{T_x^2}(x_{j,k}) = \sum_{i=1}^{r_k} contr(t_{i,k}, x_{j,k}) = \sum_{i=1}^{r_k} \frac{t_{i,k} x_{j,k} P_{j,i,k}}{\lambda_{i,k}} \tag{12}$$

If $thre_{T_x^2}(x_{j,k})$ is surpassed with T^2 statistics, i.e., $contr_{T_x^2}(x_{j,k}) > thre_{T_x^2}(x_{j,k})$, then $x_{j,k}$ is identified as a probable cause of the abnormal behavior detected with T^2 statistics.

In turn, from (9), the contribution of $x_{j,k}$ to SPE_x, $contr_{SPE_x}(x_{j,k})$, is given by (13), whereas the individual thresholds $thre_{SPE_x}(x_{j,k})$ are given by the average plus three times the standard deviation of the calculated values of $contr_{SPE_x}(x_{j,k})$ over τ_k.

$$contr_{SPE_x}(x_{j,k}) = (x_{j,k} - \hat{x}_{j,k})^2 = \tilde{x}_{j,k}^2 \tag{13}$$

If the thresholds calculated analytically for each variable $x_{j,k}$ are surpassed, i.e., $contr_{SPE_x}(x_{j,k}) > thre_{SPE_x}(x_{j,k})$, then $x_{j,k}$ is identified as a probable cause of the abnormal behavior detected with SPE statistics. Eventually, diagnosis of abnormal operating conditions is realized with knowledge of the network topology and energy appliances covered by each IED.

2.3. Sliding-window PCA

2.3.1. Description and flowchart

The introduction of a sliding window framework implies that the PCA monitoring described in the previous section is run on the fly every τ_k time instants. This approach increases the situational awareness of the analysis, as it allows for simultaneous modeling and monitoring over τ_k. Fig. 3 provides an overview of the PCA-based modeling and monitoring method presented in Section 2.2 for a generic layer k. This procedure shall be executed for $k = \{1, \cdots, \ell\}$. Assuming that the data are preprocessed and ready for evaluation, the algorithm starts with the creation of a time window of duration τ_k at the k^{th} layer, once there are enough observations acquired over τ_k. Then, an $n_k \times (p_k \times m_k)$ observation matrix is arranged and the PCA model is computed over τ_k. If an abnormal observation is detected, the algorithm proceeds with diagnosis and continues to evaluate related occurrences up to the layer $k - 1$ and the analysis proceeds with the creation of a new time window τ_k in the k^{th} layer, as long as new data are received.

2.3.2. Example

This subsection provides an example of sliding-window PCA within a multilayer implementation. The PCA is aimed at detecting hourly and daily variations in power consumption patterns. Input data consist of single-phase active power measurements gathered every minute (i.e., $f_{raw} = 1\ \text{min}^{-1}$) by a single IED (i.e., $m_{raw} = 3$ in a three-phase system). For illustration, a single-phase active power profile is plotted in Fig. 4 over distinct lengths of time (solid yellow lines). In the next paragraphs, consider that the layers $k = \{1, 2\}$ refer to hourly and daily variations in power consumption patterns displayed in the bottom and top graphs of Fig. 4, respectively.

In this scenario, consider the introduction of two time windows $\tau_1 = 1$ day and $\tau_2 = 91$ days enabling a PCA-based monitoring every day and over a season, respectively. As a result, in a season, τ_1 would be run 91 times on a sliding-time basis and τ_2, just once.

Without data organization, however, anomalies can only be detected at specific minutes, as $f_{raw} = 1\ \text{min}^{-1}$. This choice results in $n_2 = 131,040$ observations (solid yellow line at the top graph of Fig. 4) and $n_1 = 24 \times 60 = 1440$ observations (solid yellow line at the bottom graph of Fig. 4). To prevent this, observations can be filtered and resampled (solid blue line in the bottom graph of Fig. 4)) and/or re-arranged multiway (pink boxes in the top and middle graphs of Fig. 4). In $k = 1$, re-sampling of average values over an hour with $f_1 = 1\ \text{h}^{-1}$ and $\sigma_1 = \theta_1 = 1$ h results in $p_1 = 1$ from (2) and $n_1 = \frac{24}{1} = 24$ observations from (1). In $k = 2$, multiway re-arrangement to evaluate a specific day of the week with $\sigma_2 = 7$ days, $\theta_2 = 1$ day, and $f_2 = 1\ \text{h}^{-1}$ results in $p_2 = 24 \times 1 = 24$ from (2) and $n_2 = \frac{91}{7} = 13$ from (1). As a result, $\mathbf{X}_1$ is a

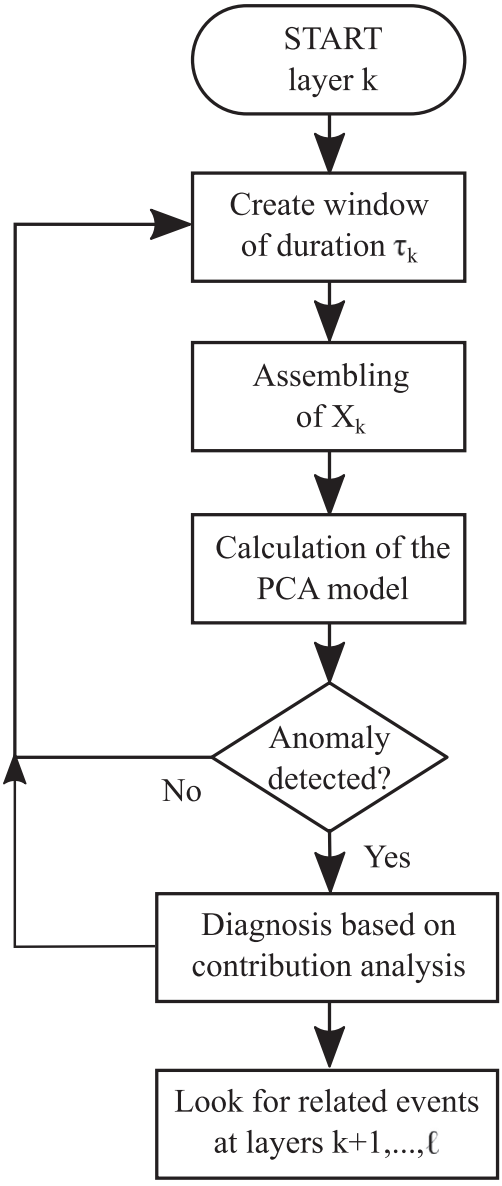

Fig. 3. Flowchart of the strategy in a generic layer k.

24×3 matrix, whereas $\mathbf{X}_2$ is a $13 \times (24 \times 3)$ matrix.

Additionally, in a multilayer implementation of PCA-based modeling and monitoring, τ_1 and τ_2 can be concatenated such that the anomalies detected with PCA-based monitoring over the shortest lengths of time (τ_1) can be related to those detected over the longest lengths of time (τ_2). This is performed as a bottom-up procedure, as schematized in Fig. 3, from $k+1$ to ℓ.

2.3.3. Example of combined analysis

A practical example is provided in Fig. 4 for a single-phase power profile. Consider that the layers $k = \{1, 2, 3\}$ refer to hourly, weekly, and seasonal variations in power consumption patterns displayed in the bottom, middle, and top graphs, respectively with $m_{raw} = 1$ and $f_{raw} = 1$ min^{-1}. The goal of identifying daily profile changes over the season (with $\tau_3 = 91$ days for exact integer calculations) with $f_3 = f_{raw}$ and $\theta_3 = \sigma_3 = 1$ min results in $n_3 = 131,040$ observations (solid yellow line at the top graph of Fig. 4), which would not be adequate for the task. As an option, the observations could be re-arranged multiway to evaluate the data from specific weekdays (e.g pink box in the second graph of Fig. 4) without re-sampling, which provides $f_2' = 1$ min^{-1}, $\theta_2 = 1$ day, $n_2' = 24 \times 60 = 1440$, $m_2' = m_{raw} = 1$, $m_2 = 1 \times 1440 = 1440$, $\tau_2 = 91$ days, $\sigma_2 = 7$ days, $f_2 = \frac{1}{7}$ day^{-1}, and $n_2 = \frac{91}{7} = 13$, or averaged every hour over a day with re-sapmpling for a joint evaluation of the data gathered at specific hours of the day (e.g., solid blue line at the bottom graph of Fig. 4) with $\tau_1 = 1$ day, which provides $\sigma_1 = 60$ min, $f_1 = \frac{1}{60}$ min^{-1}, $\theta_1 = 60$ min, and $n_1 = \frac{1440}{60} = 24$. As a result, $\mathbf{X}_1$ is a $131,040 \times 1$ matrix, $\mathbf{X}_2$ is a 1440×13 matrix, and $\mathbf{X}_3$ is a 24×1 matrix.

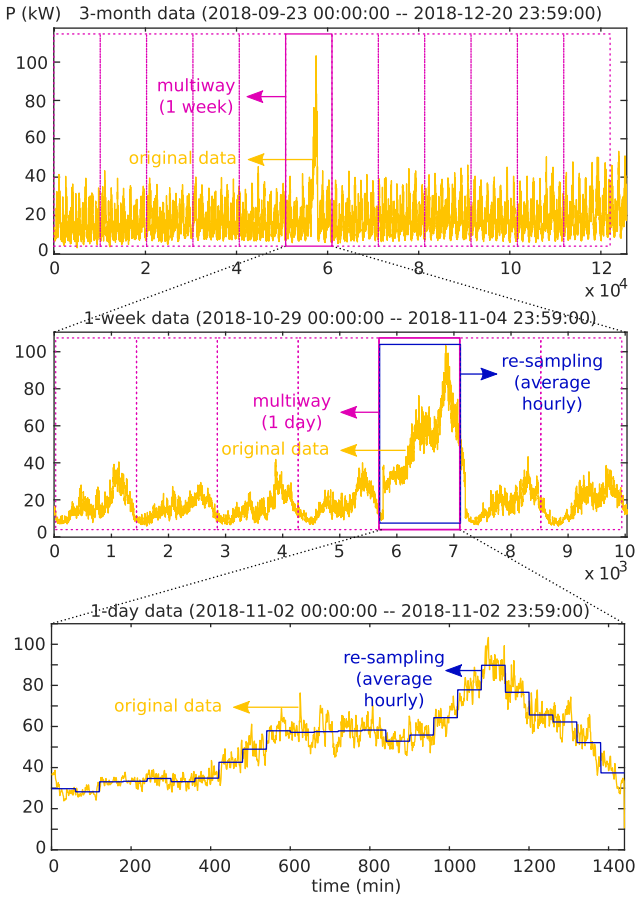

Fig. 4. Example of data organization over power profiles: original data represented by solid yellow lines, multiway re-arrangement represented by pink dashed boxes, and re-sampling with average values represented by the solid blue line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Case study

This section presents an overview of the monitoring network in Section 3.1, the main objectives and analysis description in Section 3.2, and results in Section 3.3.

3.1. Network

The presented strategy was tested in MATLAB using data from the OpenLV project at Marshfield Village, South Gloucestershire, United Kingdom. For more information about the OpenLV project and the chosen area, see [19].

A map of the approximate locations of the IEDs in this network (OpenLV substations 43, 44, 69, and 70) is displayed in Fig. 5. They are installed at LV distribution level ($415V$) in different areas of a three-phase unbalanced network, where single-phase solar PV panels are also installed, and record average values of phase voltage and line current magnitudes and active and reactive power and energy every minute (i.e., $f_{raw} = 1$ min^{-1}). Therefore, $m_{raw} = 4 \times 3 \times 6 = 72$ in total. The measurements aggregate energy consumption from distinct households

Fig. 5. Map of IED locations within the Marshfield network.

and buildings and energy production from solar PV panels whose rated power and connection points are known.

3.2. Main objectives and analysis description

Considering the time ranges displayed in Fig. 1, f_{raw} and m_{raw}, and the records of OpenLV data from July 17, 2018 to June 30, 2019, the multilayer PCA strategy is aimed at identifying and characterizing distinct abnormal behavioral patterns associated with electricity production and consumption over time, from fast load behavioral changes at minute resolution (which can be seen in the reactive power Q changing from/to inductive to/from capacitive) to slow profile changes at day resolution (which can be seen in the active power P). The layers and problems of interest are defined accordingly in Table 1, together with their corresponding f_k, θ_k, σ_k, τ_k, and the input data set. However, only events that last a few minutes or longer are visible in the data, since $f_{raw} = 1\ \text{min}^{-1}$. In the input data, P stands for active power measurements, whereas Q stands for reactive power measurements. All data are single phase quantities acquired separately at the three phases of each substation.

It is noteworthy that not all measured quantities are required for the evaluation of a specific problem of interest. For instance, phase voltage and line current magnitudes are better indicators of power quality problems, whereas active and reactive power and energy are better indicators of energy behavioral patterns. To some extent, the power and energy measurements contain redundant information, as the energy quantities derive from their corresponding power quantities. Based on the measurement-based techniques summarized in [22], and considering that power is a better indicator of average behaviors over time than energy, by definition, the usage of active and reactive power is adopted for the problems of interest defined in Table 1. Additionally, the active and reactive power measurements are further adjusted to the timescales schematized in the temporal hierarchy of decisions of Fig. 1.

Layer 1 does not require a priori re-sampling or multiway re-arrangement due to its short duration; in contrast, layers 2 and 3 do, for their long duration. Thereby, in layers 2 and 3, the IED data are averaged (over an hour in layer 2 and over a day in layer 3) to represent a single measurement with a lower sampling rate, and further re-arranged multiway in layer 3 to evaluate repetitive patterns on a specific day of the week. This previous data organization is necessary whenever the data have to be adjusted within a specific problem of interest.

The PCA models are built on the fly, as soon as new data are ready for analysis over τ_k at the k layers described in Table 1. The analysis relies on measurements collected from July 17,2018 to June 30, 2019. Moreover, power system operation over a day is further divided in two periods, before noon and after noon, as a typical day presents two different load peaks, one in the morning and the other in the afternoon (see Fig. 6). This choice is made to catch these two daily load peaks in separate within the statistical models, such that one peak does not interfere with the other. This is taken into consideration in layers 2 and 3, which contain half the number of observations of the whole day.

A confidence level $\alpha = 0.95$ is chosen for the whole analysis, as it results in a few observations surpassing the thresholds calculated analytically with (8) and (10). This means that the conclusion reached by the experiment will actually be wrong (that is, result in false positives or negatives) in 5% of the tests. Further investigation is required to discard false positives and negatives (i.e., wrong event detections and missed event detections). For instance, in layer $k = 1$, there are $n_1 = 60$ observations, of which 3 are expected to be wrongly classified; in $k = 2$, there are $n_2 = 84$ observations, of which at least 4 are expected to be wrongly classified; and in $k = 3$, there are $n_3 = 26$ observations, of which at least 1 is expected to be wrongly classified.

3.3. Results

This section presents examples of detection and diagnosis of abnormal power consumption patterns using the multilayer PCA strategy presented in Section 2, the scenarios described in Section 3, and OpenLV data recorded on November 02, 2018. The computation time of all layers is of a few milliseconds, which enables an online implementation of the methodology. The active power profiles displayed in Fig. 6 contain the abnormal behaviors under evaluation in layers 2 and 3 of Table 1, further averaged and separated in before noon and after noon. In addition, the multiway re-arrangement introduced in layer 3 allows for a comparative evaluation of weeks.

For layers 1 to 3 of Table 1, event detection results are displayed in Figs. 7–9, whereas contributions in terms of T^2 and SPE statistics for the specific events highlighted in Figs. 7–9 are illustrated graphically in Figs. 10–12 in the projection subspace (top chart) and residual subspace (bottom chart), respectively. Although there are a few points outside the square area in Figs. 7–9, this section focuses on the occurrences highlighted in Figs. 7–9 to illustrate the methodology. All graphs display the statistical thresholds calculated individually for each substation variable (i.e., active power P or reactive power Q at phases 1, 2, and 3 of substations 43, 44, 69, and 70), represented by the solid black line, together with the contributions of each substation variable to the calculated values of T^2 and SPE, represented by the column charts. Although there are more abnormal observations in Figs. 7–9 than those highlighted, the analysis focuses on those points in particular.

In these examples, the individual contributions that violate the statistical thresholds of T^2 indicate that the reactive power variables $Q_{43,1}$, $Q_{43,3}$, $Q_{44,1}$, and $Q_{69,1}$ are the main involved in the statistically abnormal behavior in Fig. 10 from 07:05 PM to 07:10 PM; and that the active power variables $P_{69,1}$, $P_{69,2}$, and $P_{69,3}$ are involved in the abnormal behavior in Fig. 11 and that none of the IEDs is involved in Fig. 12. Further evaluation of the network topology suggests that constraint violations with T^2 statistics are probably due to the solar PV panels

Table 1
Layers and problems of interest.

k	Problem of interest	Time lengths				Input			
		f_k	θ_k	σ_k	τ_k	data	n_k	m_k	p_k
1	Load behavior	1 min^{-1}	1 min	1 min	1 h	Q only	60	12	1
2	Hourly power changes	1 h^{-1}	1 h	1 h	1 week	P only	84	12	1
3	Daily power changes	1 day^{-1}	1 week	1 week	6 months	P only	26	12	7

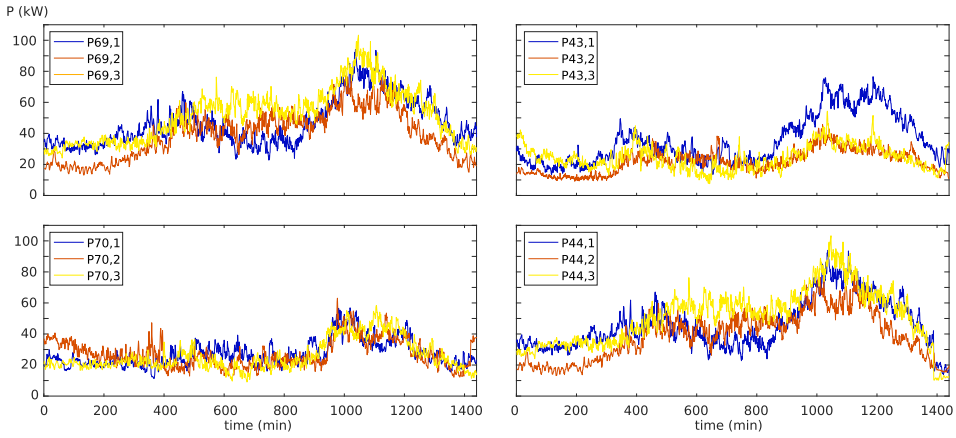

Fig. 6. Single-phase daily active power profiles recorded at substations 43, 44, 69, and 70 on November 02, 2018.

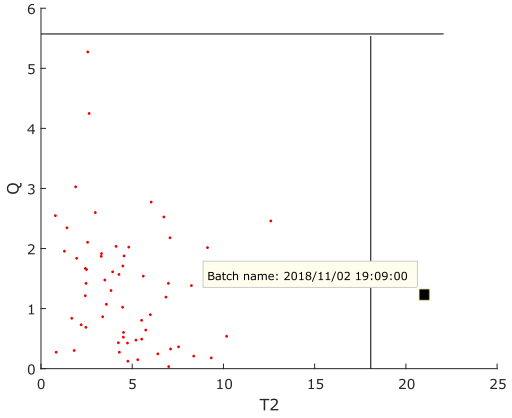

Fig. 7. SPE vs. T^2 results at layer 1: minutes between 07:00 PM and 07:59 PM on November 02, 2018.

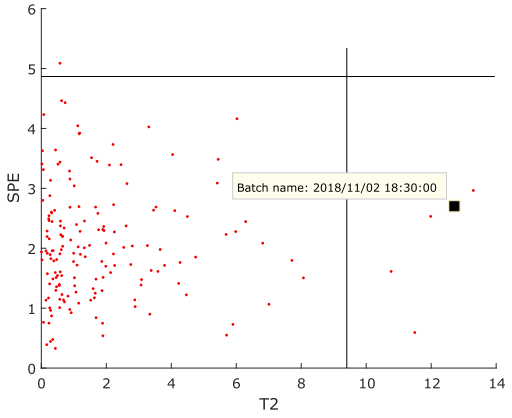

Fig. 8. SPE vs. T^2 results at layer 2: hours from October 29, 2018 to November 04, 2018.

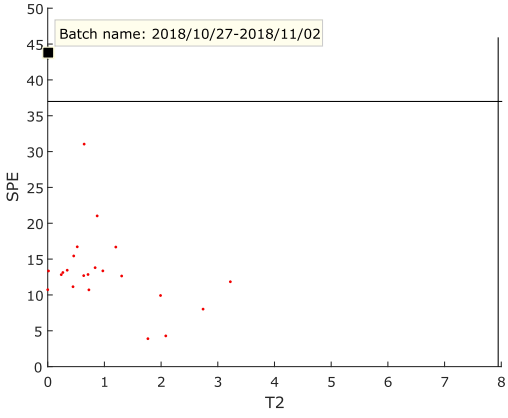

Fig. 9. SPE vs. T^2 results at layer 3: afternoons of the weeks in the 2018 Autumn term.

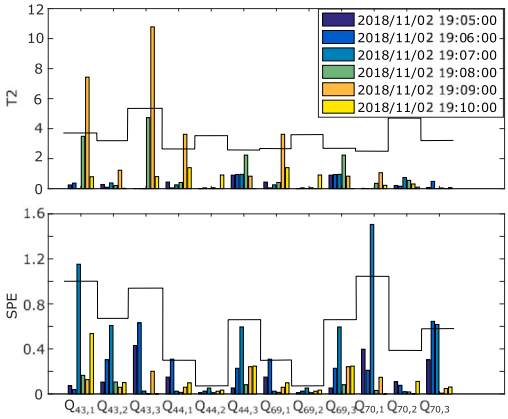

Fig. 10. Contribution analysis of T^2 (top) and SPE (bottom) at layer 1: load behavior.

connected to the feeders of the substations involved in the occurrence followed by an increase in electricity consumption. In fact, the presence of inverter-based generation changes the behavior of the network from inductive to capacitive over a few minutes (Fig. 10) while increasing the injection of active power in the network (Fig. 11), whereas an increase in electricity consumption, reflected in the active power demand, is related to changes in the standard operation of the grid over longer intervals. In turn, the individual contributions that violate the statistical thresholds of SPE indicate that the reactive power variables $Q_{43,1}$, $Q_{70,1}$, and $Q_{70,3}$ are the main involved in the abnormal behavior in Fig. 10 from 07:05 PM to 07:10 PM; that the active power variable $P_{69,3}$ is involved in the abnormal behavior in Fig. 11; and that the active power variables $P_{69,1}$, $P_{69,2}$, and $P_{69,3}$ gathered on November 02, 2018 are involved in the abnormal behavior in Fig. 12. Further evaluation of the network topology shows that constraint violations with SPE statistics are due to a high energy consumption within the coverage area of the IEDs involved in the occurrence.

4. Discussion

It can be noticed from Section 3.3 that the multilayer PCA strategy presented in this article allows for detection and isolation of different

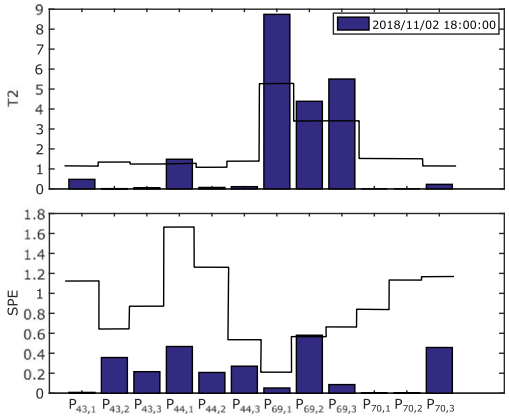

Fig. 11. Contribution analysis of T^2 (top) and SPE (bottom) at layer 2: hourly power changes.

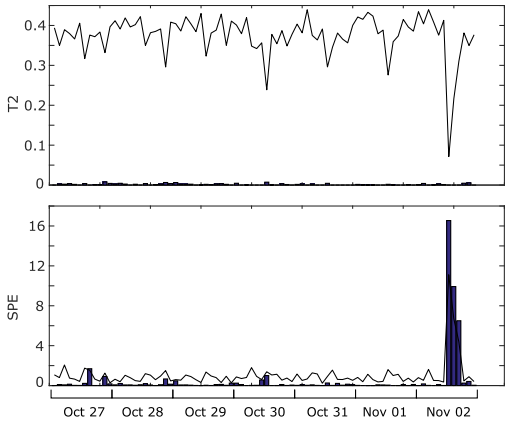

Fig. 12. Contribution analysis of T^2 (top) and SPE (bottom) at layer 3: daily power changes.

types of abnormalities in energy production and consumption patterns in parallel, taking into consideration the time ranges of distinct phenomena or problems of interest while respecting the temporal hierarchy of decisions. As an outcome, multilayer PCA is more flexible and scalable than traditional multiway PCA when it comes to the timescales, variables, and layers of interest.

The data organization enables to apply the same data set to identify different types of abnormal behaviors more effectively. This procedure provides a more appropriate sampling rate so that a specific set of variables can be used over the length of time associated with a given problem of interest. As a result, this procedure ensures good performance with an adequate scaling of the data at different layer levels in the next steps of the strategy. On the top of this, the results obtained with data organization show that it is of paramount importance to select an appropriate sampling according to the length of time of the layer so that different types of abnormal behaviors can be detected with an adequate representation over time. If this procedure was not used, the results detected at the longest layers would contain redundant information, which is not desired.

In the case study described in Section 3, variations in the normal operating conditions are attributed to the uncertain, intermittent nature of the electricity production from solar PV panels, whereas other perturbations are attributed to abnormal energy consumption patterns. Consequently, events detected with T^2 statistics are related to the energy generated by the solar PV panels, as it measures the distance of the projected data to the center of the model and is expected to present larger values associated with variations in the normal operating conditions, whereas events detected with SPE statistics are related to the energy consumption, as it measures the distance of the observation to the projection subspace and is expected to present larger values associated with changes in the correlation structure of the observations.

Moreover, it is recommended to proceed with a further investigation of any results detected with any of the T^2 or SPE statistics, as they produce violation of constraints related to major concerns about power system operation in different ways. As a matter of a fact, events detected with T^2 statistics are associated with deviations from the average standard operating conditions, whereas events detected with SPE statistics reflect changes in the correlation structure of the observations. As a result, this comprehensive analysis prevents the neglect of abnormal behaviors at distinct layer levels with the data set in use.

However, it is noteworthy that any types of abnormal behaviors can be identified only if they produce changes in the measured quantities that last enough to be recorded. This principle applies to parts of the grid that are directly or indirectly monitored by IEDs.

The results presented in Section 3.3 also show that it is possible to zoom in and out of abnormal energy production and consumption patterns through this multilayer PCA and thereby associate distinct problems of interest over time whenever they are identified at different layers. In addition, it can be noticed that, the shorter duration of the layer and the higher sampling rate of the data, the more principal components are needed to express the same information. It shows that the previous data preprocessing step is effective to address problems of interest over distinct timescales with adequate data. As a matter of a fact, for the case study presented in Section 3, the analysis requested $r_1 = 3$ hourly, $r_2 = 2$ weekly, and $r_3 = 1$ biannually.

Furthermore, the procedure of isolation enables to identify the most probable substations, lines, and electrical quantities responsible for the abnormal behavior, as it computes the influence of each variable – referred to an electrical quantity of a substation – in the calculated values of the T^2 and SPE statistics. Consequently, information about the network topology and its electrical parameters and energy appliances are helpful for the correct identification of the most probable locations and causes of the abnormal behavior.

5. Conclusion

This paper presented a strategy to identify abnormal operating conditions in power system data gathered by IEDs based on a multilayer implementation of PCA with use of T^2 and SPE statistics. The methodology is able to detect and diagnose occurrences of different nature and time spans with significant dimensionality reduction, which is advantageous to process large amounts of data collected by many IEDs installed at different locations. Moreover, this multilayer PCA enables to tackle distinct problems of interest in parallel over time, re-arranging the data to fit the scale and purpose of the analysis when required. The presented methodology was tested with OpenLV data in a case study focused on the detection, isolation, and diagnosis of abnormal energy production and consumption patterns at different timescales. The results indicate that the presented method is accurate and efficient, as long as an adequate data set is used to build the PCA model at each individual layer.

Declaration of Competing Interest

None.

Acknowledgments

The authors would like to thank the OpenLV project for providing network data. OpenLV is part funded by Ofgem's Network Innovation Competition funding, Western Power Distribution as the host Distribution Network Operator, and EA Technology through in kind contribution. This work has been supported by the European Union's Horizon 2020 research and innovation framework under the auspices of

the project *Renewable penetration levered by Efficient Low Voltage Distribution grids*, grant agreement number 773715, and University of Girona scholarship.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijepes.2020.106471.

References

[1] Stott B, Alsac O, Monticelli AJ. Security analysis and optimization. Proc IEEE 1987;75:1623–44.
[2] Sauer PW, Pai MA. Power system dynamics and stability. Prentice-Hall; 1998.
[3] IEEE, IEEE recommended practice for monitoring electric power quality, IEEE Std 1159-2009 (Revision of IEEE Std 1159-1995); 2009. p. 1–94.
[4] IEEE, IEEE guide for collecting, categorizing, and utilizing information related to electric power distribution interruption events, IEEE Std 1782-2014; 2014. p. 1–98.
[5] Löf A, Repo S, Pikkarainen M, Lu S, Pöhö T. Low voltage network monitoring in rtds environment 2013:1–5.
[6] Dedé A, Della Giustina D, Rinaldi S, Ferrari P, Flammini A, Vezzoli A. Smart meters as part of a sensor network for monitoring the low voltage grid 2015:1–6.
[7] Barbato A, Dedé A, Giustina DD, Massa G, Angioni A, Lipari G, et al. Lessons learnt from real-time monitoring of the low voltage distribution network. Sustain Energy, Grids and Networks 2018;15: 76–85. Technologies and Methodologies in Modern Distribution Grid Automation.
[8] Junior WLR, Borges FA, Veloso AF da S, Rabêlo R de AL, Rodrigues JJ. Low voltage smart meter for monitoring of power quality disturbances applied in smart grid. Measurement 2019;147:106890.
[9] Silva N, Basadre F, Rodrigues P, Nunes MS, Grilo A, Casaca A, Melo F, Gaspar L. Fault detection and location in low voltage grids based on distributed monitoring 2016:1–6.
[10] van Deursen A, Wouters P, Steennis F. Corrosion in low-voltage distribution networks and perspectives for online condition monitoring. IEEE Trans Power Deliv 2019;34:1423–31.
[11] Maaß H, Cakmak HK, Bach F, Mikut R, Harrabi A, Süß W, et al. Data processing of high-rate low-voltage distribution grid recordings for smart grid monitoring and analysis. EURASIP J Adv Signal Process 2015;2015:14.
[12] Stefan M, Lopez JG, Olsen RL. Exploring the potential of modern advanced metering infrastructure in low-voltage grid monitoring systems 2018:3543–8.
[13] Russell E, Chiang L, Braatz R. Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes. Advances in Industrial Control. London: Springer; 2000.
[14] Barocio E, Pal BC, Fabozzi D, Thornhill NF. Detection and visualization of power system disturbances using principal component analysis. In: 2013 IREP Symposium Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid; 2013. p. 1–10. doi:10.1109/IREP.2013.6629374.
[15] Liu X, Laverty DM, Best R, Li K, McLoone S. Principal component analysis of wide area phasor measurements for islanding detection - a geometric view. IEEE Trans Power Deliv 2015;30:976–85.
[16] Guo Y, Li K, Laverty DM, Xue Y. Synchrophasor-based islanding detection for distributed generation systems using systematic principal component analysis approaches. IEEE Trans Power Deliv 2015;30:2544–52.
[17] Rafferty M, Liu X, Laverty DM, McLoone S. Real-time multiple event detection and classification using moving window PCA. IEEE Trans Smart Grid 2016;7:2537–48.
[18] Ayech N, Chakour C, Harkat MF. New adaptive moving window pca for process monitoring, IFAC Proceedings Volumes 2012;45: 606–11. 8th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes.
[19] E. Technology, OpenLV – the groundbreaking project that's making local electricity data openly available; 2017. URL: http://openlv.net.
[20] C. of European Energy Regulators, Ceer benchmarking report 6.1 on the continuity of electricity and gas supply; 2018.
[21] Qin SJ, Dunia R. Determining the number of principal components for best reconstruction. J Process Control 2000;10:245–50.
[22] Arif A, Wang Z, Wang J, Mather B, Bashualdo H, Zhao D. Load modeling–a review. IEEE Trans Smart Grid 2018;9:5986–99.