Design and Implementation of a Data-Driven Approach to Visualizing Power Quality

Xiao, Fei; Lu, Tianguang ; Ai, Qian ; Wang, Xiaolong; Chen, Xinyu; Fang, Sidun; Wu, Qiuwei

Published in:
IEEE Transactions on Smart Grid

Link to article, DOI:
10.1109/TSG.2020.2985767

Publication date:
2020

Document Version
Peer reviewed version

Design and Implementation of a Data-Driven Approach to Visualizing Power Quality

Fei Xiao, Tianguang Lu, *Member, IEEE*, Qian Ai, *Senior Member, IEEE*, Xiaolong Wang, *Member, IEEE*, Xinyu Chen, *Member, IEEE*, Sidun Fang, *Member, IEEE*, and Qiuwei Wu, *Senior Member, IEEE*,

***Abstract*—Numerous underlying causes of power-quality (PQ) disturbances have enhanced the application of situational awareness to power systems. This application provides an optimal overall response for contingencies. With measurement data acquired by a multi-source PQ monitoring system, we propose an interactive visualization tool for PQ disturbance data based on a geographic information system (GIS). This tool demonstrates the spatio–temporal distribution of the PQ disturbance events and the cross-correlation between PQ records and environmental factors, leveraging Getis statistics and random matrix theory. A methodology based on entity matching is also introduced to analyze the underlying causes of PQ disturbance events. Based on real-world data obtained from an actual power system, offline and online PQ data visualization scenarios are provided to verify the effectiveness and robustness of the proposed framework.**

***Index Terms*—situation awareness, power quality, geographic information system, Getis statistics, random matrix theory, entity matching.**

I. Introduction

Power quality (PQ) has become a challenging concern to power companies and their users. The causes of PQ disturbances are complex and mostly related to the performance of equipment, including line faults, capacitor actions, and start-up/shut-down of large motors [1]. Successful visualization of power disturbances is a worthwhile goal for providing a clear overview of large PQ datasets, grasping the characteristic properties of PQ data, and detecting contingencies related to equipment failures.

With the decreased cost of power-measurement devices and rapid development of cyber-physical systems in power grids, a large volume of PQ disturbance-related data can be generated and stored [2]. Current PQ monitoring systems cover all levels of urban substations. These PQ-related measurements have potential advantages in certain applications, such as analyzing PQ characteristics [3], locating sources of PQ disturbance [4], detecting and classifying PQ disturbances [5], and monitoring online loads [6]. In [7], a Hadoop-based PQ data-processing platform is proposed to assess the severity levels of PQ disturbances and predict PQ events. Visualization is an important method for big-data analytics, particularly for revealing connections between the quantitative content of the data and human intuition. However, achieving visualization in the case of heterogeneous and diverse data (unstructured, structured, and semi-structured) is challenging. Cleaning and integrating techniques for PQ data are required to optimally support advanced analysis.

The first research on data cleaning was conducted in the United States to correct errors in social security numbers [8]. Subsequent studies focus on detection of abnormal data, deletion of duplicate data, and integration of data. Research on PQ data cleaning and integration is scant but is urgently needed for three reasons. 1) Disturbance energy (DE) caused by equipment failures propagates between lines, which may trigger the same PQ meter several times in short time. 2) The trigger mechanism of PQ meters for complex power quality events is imperfect, leading to duplicate data. 3) A PQ disturbance may be correlated with device operations, such as protective relays, capacitor banks, and large motor operations [9]. Processing of PQ and power-network operation data can contribute to determining the underlying causes of PQ events. This study proposes a novel block processing-based method (BPM) to detect and delete duplicate items in PQ data. A data-integration methodology based on entity matching is also utilized to analyze the underlying causes of PQ events that are often related to power-grid contingencies.

A number of visualization tools have been developed to quickly detect power-grid contingencies [10]–[12]. In [10], interactive 3D visualizations are explored to demonstrate contingency data that can help system operators comprehend the static security status quickly and intuitively. In [11], color contours are used to help visualize information on the magnitude of power-system bus voltages. The test results indicate that color can be effective highlighting feature to reduce the size of the search space and facilitate target detection. Visualization techniques of PQ data related to power-grid contingencies have also been developed in [13]–[17]. Several types of graphs are presented in [13] to assess

This work is supported by National Natural Science Foundation of China under Grant U1766207 and Grant U1866206, by Harvard Global Institute, by Key Research and Development Program of Shandong Province under Grant 2018GGX103048, and by The Fundamental Research Funds of Shandong University under Grant 2018TB037. (*Corresponding author: Tianguang Lu.*)

F. Xiao and Q. Ai are with Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xiaofeisjtu@163.com; aiqian@sjtu.edu.cn).

T. Lu is with the School of Engineering and Applied Sciences and Harvard China Project, Harvard University, Cambridge, MA 02138, USA (e-mail: tlu@seas.harvard.edu).

X. Wang is with Shandong University, Jinan 250061, China (e-mail: tzy@sdu.edu.cn).

X. Chen is with Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xchen2019@hust.edu.cn).

S. Fang is with Nanyang Technological University, 639798 Singapore (e-mail: fangston@foxmail.com).

Q. Wu is with Technical University of Denmark (DTU), Kgs. Lyngby 2800, Denmark (e-mail: qw@elektro.dtu.dk).

the PQ risk in multiple locations, but data-visualization theory has been rarely presented. In [14], the theoretical foundations of data visualization are outlined, and its applications to PQ data are demonstrated with examples. In [15], the representations of voltage-space vectors, namely, cylindrical and polar representations, are introduced for visualization to clarify the evolution of global voltage dips over time. Reference [16] offers a visualization tool for online tracking of voltage-sag events and predicting future incidents. In [17], a quadtree-based map is used in a visualization algorithm to display PQ data without sacrificing user privacy. In the existing literature, few studies report on the integration of geographical information data and visualization of the correlations between PQ data and external-network factors (e.g., weather and temperature).

As a result of experience/hypothesis-based approaches, traditional PQ data-visualization tools can only partially discover the patterns hidden in PQ data. A data-driven approach will fundamentally overcome this shortcoming [18]. In this study, we developed a data-driven method that relies on the analysis of preprocessing and correlation algorithms to create visualizations on the basis of the characteristics of empirically or mathematically derived data.

The study of data-driven approaches to power system visualization has recently drawn attention. Specifically, research has been conducted on the visualization of the correlation among power-grid measurements based on a spatial autocorrelation index and random matrix theory (RMT) [18]–[27]. Given that variations of renewable energy in neighboring sites are strongly correlated, geographical autocorrelation parameters are calculated to help predict the output of wind and solar power [18]–[20]. The abnormal events must be detected through correlation analysis to achieve enhanced situational awareness. In [24], a principal component analysis (PCA)-based statistical monitoring framework is proposed for islanding detection by visualizing Hotelling's T^2 and Q statistics. On this basis, a moving window PCA (MWPCA) approach [25] is developed for classifying multiple cascading events. However, the large principal components calculated by PCA-based method cannot obtain all information from data sources. In [26], the mean spectral radius (MSR) index calculated by RMT is leveraged to visualize the correlations of the whole power grid measurements. Reference [27] uses an augmented matrix to fuse data from different sensor types and visualizes the correlation by using a ring graphic. However, few tools are available to visualize the correlation analytical results obtained by RMT.

In accordance with the previous discussion, GIS-based spatial autocorrelation is used in this study to illustrate the spatial density of PQ events. Although the geographic autocorrelation parameters can reveal the correlation between neighboring PQ events, it cannot customize the dynamically created displays to meet users' specific needs. This study extends the RMT-based correlation analysis method [27]. In our proposed method, an RMT-based data-analysis-and-display framework is designed to reveal correlations between the selected factors (e.g., weather and temperature) and the PQ level (i.e., the number of PQ records). The reasons for selecting RMT are as follows: 1) PQ disturbances are usually regarded as random events, and RMT-based method can effectively recognize valid information from random events. 2) RMT can reveal the global-association characteristics in real data, and it is suitable for the analysis of correlative PQ events caused by spatial propagation of DE.

The key contributions of this study are threefold:

1) A BPM-based PQ data cleaning methodology, which comprises blocking and deleting phases, is designed to detect duplicate PQ records. Especially for complex PQ disturbance, the proposed cleaning methodology has higher computational efficiency and detection accuracy than existing methods. The proposed methodology based on entity matching can also efficiently analyze the underlying causes of PQ sources by considering the temporal and spatial properties of PQ records.

2) The spatial-autocorrelation local index of Getis statistics is used for the first time to illustrate the propagation properties of PQ DE. This technique enables the analysis and visualization of the spatial concentration and regional effect of PQ events. The proposed GIS-based graph partitioning approach can also adjust the visualization resolution to satisfy various application scenarios.

3) A novel RMT-based visualization method for PQ disturbances is proposed, including PQ meter allocation, GIS-based data source integration, RMT-based correlation analysis, and visualization. Compared with the conventional correlation analysis method, the proposed method is based on the spatial-temporal characteristics of PQ events. Combined with RMT, this method exhibits higher sensitivity to correlation analysis and is robust against random PQ disturbance data and error measurement. The proposed visualization method can also realize offline and online correlation analyses.

The rest of this paper is organized as follows. The basic modules for preprocessing, analyzing, and visualizing PQ data are described in Section II. Case studies based on real-world data are presented in Section III, followed by the concluding remarks in Section IV.

II. Basic Methodology

The proposed framework for visualizing PQ disturbances comprises four modules (Fig. 1). The first and second modules involve PQ data preprocessors, including blocking and deleting stages for merging duplicated PQ records. In the third module, we determine the PQ-related power system operations by using entity matching-based data integration. The fourth module includes the visualization of preprocessed PQ disturbance data. This module comprises one type of graph partition approach and three types of visualization tools: partitioning geographic information graph into pixels by locating the PQ meters and visualizing (a) the underlying cause analysis, (b) spatio–temporal distribution, and (c) cross-correlations of PQ disturbances. In particular, visualization of the spatio–temporal distribution contributes to assessing the severity of PQ events caused by various faults. The cross-

correlation visualization can also illustrate the correlation between PQ events and environmental factors. For example, a strong correlation can be observed between weather conditions and PQ events caused by transmission line faults [9]. The following sub-sections describe the specific steps in the four modules.

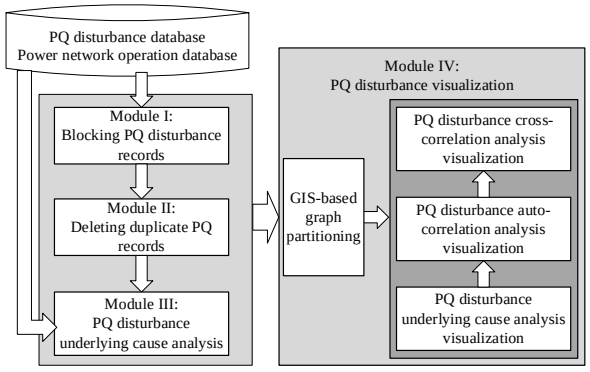


Fig. 1. PQ data cleaning and visualization framework.

A. Module I: Blocking PQ records

The same disturbance source is simultaneously recorded at several neighboring PQ measurement sites because of the DE propagation characteristics. We then divide the PQ disturbance records into block collection $\boldsymbol{B}$ on the basis of the index of the beginning time [7]. To facilitate the usage of subsequent modules, we define each block of records as a basic PQ event.

The quality of $\boldsymbol{B}$ is estimated in terms of two competing criteria: efficiency and effectiveness. The former is directly related to its aggregate cardinality $\| \boldsymbol{B} \|$, which denotes the total number of comparisons $\| \boldsymbol{B} \| = \sum_{b_i \in \boldsymbol{B}} \| b_i \|$, where $\| b_i \|$ is the individual cardinality of b_i, i.e., total number of comparisons entailed in ith blocks. The effectiveness of the blocking phase depends on the cardinality of the set $C(\boldsymbol{B})$ of detectable matches, i.e., pairs of duplicate PQ records compared with at least one block. Therefore, the pair completeness (PC) and reduction ratio (RR) are used for estimating the effectiveness and efficiency for $\boldsymbol{B}$ [28].

Pair completeness: PC assesses the portion of duplicates that share one block and, thus, can be detected. PC is formally defined as $PC(\boldsymbol{B}) = |C(\boldsymbol{B})| / |C(\boldsymbol{R})|$, where $|C(\boldsymbol{R})|$ is the number of duplicates in the sets of PQ disturbance records $\boldsymbol{R}$. PC takes values in the interval [0, 1], with high values indicating high effectiveness for $\boldsymbol{B}$.

Reduction ratio: RR measures to which degree efficiency is enhanced with respect to a baseline block $\boldsymbol{B}_{bs}$. RR is defined as $RR(\boldsymbol{B}, \boldsymbol{B}_{bs}) = 1 - \| \boldsymbol{B} \| / \| \boldsymbol{B}_{bs} \|$ and $\| \boldsymbol{B} \| \leq \| \boldsymbol{B}_{bs} \|$. The values of RR are in the interval [0, 1] with high values denoting high efficiency for $\boldsymbol{B}$.

A clear trade-off can be observed between the effectiveness and the efficiency of the blocking phase. When a number of comparisons are executed, the effectiveness is intensified, but the efficiency is lowered, and vice versa. The goal of the blocking phase is to maximize the PC and RR. This phase is formulated as a minimum problem $\min Blk(\phi)$ for convenience of calculation. $Blk(\phi)$ is defined in (1),

$$Blk(\phi) = W_1 \frac{1}{PC(\boldsymbol{B})} + W_2 \frac{1}{RR(\boldsymbol{B})}, \tag{1}$$

where ϕ is a correlation threshold used to separate PQ records, and W_1 and W_2 are the weighting factors.

The blocking phase is described in **Algorithm I**. The temporal correlation record is defined as

$$Cor(r_i, r_j) = \left| T(r_i) - T(r_j) \right|, \tag{2}$$

where r_i and r_j are the ith and jth PQ disturbance records; $T(r_i)$ and $T(r_j)$ denote the recording time of r_i and r_j, respectively [29]. The pseudo-code of **Algorithm I** is presented below.

Line 1: The records of the PQ disturbance are sorted by the index of the beginning time.

Line 2: The algorithm parameters are initialized before searching for blocks. The boundary pairs for each block of PQ records saved in S_B, are identified by calculating the temporal correlation of adjacent records. The moving-window technique is adopted to reduce the computational burden, and its corresponding starting and ending indices are w_1 and w_2 [30].

Lines 3–4: The optimal PQ record blocking parameter is obtained. The minimum problem $\min Blk(\phi)$ is solved by Pareto method. The fuzzy satisfying method is then used to determine the weighting factors and Pareto optimal solution, namely, the correlation threshold ϕ_{op}. The baseline block $\boldsymbol{B}_{bs}$ is obtained when $\phi = \phi_{\max}$, and index PC is calculated by using Module II.

Lines 7–12: The temporal correlation of the sorted adjacent PQ records in the same moving window is calculated using (2). If $Cor(\hat{r}_{w_1}, \hat{r}_{w_2}) > \phi_{op}$, then $\hat{r}_{w_1}$ and $\hat{r}_{w_2}$ are placed into different blocks. If $Cor(\hat{r}_{w_1}, \hat{r}_{w_2}) \leq \phi_{op}$, then $\hat{r}_{w_1}$ and $\hat{r}_{w_2}$ are placed in the same block. The $\hat{r}_{w_1}$ position is defined as the boundary between two adjacent blocks.

Algorithm I: Blocking PQ Disturbance Records

Input: PQ event records r_i $(i = 1, 2, \ldots, n_r)$, number of records n_r, and range of correlation threshold $[\phi_{min}, \phi_{max}]$.

Output: Sets of boundary pairs S_B.

1: Sort r_i by the beginning time of the PQ records
2: Initialize: $S_B = \{\}$, $w_1 = 1$, and $w_2 = 2$
/* Determine the data blocking parameter */
3: Solve the function $\min Blk(\phi)$, $\phi \in [\phi_{\min}, \phi_{\max}]$
4: Obtain the optimal correlation threshold ϕ_{op} by using fuzzy satisfying method
/* Search blocks */
5: **while** $w_2 < n_r$ **do**
6: Save the starting position of block $S_B = S_B + \{w_1\}$
7: **if** $Cor(\hat{r}_{w_1}, \hat{r}_{w_2}) \leq \phi_{op}$, **then**
8: Reposition the window $w_1 = w_1 + 1, w_2 = w_2 + 1$
9: **Else**
10: Save the ending position of the block $S_B = S_B + \{w_1\}$
11: Reposition the window $w_1 = w_1 + 1, w_2 = w_2 + 1$
12: **end if**
13: **end while**
14: $S_B = S_B + \{n_r\}$

13: **return** S_B

B. Module II: Deleting duplicate PQ records

Power DE is rapidly reflected from the terminal of power lines. The PQ meters are triggered multiple times in a short time, which may easily result in duplicate data records. The criterion for determining duplicate records is given as follows:

$$Comp(r_i, r_j) = \begin{cases} 1 & \text{if } L(r_i) = L(r_j) \text{ and } D(r_i) = D(r_j) \text{ and } M(r_i) = M(r_j), \\ 0 & \text{else} \end{cases} \quad (3)$$

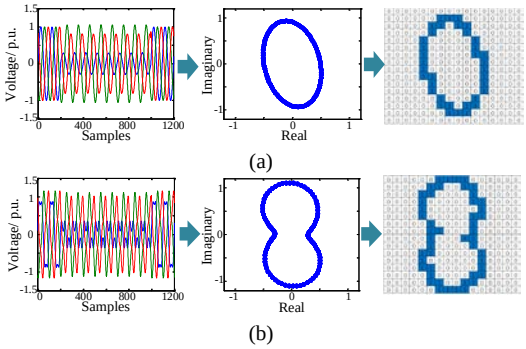
where $L(r_i)$ and $L(r_j)$, $D(r_i)$ and $D(r_j)$, and $M(r_i)$ and $M(r_j)$ represent the locations (i.e., buses), duration times, and magnitudes of the voltage of the ith and jth PQ disturbance records. However, the magnitudes of the voltage $M(r_i)$ are affected by the sampling frequency and measurement deviation. This study presents a robust PQ duplicate record detection method to solve this problem. In this method, the 2D representations of the PQ waveform replace the magnitudes of the voltage. Initially, $D(r_i)$ is calculated by the scaling coefficient energies of the phase voltages [29]. On this basis, the 2D representations of the PQ waveform are extracted from the PQ disturbance duration.

The space vector is formed with the first two components of Clarke transformation [31] in (4) to obtain the 2D representations.

$$s_v(t) = y_\alpha(t) + jy_\beta(t) = \frac{2}{3}\begin{bmatrix} 1 & e^{j2\pi/3} & e^{j4\pi/3} \end{bmatrix}\begin{bmatrix} v_a(t) \\ v_b(t) \\ v_c(t) \end{bmatrix}, \quad (4)$$

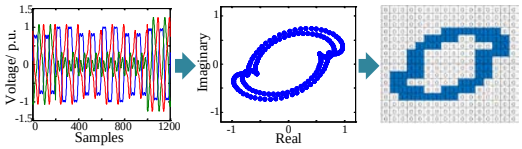
where $y_\alpha(t)$ and $y_\beta(t)$ are the transformed vectors; and $v_a(t)$, $v_b(t)$, and $v_c(t)$ are the three phase-to-neutral voltages. For a sinusoidal balanced voltage, the space vector is a circle centered around the origin in the complex plane. PQ disturbance leads to elliptical distortion.

The duplicate record detection is facilitated by converting 2D graphics into vectors comprising binary values $\boldsymbol{m}_n$. Fig. 2 demonstrates that the typical single and complex PQ disturbances are converted into complex planes forming a 2D representation with $n_t \times n_t$ components, where $n_t = 16$. Therefore, the comparison of magnitudes of the voltage of the PQ disturbance records can be converted into a comparison of 256D binary vectors, namely, $M(r_i) = \boldsymbol{m}_n$.



(a)

(b)



(c)

Fig. 2. Feature extraction of PQ data. (a) Voltage sag. (b) Voltage sag + harmonic. (c) Voltage sag + harmonic + voltage fluctuations.

Unlike existing point matching methods [32], the proposed method uses the shape features of the converted PQ data to avoid the effect of sampling rate. Moreover, this method has higher computational efficiency and robustness compared with waveform decomposition-based methods [29]. **Algorithm II**, which is used to delete the duplicate PQ data and divide the complex ones, is laid out as follows:

Lines 2–9: Whether the selected record r_ρ is repeated in the set of preprocessed PQ records S_P is checked by (3). If yes, then the iteration is continued. Otherwise, the record r_ρ is saved in S_P.

Lines 10–15: The complex PQ disturbances are decomposed into one fundamental component and several PQ disturbances $\Delta r_{i,j}$ by using matching pursuit (MP) algorithm with damped sinusoid representing power DE signals [33]. The PQ disturbance source locating algorithm [7] verifies whether the relative directions of the PQ disturbances $D_E(\Delta r_{i,j})$ and set of PQ records $D_E(S_{P,k})$ belong to the same disturbance source. The decomposed identical-source PQ disturbances are then divided into same block, and its information of beginning time and location will support the following underlying cause analysis.

Algorithm II: Deleting Duplicate PQ Disturbance Records

Input: PQ records saved in the same block.
Output: Preprocessed PQ records.

1: Initialize the set of preprocessed PQ records $S_P = \{\}$, $S_{P,k} = \{\}$, the number of PQ records n_b, and $\rho = 1$
/* Delete duplicate PQ disturbance records */
2: **while** $\rho < n_b$ **do**
3: Normalize the phase voltages of record r_ρ
4: Calculate the $D(r_\rho)$ and $M(r_\rho)$ by using scaling coefficient energies and space vector
5: **if** $Comp(r_\rho, S_P) = 0$, **then**
6: $S_P = S_P + \{r_\rho\}$
7: **end if**
8: Update $\rho = \rho + 1$
9: **end while**
/* Divide the complex PQ disturbances */
10: **for each** $r_i \in S_P$ **do**
11: Decompose the record r_i into $\Delta r_{i,j}$ by using MP
12: **if** $D_E(\Delta r_{i,j}) = D_E(S_{P,k})$, **then**
13: $S_{P,k} = S_{P,k} + \{\Delta r_{i,j}\}$
14: **end if**
15: **end for**
16: **return** $S_{P,k}$, $\Delta r_{i,j}$

C. Module III: Underlying cause analysis of PQ disturbances

In data preprocessing, the cleaned PQ disturbance records with identical-source are clustered into the same block. Considering the underlying cause of PQ disturbances, the records belonging to the same block are assumed to be related to the same device operations, such as a protective relay operation, a circuit breaker tripping, a shunt capacitor bank switching, or a large motor starting. In this study, the correlation between the PQ disturbance and the device operation is analyzed on the basis of the entity-matching method.

1) Entity resolution

Each PQ disturbance event is regarded as a real-world entity because of the imbalance in the number of PQ disturbance and power-network operation records. The sets of attribute names $\boldsymbol{N}$ and values θ are defined to match the power-network operation and PQ data.

DEFINITION 1. *An entity collection ε_i is a tuple $< N_i, \theta_i, P_i >$, where $N_i \subseteq \boldsymbol{N}$ is the set of available attribute names appearing, $\theta_i \subseteq \boldsymbol{\theta}$ is the set of utilized values, and $P_i \subseteq \boldsymbol{P}$ is the set of entity profiles. An entity profile P_i is a tuple $< i, \boldsymbol{A}_{Pi} >$, where $\boldsymbol{A}_{Pi}$ is the corresponding set of name–value pairs $< n, o >$, with $n \in N_i$ and $o \in \theta_i$.*

In two individual entity collections, $\boldsymbol{\varepsilon}_1$ (i.e., preprocessed PQ disturbance records) and $\boldsymbol{\varepsilon}_2$ (i.e., power network operation data), two entity profiles, namely, $p_1 \in \boldsymbol{\varepsilon}_1$ and $p_2 \in \boldsymbol{\varepsilon}_2$, are defined to be matching if they refer to the same PQ disturbance event (i.e., a real-world entity). $p_1 \equiv p_2$ denotes the relationship between the two collections. All matching entities in collections $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ must be identified to analyze the related power-network operations of PQ disturbance events. This condition is a problem of quadratic time complexity because the naive solution compares each entity from one collection with all entities from the other. Approximation techniques reduce a few comparisons to ensure scalability, sacrificing accuracy within a limited and controllable range. In the following, we propose a blocking-based entity resolution.

2) Blocking-based entity matching

The goal of blocking is to make entity resolution scalable by grouping similar entities (i.e., PQ disturbance and power-network operation data) into blocks such that they suffice to stimulate comparisons only within entities in the same block. Blocks are constructed according to a blocking scheme that comprises two phases: 1) a transformation function f_t that determines the blocking representation from each entity profile (in this study, the time stamp and location of records are leveraged for blocking) and a constraint function f_c that decides whether entity profiles are to be placed in the same block or not.

DEFINITION 2. *Given two entity collections, namely, $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$, a blocking scheme comprises a transformation function $f_t : \varepsilon_1 \cup \varepsilon_2 \mapsto \boldsymbol{E}$ and a constraint function $f_c : \boldsymbol{E} \mapsto \{true, false\}$, where $\boldsymbol{E}$ represents the space of all possible blocking representations for the given entity profiles.*

The proposed blocking-based methodology comprises three steps, as follows:

Step 1: Determine the set of global identifiers. This method integrates spatio–temporal information on each entity into a group of similar entities. The name–value pairs $< n, o >$ for an entity profile P_i are denoted as n = {*location of entity, time stamp of entity*} and its corresponding value $o = \{l_c, t_c\}$.

Step 2: Find the matching entities. Modules I and II are utilized to cluster processed PQ data into different blocks. The global identifiers of each PQ record block are expanded as $\hat{v} = \{\boldsymbol{l}_c, t_c, t_s\}$, where $\boldsymbol{l}_c$ denotes the set of locations of all records in one block and t_c and t_s represent the time stamp of the first and the last record in one block, respectively.

The functionality of entity matching is outlined in **Algorithm III**. Additional remarks are presented here.

Lines 3–5: The placement of the selected power network operation record r_j in the block S'_R is checked. If yes, then record r_j is saved in S'_R. Otherwise, the iteration is continued. This part works as a constraint function f_c. The time threshold φ is used to reduce the time-stamp error existing in different monitoring systems.

Algorithm III: Matching Entities

Input: The block of PQ records S_R, corresponding values of the global identifier $\hat{v} = \{\boldsymbol{l}_c, t_c, t_s\}$, power network operation entity collections $\boldsymbol{\varepsilon}_2$, and time threshold φ.

Output: Sets of power network operation matching entities S'_R.

1: Initialize $S'_R = S_R$
2: **for each** $r_j \in \boldsymbol{\varepsilon}_2$ **do**
3: **if** $t_c - \varphi < T(r_j) < t_s + \varphi$ **and** $L(r_j) \in \boldsymbol{l}_c$, **then**
4: $S'_R = S'_R + \{r_j\}$
5: **end if**
6: **end for**
7: **return** S'_R

Step 3: Merge the results of entity matching. The first recorded operating device is regarded as the PQ disturbance-related device because of the continuous operations of protective relays.

D. Module IV: PQ disturbance visualization

GIS-based PQ data visualization is of importance for operators to improve the PQ level, formulate correlative solution measures, and establish technical parameters of the controlling device. Four phases are conducted in this module.

1) Phase I: GIS-based graph partitioning

To encourage the application of GIS tools to perform the following advanced visualization, the GIS-based partition algorithm is applied to divide network graph into square segments, i.e., pixels, as shown in Fig. 3(a). In view of the sparseness of the spatial distribution of PQ meters and the accuracy of the visualization, the multi-size pixels are obtained by the proposed partition algorithm, which comprises the following two steps:

Step 1: Calculate the density of each pixel. The density value is denoted as the total number of PQ meters in each pixel.

Step 2: Partition pixels. If the densities of pixel exceed the threshold τ, then the pixel is divided into four same-size pieces, until the densities of all pixels are below the threshold.

The functionality of pixel partition is outlined in **Algorithm IV**. Additional remarks are presented here.

Lines 2–7: Each pixel is partitioned continuously until the density of all pixels is less than the threshold τ. The function $Den(\xi_{\text{sub}i})$ denotes the density of pixel $\xi_{\text{sub}i}$.

Lines 9–10: The function $Par(\xi^{\text{cur}})$ is leveraged for partitioning each pixel. Functions ξ^{cur} and $\xi^{\text{cur}}_{\text{sub}i}$ are the current pixel and its sub regions, respectively.

Algorithm IV: Partitioning Pixels

Input: The locations of PQ meters, GIS-based graph ξ and partition threshold τ.

Output: Results of partitioned graph $\overline{\xi}_{\text{sub}}$.

1: Divide the complete graph ξ into four segments $\xi_{\text{sub}i}, i = 1, \ldots, 4$
2: **for each** $\xi_{\text{sub}i} \in \xi$ **do**
3: Calculate the densities of pixel $Den(\xi_{\text{sub}i})$
4: **while** $Den(\xi_{\text{sub}k}) > \tau$ **and** $\xi_{\text{sub}k} \in \xi_{\text{sub}i}$ **do**
5: Implement $Par(\xi_{\text{sub}k})$
6: **end while**
7: **end for**
8: **return** $\overline{\xi}_{\text{sub}}$
Function $Par(\xi^{\text{cur}})$
9: Divide pixel ξ^{cur} into four segments $\xi^{\text{cur}}_{\text{sub}i}$
10: **return** $\xi^{\text{cur}}_{\text{sub}i}$

We present a simple analytical example in Fig. 3(b) to illustrate the general idea of the pixel partition functionality. We assume that the threshold $\tau = 3$. The left part in Fig. 3(b) is a complete graph where the locations of the PQ meters are denoted as red dot. The density of the graph is $Den(\xi) = 7$, which is larger than the threshold. On this basis, this graph is divided into four pixels (middle part of Fig. 3(b)). Only the density of pixel III is larger than the threshold. Accordingly, pixel III is further divided into four small pixels. The densities of all pixels in the right part of Fig. 3(b) are below the threshold. Therefore, the GIS-based visual maps with different resolutions can be obtained through changing the threshold value τ.

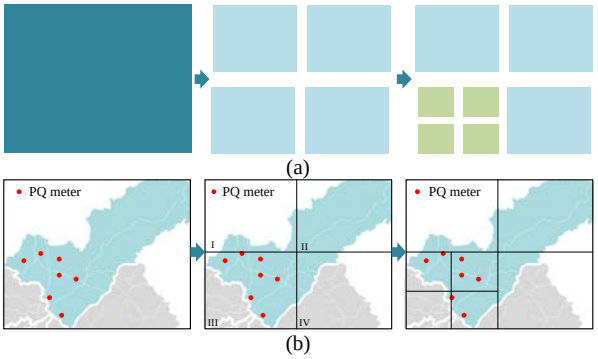


Fig. 3. Illustration of the pixel partition. (a) Pixel partition schematic. (b) Analytical example.

2) Phase II: Underlying cause-analysis visualization

The PQ records should be initially classified according to device operation, which is acquired from Module III. The PQ event frequency related to a specific device operation at every region in a time span of one month is visualized with a bar chart in this phase.

3) Phase III: Auto-correlation analytical visualization using Getis statistics

The partitioned GIS-based graph and PQ meter location are used in this phase to visualize the spatio–temporal distribution of the PQ records and clearly illustrate the records. The proposed visualization tool for spatio–temporal distribution can analyze the PQ data recorded with a time resolution of several seconds to as much as one year.

Considering the regional character of the PQ disturbance events, the spatial autocorrelation local index, i.e., Getis statistics, is calculated to illustrate the spatial concentration of PQ disturbances [21].

Getis statistics are defined as follows:

$$G_i^*(d) = \frac{\sum_j \omega_{ij}(d)x_j - W_i^*\overline{x}}{s\left[\frac{W_i^*(m_p - W_i^*)}{m_p - 1}\right]^{1/2}}, \tag{5}$$

where the matrix of spectral weights $\{\omega_{ij}(d)\}$ is binary and symmetric with a weight equal to unity ($\omega_{ij} = 1$) for all the pixels found within distance d of the ith pixel considered and a weight equal to zero ($\omega_{ij} = 0$) for all the pixels found outside d; $\Sigma\omega_{ij}(d)x_j$ is the sum of the varying values within a distance d of the ith pixel; m_p is the total number of pixels; and W_i^*, $\overline{x}$, and s are the number of pixels within the distance d, the global mean of x, and the variance of x, respectively, which are expressed as follows:

$$W_i^* = \sum_j \omega_{ij}(d), \tag{6}$$

$$\overline{x} = \frac{\sum_j x_j}{m_P}, \tag{7}$$

$$s^2 = \frac{\sum_j x_j^2}{(m_P - \overline{x}^2)}. \tag{8}$$

A cluster of pixels above the average digital counts produces mostly positive G_i^* values. By contrast, the pixels below the average digital counts produce mostly negative G_i^* values. A small threshold τ is set to visualize the details of the spatio–temporal distribution of the PQ records.

4) Phase IV: Cross-correlation analytical visualization using RMT

The aforementioned modules are implemented to obtain the tagged PQ data, which are used as input matrices for the cross-correlation analysis. Before constructing the input matrices for the RMT-based correlation analysis, the PQ data matrix is formed according to the topological locations of the metered buses to reflect the system configuration [34]. Fig. 4 illustrates the PQ data cross-correlation analysis and visualization. The details of each step are as follows:

Step 1: The network is divided into small pixels according to the location of the PQ meters by using **Algorithm IV**.

With the GIS-based graph division, the PQ meters located in the same pixels are merged as a group of meters.

Step 2: The PQ meters are allocated in pixels in the PQ data matrix in accordance with the following two rules: the merged PQ meters should be allocated in a pixel in the PQ data matrix, and the PQ meters from the neighboring areas should be allocated next to each other in the PQ data matrix.

Step 3: Steps 1–2 are repeated until all PQ meters are determined in the PQ data matrix.

Step 4: A split window consisting of neighboring pixels of the PQ data matrix is utilized to construct a regional PQ data sample. For example, one split window consists of nine nearest neighboring pixels in Fig. 4. The augmented matrix is obtained by combining the environmental factors. This matrix serves as the data source $\mathbf{\Omega}$ for real-time correlation analysis. The details of this step are as follows:

A split window covers δ pixels, and the regional PQ data sample $\mathbf{U}$ is formulated as follows:

$$\mathbf{U} = \left[\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \cdots, \mathbf{u}^{(t)}\right] + \boldsymbol{\eta}_u, \tag{9}$$

where $\mathbf{u}^{(k)} = [u_1^{(k)}, u_2^{(k)}, \cdots, u_\delta^{(k)}]^{\mathrm{T}}$ denotes the vector of the number of PQ records at time step k, and t is the study duration. $\boldsymbol{\eta}_u$ is the real-time PQ data uncertain matrix, which satisfies normal distribution. The sequence of related factors is formulated as follows:

$$\mathbf{f} = \left[f^{(1)}, f^{(2)}, \cdots, f^{(t)}\right], \tag{10}$$

where $f^{(k)}$ is the value of the selected factors at time step k.

An augmented matrix is constructed to analyze the correlation between the PQ records and the selected factors. The dataset matrix is formulated as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{U} \\ \mathbf{F} \end{bmatrix}, \tag{11}$$

where $\mathbf{F}$, transformed from $\mathbf{f}$, is denoted as

$$\mathbf{F} = \begin{bmatrix} \mathbf{f} \\ \mathbf{f} \\ \vdots \\ \mathbf{f} \end{bmatrix}_{\delta \times t} + \boldsymbol{\eta}_f, \tag{12}$$

where $\boldsymbol{\eta}_f$ is the measurement error matrix, which is represented as a white Gaussian noise matrix.

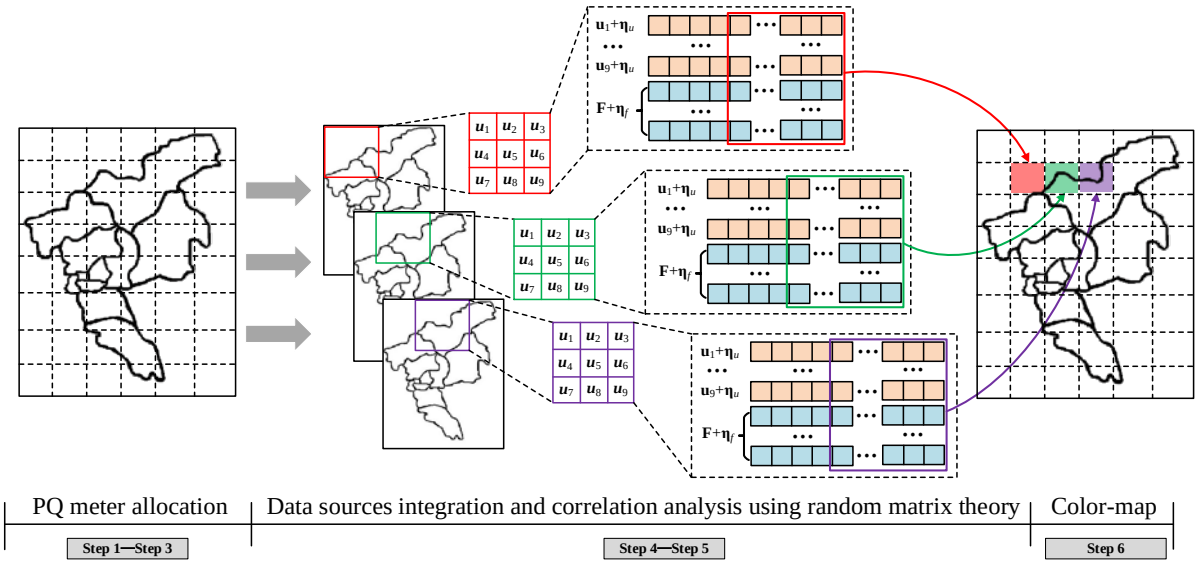


Fig. 4. Cross-correlation analysis and visualization based on RMT.

Step 5: The cross-correlation indicator is calculated in this step on the basis of the data source $\mathbf{\Omega}$. The details of the calculation are illustrated as follows:

A raw data matrix $\hat{\mathbf{X}} \in \mathbb{C}^{H \times T_w}$ is obtained from $\mathbf{\Omega}$ [27]. $\hat{\mathbf{X}}$ is converted into a normalized non-Hermitian matrix $\tilde{\mathbf{X}}$ by using (13).

$$\tilde{x}_{i,j} = (\hat{x}_{i,j} - \mu(\hat{\mathbf{x}}_i)) \frac{\sigma(\tilde{\mathbf{x}}_i)}{\sigma(\hat{\mathbf{x}}_i)} + \mu(\tilde{\mathbf{x}}_i), \tag{13}$$

where $\hat{\mathbf{x}}_i = [\hat{x}_{i,1}, \hat{x}_{i,2}, \cdots, \hat{x}_{i,T_w}]$, $\mu(\tilde{\mathbf{x}}_i) = 0$, and $\sigma(\tilde{\mathbf{x}}_i) = 1$ for $i = 1, 2, \cdots, H$, and $j = 1, 2, \cdots, T_w$. The singular value equivalent $\tilde{\mathbf{X}}_u$ of $\tilde{\mathbf{X}}$ is calculated with a Haar unitary matrix [27]. The matrix product for multiple arbitrarily assigned normalized non-Hermitian matrices $\tilde{\mathbf{X}}_i$ $(i = 1, 2, \ldots, L)$ is obtained using $\hat{\mathbf{Z}} = \prod_{i=1}^{L} \tilde{\mathbf{X}}_{u,i}$ $(i = 1, 2, \ldots, L)$. For simplicity, L is set equal to one in this study. The standard matrix product $\tilde{\mathbf{Z}}$ is converted from $\hat{\mathbf{Z}}$ by using the following equation:

$$\tilde{\mathbf{z}}_i = \frac{\hat{\mathbf{z}}_i}{\sqrt{H}\sigma(\hat{\mathbf{z}}_i)}, \tag{14}$$

where $\tilde{\mathbf{z}}_i = [\tilde{z}_{i,1}, \tilde{z}_{i,2}, \cdots, \tilde{z}_{i,H}]$, and $\hat{\mathbf{z}}_i = [\hat{z}_{i,1}, \hat{z}_{i,2}, \cdots, \hat{z}_{i,H}]$.

The index of MSR often achieves an improved performance for dealing with the asymptotic distribution of the eigenvalues of large rectangular random matrices. The MSR for the corresponding standard matrix product $\tilde{\mathbf{Z}}$ is formulated as follows:

$$\kappa_{\mathrm{MSR}} = \frac{1}{H} \sum_{i=1}^{H} |\lambda_{\tilde{\mathbf{Z}},i}|, \tag{15}$$

where $\lambda_{\tilde{\mathbf{Z}},i} (i = 1, 2, \cdots, H)$ represent the eigenvalues of $\tilde{\mathbf{Z}}$, and $|\lambda_{\tilde{\mathbf{Z}},i}|$ is the radius of the eigenvalue $\lambda_{\tilde{\mathbf{Z}},i}$ on the

complex plane. The ring law [23] stipulates that the empirical spectrum density (ESD) almost surely converges to $f_{\mathrm{ESD}}(\lambda_{\tilde{\mathbf{Z}}})$ to the limit given as:

$$f_{\mathrm{ESD}}(\lambda_{\tilde{\mathbf{Z}}}) = \begin{cases} \dfrac{1}{\pi\gamma L} |\lambda|^{(2/L-2)}, (1-\gamma)^{L/2} \le |\lambda| \le 1 \\ 0, \qquad\qquad \text{otherwise} \end{cases}, \quad (16)$$

as $H, T_w \to \infty$ with the ratio $H/T_w = \gamma \in (0,1]$. If the size of the raw data matrix is determined, then the ESD limits can be predicted. In this study, if $\kappa_{\mathrm{MSR}} < (1-\gamma)^{L/2}$, then the corresponding pixels in the GIS-based graph are determined as the PQ sensitive regions. Although the asymptotic convergence in RMT is considered under infinite dimensions, the asymptotic results are remarkably accurate for relatively moderate matrix sizes, such as tens [23]. On this basis, the correlation indicator between PQ events and selected factors is designed as (17) to assess the sensitivity. The indicator definition is formulated as follows:

$$d_\kappa = \begin{cases} 0, & (1-\gamma)^{L/2} \le \kappa_{\mathrm{MSR}} \le 1 \\ \dfrac{(1-\gamma)^{L/2} - \kappa_{\mathrm{MSR}}}{(1-\gamma)^{L/2}}, & \kappa_{\mathrm{MSR}} < (1-\gamma)^{L/2} \end{cases}, \quad (17)$$

where d_κ is the correlation indicator, and $(1-\gamma)^{L/2}$ is the limit of ESD.

Step 6: The magnitude of the correlation indicator is distinguished using colors. Steps 4–5 are repeated until the split window location is at the end of the PQ data matrix. The step size of the split-window's movement across the input is set to one in the study.

As mentioned in the previous discussion, **Algorithm V**, which is used to visualize the correlation analytical results of the PQ data, is illustrated as follows.

Algorithm V: Visualizing Correlation Analytical Results

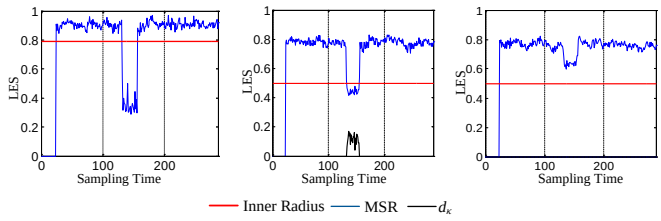
Input: PQ meters c_i $(i = 1,2,\ldots,n_c)$, number of PQ meters n_s, pixels q_i $(i = 1,2,\ldots,n_q)$, number of pixels n_q, value of selected factors $f^{(k)}(k = 1,2,\ldots,t)$, and number of PQ records $u_{ci}^{(k)}(k = 1,2,\ldots,t)$.

Output: Color-map representing the correlation between PQ data and factors.

/* Determine PQ data matrix */
Initialize: the set of PQ meters in each pixel $S_{qi} = \{\}$ $(i = 1,2,\ldots,n_q)$ and the PQ data matrix elements (i.e., total number of PQ records in each pixel $u_{sum,qi}^{(k)} = 0$)
for $i = 1$ **to** n_c
  for $j = 1$ **to** n_q
    if $L(c_i) \in q_j$ **then**
      Save the PQ meter $S_{qj} = S_{qj} + \{c_i\}$
      Update $u_{sum,qj}^{(k)} = u_{sum,qj}^{(k)} + u_{ci}^{(k)}$ $(k = 1,2,\ldots,t)$
    end if
  end for
end for
/* Obtain PQ data correlation indictor */
For each regional PQ data sample
  Form the regional PQ data sample $\mathbf{U} \in \mathbb{C}^{\delta \times t}$ by using $u_{sum,qi}^{(k)}$ and selected factor vector $\mathbf{f} \in \mathbb{C}^{1 \times t}$
  Construct an augmented matrix $\mathbf{A} \in \mathbb{C}^{2\delta \times t}$ by using (9) to (12)
  Acquire the standard matrix product $\tilde{\mathbf{Z}}$ from $\mathbf{A}$
  Calculate the κ_{MSR} from $\tilde{\mathbf{Z}}$ by using (15)
  Acquire the correlation indictor d_κ by using (17)
return cross-correlation results with color-map

A simple analytical example is designed to verify the effectiveness of the proposed cross-correlation analytical method. In this example, the status matrix consists of a frequency of PQ disturbances from nine pixels. The sampling rate is five minutes. Let $\delta = 9$, $t = 288$, $H = 9$, $T_w = 24$, and the parameters of PQ data uncertain matrix and weather data measurement error matrix are $\mu(\boldsymbol{\eta}_u) = 0.03$, $\sigma(\boldsymbol{\eta}_u) = 0.01$, $\mu(\boldsymbol{\eta}_f) = 3$ and $\sigma(\boldsymbol{\eta}_f) = 1$.

Fig. 5(a) shows that the U-shaped curve is from sampling time t_p = 132 to t_p = 156. This outcome indicates abnormal PQ events occurring at t_p = 132. Fig. 5(b) manifests that when we augment the weather conditions, such as the real-time regional rainfall, the minimum κ_{MSR} inside the abnormal PQ event duration is 0.421, which is deviated from the predicted ring, and $d_\kappa = 0.170$. This manifestation indicates correlations between the weather conditions and the abnormal PQ events. By contrast, Fig. 5(c) presents that when we augment the load data of the corresponding area, κ_{MSR} remains in the predicted ring throughout the abnormal PQ event duration. This outcome indicates poor correlations between the load data and the PQ events. Therefore, the severe weather condition is the main cause of these abnormal events.



(a) (b) (c)

Fig. 5. $\kappa_{\mathrm{MSR}} - t$ curves of various data sources: (a) status matrix; (b) augmented matrix, including weather conditions; (c) augmented matrix, including regional loads.

III. Case Studies

A. Data characteristics

The input data consist of the real-world PQ and power-network operation data measured in a practical power grid in a city. Tables I and II present the detailed information on the tested data and its structure.

TABLE I
Testing Data

Data type	PQ data	Power-network operation data
Time (month/day/year)	1/1/15–12/31/17	7/1/16–12/31/16
Number of records	30,053	404,684,813

TABLE II
Data Structure

PQ data structure				
Record No.	Substation	Types of PQ	Voltage and current waveforms	Time stamp
Power-network operation data structure				
Operation No.	Substation	Content of device alarms and operations		Time stamp

The PQ meters are located at high-voltage, medium-voltage, and low-voltage substations covering the entire region of the city. The power-network operation data were collected from equipment monitoring and energy management system. The following equipment information is leveraged for matching the PQ data: 1) power line, transformer, and bus protective device tripping message; 2) power line, transformer, and bus devices failure alarms; and 3) action information of auxiliary equipment, such as shunt capacitor bank and transformer taps. Thus, eight types of underlying causes are analyzed. All simulations were performed using an Intel Core i7-5500U 2.4 GHz CPU with 16 GB RAM.

B. Data preprocess

Modules I and II are used to classify the possible identical-source PQ disturbance records and delete duplicate data. Considering that the occurrence time of identical-source PQ disturbance records is close to each other, the range of correlation thresholds is set to ϕ_{min} = 50 ms and ϕ_{max} = 1000ms. The Module I blocking phase is conducted through minimizing (1), and 11 Pareto optimal solutions are shown in Table III. The fuzzy satisfying method is utilized to select the best solution among the obtained Pareto optimal sets. The last column of Table III clearly indicates that the optimal solution is #3, with the maximum weakest membership function of 0.713. The optimal correlation threshold is ϕ_{op} = 580 ms, and the index of pair completeness and reduction ratio are $PC(\boldsymbol{B}) = 0.983$ and $RR(\boldsymbol{B}) = 0.063$, respectively.

TABLE III
PARETO OPTIMAL SOLUTION OF THE EFFECTIVENESS AND EFFICIENCY FOR BLOCKING PQ RECORDS

#	W_1	W_2	$J_1 = 1/PC(\boldsymbol{B})$	$J_2 = 1/RR(\boldsymbol{B})$	$J_{1,pu} = \frac{J_{1,\max} - J_1}{J_{1,\max} - J_{1,\min}}$	$J_{2,pu} = \frac{J_{2,\max} - J_2}{J_{2,\max} - J_{2,\min}}$	min ($J_{1,pu}$, $J_{2,pu}$)
1	1	0	1.008	47.619	1.000	0.000	0.000
2	0.9	0.1	1.010	23.256	0.990	0.547	0.547
3	**0.8**	**0.2**	**1.017**	**15.873**	**0.955**	**0.713**	**0.713**
4	0.7	0.3	1.080	11.765	0.653	0.805	0.653
5	0.6	0.4	1.094	7.407	0.584	0.903	0.584
6	0.5	0.5	1.116	6.061	0.478	0.933	0.478
7	0.4	0.6	1.138	4.566	0.374	0.967	0.374
8	0.3	0.7	1.175	4.525	0.193	0.968	0.193
9	0.2	0.8	1.199	3.953	0.077	0.981	0.077
10	0.1	0.9	1.209	3.497	0.028	0.991	0.028
11	0	1	1.215	3.086	0.000	1.000	0.000

To show the advantage of duplicate data detection method proposed in Module II, more comparisons are performed. The proposed 2D representation-based method and wavelet transformation (WT)-based method [29] are leveraged to detect duplicate records. The two types of duplicate samples are generated as follows.

1) *Single PQ Disturbance:* seven types of single PQ disturbances, i.e., voltage sag, voltage swell, voltage interruption, impulsive transient, oscillation transient, harmonic, and flicker are randomly generated.

2) *Complex PQ Disturbance:* combinations of two disturbances, e.g., voltage sag and harmonic in Fig. 2(b), or combinations of three disturbances, e.g., voltage sag, harmonic, and voltage fluctuations in Fig. 2(c), are randomly generated.

Each pair of duplicate PQ data samples differs in beginning time and sampling frequency. The results are shown in Table IV, and a few conclusions can be drawn.

TABLE IV
COMPARISON OF DIFFERENT DUPLICATE RECORD DETECTION METHODS

Disturbance type	2D representation based method		WT based method	
	DA	ACT	DA	ACT
	(%)	(seconds)	(%)	(seconds)
Single PQ disturbance	100	0.118	100	0.178
Complex PQ disturbances	99.3		98.7	

1) The detection accuracy (DA) of 2D representation-based method is higher than that of the WT-based method. The main reason is that the different sampling frequencies of the duplicate disturbance data affect the results of wavelet decomposition.

2) The proposed method has shorter average computational time (ACT) than the WT-based method due to the former's ability to process three-phase voltage data simultaneously; thus, analysis of each phase data one by one is avoided.

On the basis of Modules I and II, the results of real-world PQ data cleaning are shown in Fig. 6. The temporal distribution of PQ records for the entire region of the city shows a remarkable difference from one month to another. A total of 11,213, 10,071, and 5,639 PQ disturbance records are measured by PQ meters after data cleaning in 2015, 2016, and 2017, respectively. The various possible reasons for the decrease of PQ records over the years include the following: 1) the increase in the number of individual PQ compensation devices (i.e., static VAR compensator and dynamic voltage

restorer) improves voltage conditions and reduces the total harmonic distortion at individual delivery points; 2) the PQ-record false-positive rate is reduced with old meters replaced by advanced meters.

Modules I and II can effectively delete duplicate PQ data. Moreover, the accuracy of the PQ data depends on the behavior of the PQ metering devices located in the transducers and signal conditioning (T&C) and analog-to-digital converter (ADC) stages. Thus, considering the uncertainties introduced by each component of the measurement system and the propagation of their effects through the measurement chain is necessary.

The T&C block is assumed to contribute to the measurement uncertainty with a gain error and a time delay, where η_{TG} and η_{TD} denote its probability distribution function. The ADC block contributes to the measurement uncertainty with the gain errors, offset, and quantization respectively denoted as η_{AG}, η_{AO}, and η_{AQ}, as shown in Fig. 7 [35].

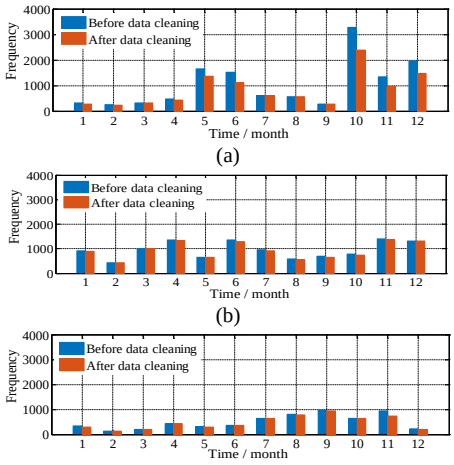


Fig. 6. Frequency of PQ disturbance events in (a) 2015, (b) 2016, and (c) 2017.

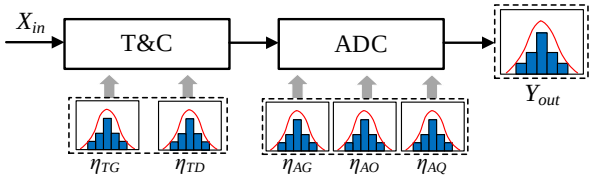


Fig. 7. Characterization of the uncertainty contributions in PQ meters

Assuming that the uncertainty of devices satisfies the normal distribution, the mathematical model for the estimation of uncertainty in the measurement of the PQ data is in the form of the multiplication product of particular variable quantities, as follows [36]:

$$\eta_u = \sqrt{\eta_{TG}^2 + \eta_{TD}^2 + \eta_{AG}^2 + \eta_{AO}^2 + \eta_{AQ}^2}\,, \tag{18}$$

where η_u is the probability distribution function of uncertain PQ data. For the difficulty of obtaining the uncertain parameters of each device, the combined uncertainty of the PQ data, as shown in Fig. 7, is calculated using the Monte Carlo method. To make the test environment similar to the site, input voltage signal X_{in} is set from 10% to 180% of primary rated voltage with a frequency band from DC to 12 kHz [35]. The uncertainties of PQ events frequency measurement are illustrated in Table V. Such uncertainties are obtained after calculating 10 times for each test. Therefore, the real-time PQ data uncertain matrix $\boldsymbol{\eta}_u$ used in the cross-correlation analytical visualization is related to the size of the split window and sampling rate.

TABLE V
EXPERIMENTAL RESULTS OF PQ DATA UNCERTAINTY

Number of Samples	1000	2000	3000	5000
$\mu(\boldsymbol{\eta}_u)$	3.1	5.8	9.1	15.2
$\sigma(\boldsymbol{\eta}_u)$	1.2	2.0	2.9	5.2

C. *Visualizing PQ disturbance events*

1) *Underlying cause-analysis visualization*

From December 1 to 31, 2016, 241 PQ disturbance events matched with power network operation records were calculated by Module III, as shown in Table VI. The matching results indicate that the total number of PQ events related to transmission line protection tripping is the largest, and these events are mostly recorded in Region #XI. Therefore, on the basis of the testing data, that line fault is the most probable cause of PQ disturbances. In addition, the results can remind the distribution network operators in Region #XI to increase the frequency of field investigations to improve the security and stability of power lines. Fig. 8 illustrates the spatial distribution of PQ events related to operation devices. To estimate the accuracy of the proposed scheme for the analysis of underlying causes, the relative directions of these 241 PQ disturbance events are calculated [7]. After comparing the device operation information and relative position of the disturbance sources, 23 disturbance events, which are inconsistent with the matching results, are found. Therefore, the accuracy of this method is estimated as follows: $(1-23/241)\% = 90.46\%$.

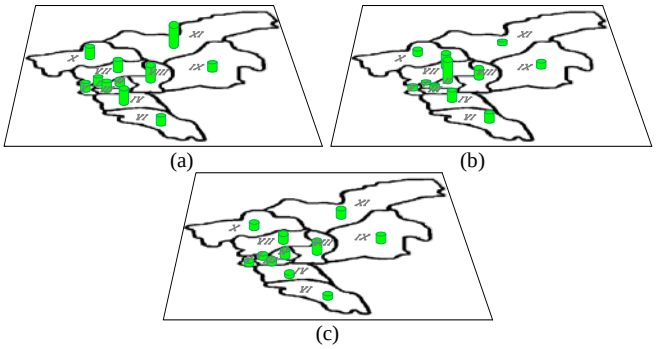


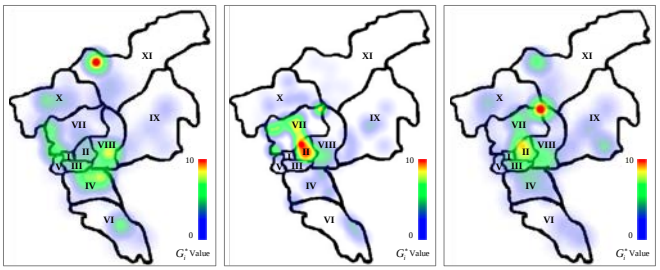
Fig. 8. Number of blocks of PQ disturbance records related to device operation. (a) Transmission line protection. (b) Transformer protection. (c) Bus protection.

TABLE VI
RESULTS OF UNDERLYING CAUSE ANALYSIS FOR PQ DISTURBANCE EVENT

Region	#I	#II	#III	#IV	#V	#VI	#VII	#VIII	#IX	#X	#XI
Transmission line protection	6	5	8	13	6	6	9	16	6	10	22
Transformer protection	0	20	2	8	2	7	15	10	6	4	3
Bus protection	2	6	3	3	1	3	8	12	6	5	8

2) Spatio–temporal distribution of PQ records

The GIS data and location of PQ meters are used in this phase to illustrate the spatio–temporal distribution of PQ records related to one specific device operation. The threshold τ is set to 2, and the visualization results are illustrated in Fig. 9. Compared with the spatial distribution of PQ events in Fig. 8, the GIS-based visualization can demonstrate the affected region of PQ events related to device faults. Specific PQ control devices can be accurately deployed based on these results.



(a) (b) (c)

Fig. 9. Spatial distribution of PQ disturbance records related to the operation of protection devices. (a) Transmission line protection. (b) Transformer protection. (c) Bus protection.

3) Correlation analysis visualization

Case I: Correlation analysis between weather condition and transmission line protection-related PQ records (07/01/16–07/31/16)

Fig. 10(a) demonstrates a strong correlation between PQ events and weather condition in Region #IX. Specifically, three weather conditions, namely, major storm disaster, adverse weather, and normal weather, are incorporated in this study [7]. These concepts require explanation due to the following reasons. 1) The total length of 220-kV overhead lines in Region #IX is larger than in other regions and 2) summer has many thunderstorms, which can easily cause an electrical short circuit and result in a PQ disturbance event. The maximum value of the correlation indicator in a weather-sensitive region, such as Region #IX, is 0.726.

Case II: Correlation analysis between temperature and PQ disturbance records (01/01/16–12/31/16)

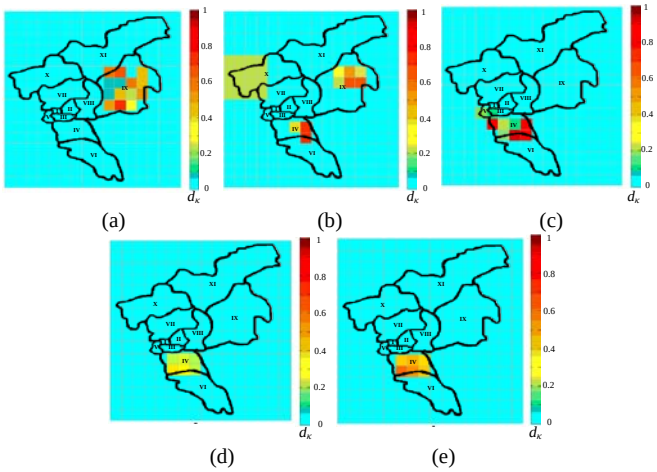
The correlation indicator for case II is shown in Fig. 10(b). This figure clearly demonstrates the temperature-sensitive region in the 2016. The results can help operators focus on the weak PQ regions and determine a control strategy of power converters to improve the PQ level. Fig. 10(b) also shows a strong correlation between PQ events in Regions IV, IX, and X and environmental temperature.

Case III: Real-time analysis of the correlation between temperature and PQ-disturbance records - 05/04/15 20:00:00.

Different from Cases I and II, the higher resolution data (e.g., the sampling rate is one hour) and moving-window technique [27] are utilized in Case III to achieve real-time correlation analysis. As shown in Fig. 10(c), when abnormal events destroy the original correlation, the corresponding area of abnormal event occurrence can be displayed using this method. Thus, detecting abnormal events promptly and assessing the severity of events (e.g., the impact area of an event) using this method are inconvenient for operators. The threshold τ is set to 5 in this section to reduce computational burden, and the PQ data uncertain matrix and temperature measurement error matrix satisfies the following parameters $\mu(\boldsymbol{\eta}_u) = 0.18$, $\sigma(\boldsymbol{\eta}_u) = 0.06$, $\mu(\boldsymbol{\eta}_f) = 2$, $\sigma(\boldsymbol{\eta}_f) = 0.1$.

Case IV: Correlation analysis between total current distortion rate and load level - 06/01/17

On the 1st June 2017, a severe current harmonic event recorded by the PQ meters occurred at Region #IV. Figs. 10(d) and 10(e) demonstrate that the total current distortion rate has a stronger correlation with substation load level compared with the total regional load level. The on-site detection shows that the harmonic current source of Region #IV is an industrial user, powered by the special line of the analyzed substation. A high load level results in large current distortion rate. Therefore, the proposed visualization method can be used to locate the PQ disturbance source.



(a) (b) (c) (d) (e)

Fig. 10. PQ data correlation analysis visualization. (a) Case I. (b) Case II. (c) Case III. (d) Case IV: the selected factor is the total regional load level. (e) Case IV: the selected factor is the substation load level.

The performance of the proposed method was compared with that of existing approaches, such as the Spearman method and the entropy method [7], in terms of the sensitivity to correlation analysis (i.e., variance values of the correlation coefficients) based on Case I. To obtain comparable results, the average values of correlation coefficients between PQ records and weather conditions in each pixel are calculated using existing approaches. The variance values of the correlation coefficients with variable pixels in one split window are presented in Table VII, from which the RMT-

based method has higher variance values than other methods. The number of pixels in each split window is set to nine, considering the computational burden. The results of the actual power system case validate the application of the proposed method in analyzing the correlation between the PQ level and various factors. Furthermore, the RMT-based method is more robust in dealing with data containing uncertainties and measurement errors [23].

TABLE VII
SENSITIVITY COMPARISON AMONG CORRELATION ANALYSIS METHODS

Methods	1 pixel	9 pixels	25 pixels
Spearman	0.3182	0.2041	0.1236
Entropy	0.3641	0.3054	0.2124
Random matrix theory	0.2179	0.4324	0.4082

D. Comparison with existing data-driven method

To benchmark the performance of the RMT-based real-time detection analysis against the existing MWPCA [25], the case studies presented in the previous section were used. Specifically, a sample of 50 PQ disturbance events caused by transmission line faults was used in a comparative study; $H = 9$, $T_w = 24$, and the sampling rate is five minutes. Tables VIII and IX provide a summary of the performance of RMT- and MWPCA-based detection methods for abnormal events. The number of accurate detections expressed as a percentage of the total number of events, that is, DA, is presented. Specifically, the impact of pixel size on the performance of RMT and MWPCA is presented in Table VIII. Moreover, to illustrate the impact of window size T_w on performance, a comparative study for the same case with the fixed pixel size, that is, $\tau = 5$, is presented in Table IX.

TABLE VIII
COMPARISONS OF RMT AND MWPCA WITH VARIABLE PIXEL NUMBER

Methods	RMT			MWPCA		
Threshold τ	5	10	15	5	10	15
Number of Pixels	166	121	76	166	121	76
DA (%)	100	100	96	100	96	88
ACT (seconds)	1.12	0.66	0.48	1.07	0.62	0.44

TABLE IX
COMPARISONS OF RMT AND MWPCA WITH VARIABLE WINDOW SIZE

Methods	RMT			MWPCA		
Window Size T_w	12	24	36	12	24	36
Fault Detection Time (seconds)	− 37	− 37	− 37	− 25	− 25	− 25
DA (%)	98	100	100	96	100	100
ACT (seconds)	0.49	1.12	1.91	0.42	1.07	1.87
Sensitivity Measurement	√			×		
GIS-based Visualization	√			×		

1) *Abnormal event detection ability:* Table VIII shows that the RMT- and MWPCA-based approaches can detect the PQ disturbance events caused by transmission line faults. Moreover, with the increase in partition threshold, the total number of pixel decreases. The proposed RMT-based method has robust detection capacity under different visual map resolutions. Table IX shows that the window size has little effect on the detection ability of RMT- and MWPCA-based approaches. The reason is that most of the PQ events occur in a concentrated time, except for certain long-duration PQ events caused by cascading failures.

The fault detection time in Table IX is the average time for the RMT- and MWPCA-based approaches to detect transmission line faults following its occurrence. The precise onset of events is unknown due to the absence of ground truth information. Consequently, the reference time for event onsets is defined as the protection devices' operation time. The results show that RMT- and MWPCA-based approaches can sometimes detect the fault before this time. Although both approaches can strongly detect abnormal PQ disturbance events, the MWPCA-based method lacks an effective quantitative index and visualization scheme for illustrating the severity and location of abnormal events.

2) *Computation ability:* The computational complexity of the RMT-based method is approximately equal to the MWPCA-based method for the real-time correlation analysis. Tables VIII and IX illustrate that the average computational time is closely related to the number of pixels and size of real-time PQ data window. The details of computation complexity are discussed in the following section.

E. Computational Complexity

The major computational burden in the proposed data analysis and display framework results from the following: 1) duplicate record deletion, 2) PQ disturbance underlying cause analysis, and 3) PQ disturbance cross-correlation analysis.

1) *Duplicate record deletion computation complexity:* Assuming that the average number of records in one PQ data block is $\overline{w}$ and the total number of records is n_r , the proposed duplicate PQ disturbance record deleting algorithm has $O(n_r(\overline{w}-1)/2)$ complexity. The traditional naïve sorted-neighborhood method [30] has $O(n_r\overline{w})$ complexity for the same condition. Hence, the proposed data preprocess scheme has advantages on computational complexity compared with traditional methods.

2) *PQ disturbance cause analysis computation complexity:* Assuming the size of PQ data block and the candidate power operation collection are $\overline{w}$ and w_{OP} , the proposed entities matching algorithm has $O(\overline{w}w_{OP})$ complexity. Hence, computational complexity scales linearly with both size of PQ data and power operation data.

3) *PQ disturbance correlation analysis computation complexity:* For moderate number of pixels (e.g., nine) computational complexity is minor issue for RMT-based correlation analysis. For example, assuming that the split window size is $H \times T_w$, the RMT-based method using parallel algorithm [37] has $O(H^2 k_H)$ complexity for calculating eigenvalues in real-time correlation analysis, where k_H is the parameter of the simplified calculation model [37]. On the contrary, the MWPCA-based approach using NIPALS algorithm [38] has $O(H^2)$ complexity for estimating the largest principal components. Therefore, the computational complexity of the proposed method is approximately equal to that of the MWPCA-based method.

For larger data source, the relatively small size of the split window and advanced distributed computation technology can be utilized to reduce computational time [39-41].

IV. Conclusion

PQ disturbances in power grids are a concern for operators and customers. Our study investigated the spatio–temporal distribution of PQ records based on real-world data. We used three modules (I to III) to achieve PQ data cleaning and analyze the underlying cause of PQ disturbances. On the basis of these preprocessed data, Module IV visualized the correlation between PQ disturbance records and environmental factors on a city level combined with GIS-based data.

The case studies lead to the following conclusions. (i) A methodology based on entity matching can reliably establish the relationship between PQ events and device operations. (ii) The proposed GIS-based graph partitioning approach can adjust the visualization resolution to satisfy various application scenarios. (iii) Our visualization module can also be used to analyze and visualize the correlation between PQ disturbances and selected factors. (iv) The use of our proposed method allows the operators to detect and locate abnormal events in real time easily, and assess the severity of events. Moreover, the performance of the proposed method was better than that of other benchmarks in terms of the sensitivity to correlation analysis and abnormal event detection accuracy.

Reference

[1] S. Mishra, C. N. Bhende, B. K. Panigrahi., "Detection and Classification of Power Quality Disturbances Using S-Transform and Probabilistic Neural Network," *IEEE Trans. Power Del*. vol. 31, no. 1, pp. 280-287, Jan. 2008.

[2] J. D. L. Ree, V. Centeno, J. S. Thorp, and A. G. Phadke, "Synchronized phasor measurement applications in power systems," *IEEE Trans. Smart Grid*., vol. 1, no. 1, pp. 20-27, Apr. 2010.

[3] D. Perera, L. Meegahapola, S. Perera, *et al*., "Characterisation of flicker emission and propagation in distribution networks with bi-directional power flows", *Renewable Energy*., vol. 63, no. 8, pp. 172-180, Mar. 2014.

[4] A. A. P. Bíscaro, R. A. F. Pereira, M. Kezunovic, *et al*., "Integrated fault location and power-quality analysis in electric power distribution systems", *IEEE Trans. Power Del*. vol. 31, no. 2, pp. 428-436, Aug. 2016.

[5] S. Hasheminejad, S. Esmaeili, S. Jazebi., "Power quality disturbance classification using S-transform and hidden Markov model," *Electric Machines & Power Systems*, vol. 40, no. 10, pp. 1160-1182, Jul. 2012.

[6] R. F. Yuan, Q. Ai, X. He., "Research on dynamic load modelling based on power quality monitoring system", *IET Gener. Transm. Distrib*., vol. 7, no. 1, pp. 46-51, Jan. 2013.

[7] F. Xiao, Q. Ai, "Data-driven multi-hidden markov model-based power quality disturbance prediction that incorporates weather conditions," *IEEE Trans. Power Syst*. vol. 34, no. 1, pp. 402-412, Jan, 2019.

[8] H. Galhardas and D. Florescu. "An extensible framework for data cleaning," in *Proc. IEEE International Conference on Data Engineering*. San Diego, California, 2000, pp. 312.

[9] B. A. Carreras, D. E. Newman, I. Dobson *et al*., "Evidence for self-organized criticality in a time series of electric power system blackouts", *IEEE Trans. Circuits Syst. I, Reg. Papers*., vol. 51, no. 9, pp. 1733-1740, Sept. 2004.

[10] Y. Sun and T. J. Overbye, "Visualizations for power system contingency analysis data," *IEEE Trans. Power Syst*., vol. 19, no. 4, pp. 1859–1866, Nov. 2004.

[11] T. J. Overbye, D. A. Wiegmann, A. M. Rich, *et al*. "Human factors aspects of power system voltage contour visualizations," *IEEE Trans. Power Syst*., vol. 18, no. 1, pp. 76-82, Feb. 2003.

[12] J. Zhu, E. Zhuang, C. Ivanov, *et al*. "A data-driven approach to interactive visualization of power systems," *IEEE Trans. Power Syst*., vol. 26, no. 4, pp. 2539-2546, Nov. 2011.

[13] E. Gasch, J. Meyer, P. Schegner, K. Schmidt, "Web-based platform for exchanging harmonic emission measurements of electronic equipment," in *Proc. IEEE International Conference on Harmonics and Quality of Power*. Hong Kong, China, 2012, pp. 943-948.

[14] C. Kattmann, S. Tenbohlen. "Visualizatiion of power quality data," in *Proc. IEEE Manchester PowerTech*. Manchester, UK, 2017, pp. 1-5.

[15] E. Gomezlazaro, J. A. Fuentes, A. Molinagarcia, *et al*. "Characterization and visualization of voltage dips in wind power installations". *IEEE Trans. Power Del*., vol. 24, no. 4, pp. 2071-2078, Oct. 2009.

[16] D. Macaya, J. Meléndez, J. Sánchez, *et al*. "Visual management of sags and incidents gathered in distribution substations for power quality management," in *Proc. IEEE Electrical Power Quality and Utilisation*. Barcelona, Spain, 2007, pp. 1-4.

[17] A. Christe, S. Negrashov, P. Johnson. "Open power quality: an open source framework for power quality collection, analysis, visualization, and privacy," in *Proc. IEEE PES Innovative Smart Grid Technologies Conference*., Minneapolis, USA, 2016, pp. 1-5.

[18] A. Murata, H. Yamaguchi, and K. Otani, "A method of estimating the output fluctuation of many photovoltaic power generation systems dispersed in a wide area," *Elect. Eng. Jpn.* vol. 166, no. 4, pp. 9–19, Mar. 2009 [Online]. Available: http: //onlinelibrary. wiley. com/ doi/ 10. 1002/eej. 20723/abstract.

[19] T. Lu, Z. Wang, J. Wang, Q. Ai, and C. Wang, "A data-driven Stackelberg market strategy for demand response-enabled distribution systems," *IEEE Trans. Smart Grid*, 2019, 10, (3), pp.2345-2357.

[20] T. Lu and Q. Ai. "Interactive energy management of networked microgrids-based active distribution system considering large-scale integration of renewable energy resources," *Applied Energy*, vol. 163, pp. 408-422, 2016.

[21] M. Wulder, "Optical remote sensing techniques for assessment of forest inventory and biophysical parameters," *Progr. Phys. Geography*, vol. 22, no. 4, pp. 449–476, 1998.

[22] Z. Bai, and J. W. Silverstein, Spectral analysis of large dimensional random matrices, 2nd ed. Spring Street, NY, USA: Springer, 2010.

[23] X. He., Q. Ai., R. C. Qiu., et al.: "A big data architecture design for smart grids based on random matrix theory", *IEEE Trans. Smart Grid*., 2017, 8, (2), pp. 674-686.

[24] X. Liu., D. M. Laverty., R. J. Best., et al.: "Principal component analysis of wide-area phasor measurements for islanding detection—A geometric view," *IEEE Trans. Power Del*., vol. 30, no. 2, pp. 976–985, Apr. 2015.

[25] M. Rafferty., X. Liu., D. M. Laverty., et al.: "Real-time multiple event detection and classification using moving window PCA", *IEEE Trans. Smart Grid*, 2016, 7, (5), pp. 2537-2548.

[26] F. Xiao, Q. Ai., "Electricity Theft Detection in Smart Grid Using Random Matrix Theory[J]. *IET Gener. Transm. Distrib.*, vol. 12, no. 2, pp. 371-378, .Feb. 2018.

[27] X. Xu., X. He., Q. Ai., et al.: "A correlation analysis method for power systems based on random matrix theory", *IEEE Trans. Smart Grid.*, 2017, 8, (4), pp. 1811-1820.

[28] G. Papadakis., G. Koutrika., T. Palpanas., et al.: "Meta-blocking: taking entity resolutionto the next level", *IEEE Trans. Knowl Data En*., 2013, 26, (8), pp. 1946-1960.

[29] F. B. Costa, J. Driesen, "Assessment of voltage sag indices based on scaling and wavelet coefficient energy analysis," *IEEE Trans. Power Del*. vol. 28, no. 1, pp. 336-346, Jan. 2013.

[30] S. Yan, D. Lee, M. Y. Kan, et al., "Adaptive sorted neighborhood methods for efficient record linkage", in Proc, ACM/IEEE-CS Digital Libraries, Vancouver, BC, Canada, 2007, pp. 185-194.

[31] J. Aller., A. Bueno., T. Paga.: "Power system analysis using space vector transformation," *IEEE Trans. Power Syst.*, 2002, 17, (4), pp. 957–965.

[32] N. R. Watson., C. K. Ying., C. P. Arnold.: "A global power quality index for aperiodic waveforms", Ninth International Conference on Harmonics and Quality of Power. Proceedings. IEEE, 2000, 3, pp. 1029-1034.

[33] T. X. Zhu.: “Detection and characterization of oscillatory transients using matching pursuits with a damped sinusoidal dictionary,” *IEEE Trans. Power Del.*, vol. 22, no. 2, pp. 1093-1099, 2007.

[34] L. Huilian, J.V. Milanovic, R. Marcos, et al. “voltage sag estimation in sparsely monitored power systems based on deep learning and system area mapping,” *IEEE Trans. Power Del*., vol. 33, no. 6, pp. 3162-3172, Dec, 2018.

[35] A. Ferrero, S. Salicone.: “A Monte Carlo-like approach to uncertainty estimation in electric power quality measurements,” *COMPEL-The international journal for computation and mathematics in electrical and electronic engineering.*, vol. 23, no. 1, pp. 119-132, 2004.

[36] L. Arsov, M. Cundeva-Blajer, Z. Grkov, et al.: “Estimation of uncertainty in measurement of power quality characteristics with a virtual measurement instrument,” in *Proc. IEEE International Instrumentation and Measurement Technology*. Graz, Austria, 2012, pp. 2752-2757.

[37] K. B. Yu.: “Recursive updating the eigenvalue decomposition of a covariance matrix,” *IEEE Trans. Signal Processing.*, vol. 39, no. 5, pp. 1136-1145, 1991.

[38] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometr. Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

[39] Tianguang Lu, Qian Ai, Zhaoyu Wang.: "Interactive game vector: an operation-based pricing mechanism for smart distribution systems with coupled-microgrid". *Applied Energy*, 2018, vol. 212, pp. 1462-1475.

[40] Tianguang Lu, Zhaoyu Wang, Qian Ai, Wei-Jen Lee.: "Interactive model for energy management of clustered microgrids". *IEEE Transactions on Industry Applications*, 2017, no. 3, vol. 53, pp. 1739-1750.

[41] Tianguang Lu, Wei-Jen Lee, Qian Ai, Songtao Lu.: "A priority decision making-based biding strategy for interactive aggregators". *IEEE Transactions on Industry Applications*, 2018, no. 6, vol. 54, pp. 5569-5578.