IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 1, 2017, accepted February 19, 2017, date of publication March 22, 2017, date of current version May 17, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2679006

Real-Time Detection of Power System Disturbances Based on *k*-Nearest Neighbor Analysis

LIANFANG CAI[1], NINA F. THORNHILL[1], (Senior Member, IEEE), STEFANIE KUENZEL[2], (Member, IEEE), AND BIKASH C. PAL[3], (Fellow, IEEE)

[1]Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, U.K.
[2]Department of Electronic Engineering, Royal Holloway, University of London, London TW20 0EX, U.K.
[3]Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K.

Corresponding author: L. Cai (l.cai@imperial.ac.uk)

This work was supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/L014343/1.

ABSTRACT Efficient disturbance detection is important for power system security and stability. In this paper, a new detection method is proposed based on a time series analysis technique known as *k*-nearest neighbor (*k*NN) analysis. Advantages of this method are that it can deal with the electrical measurements with oscillatory trends and can be implemented in real time. The method consists of two stages, which are the off-line modeling and the on-line detection. The off-line stage calculates a sequence of anomaly index values using *k*NN on the historical ambient data and then determines the detection threshold. Afterward, the on-line stage calculates the anomaly index value of presently measured data by readopting *k*NN and compares it with the established threshold for detecting disturbances. To meet the real-time requirement, strategies for recursively calculating the distance metrics of *k*NN and for rapidly picking out the *k*th smallest metric are built. Case studies conducted on simulation data from the reduced equivalent model of the Great Britain power system and measurements from an actual power system in Europe demonstrate the effectiveness of the proposed method.

INDEX TERMS Disturbance detection, power system, security, stability, *k*-nearest neighbor (*k*NN), anomaly index, real-time.

I. INTRODUCTION

In the past decade, detecting power system disturbances has emerged as a new and promising research area, because efficient disturbance detection plays a crucial role in understanding the system behavior and improving the system operating stability margin. According to [1], a power system disturbance has been described as ''a sudden change or sequence of changes in one or more of the power system parameters'', which may be large or small. Large disturbances can stress the power system so severely that the stability is lost, while small disturbances may gently push the power system into another operating condition. Presently, disturbance detection and analysis are needed in Wide-Area Monitoring and Control Systems (WAMCS) which collect data using phasor measurement units (PMUs) from different locations with time synchronized through global positioning systems (GPS) [2].

Numerous advanced measurement instruments provide abundant measurements for the development of data-driven disturbance detection. Most of the reported data-driven approaches assume that a disturbance is distinct from the normal trend of measurements by its amplitude in the time domain or by its scale in the time-frequency domain. For instance, statistical analysis methods [3]–[6] make use of differences in time-domain amplitude, while the wavelet transform methods [7]–[9] usually exploit differences in scale to detect disturbances which map to the wavelet coefficients of high amplitude in the lower scales. However, such assumption cannot be well met all the time, as indicated in [10] and exemplified in [11], especially for the cases in power systems where electrical measurements often exhibit oscillatory or cyclical characteristics [12].

Recently, a univariate detection method based on a time series analysis technique known as *k*-nearest neighbor (*k*NN) analysis was presented in [10], framing disturbance detection as an anomaly detection problem and solving it with *k*NN. This method does not require the relative amplitude or the

wavelet coefficients of disturbances to be markedly different from the overall trend and thus is more generic. Soon afterwards, it was further extended into a multivariate detection method [11], since the identification of a disturbance can be difficult in the measurements of an individual variable with strong oscillatory trends and the presence of the same disturbance in the measurements of different variables can be jointly explored for an improved outcome. However, the methods in [10, 11] were developed for off-line analysis and cannot be effectively implemented in real time for an on-line application.

Motivated by the above analysis, in this work, our main contribution is to propose a detection method based on kNN, which can be performed online to detect power system disturbances in real time. The real-time implementation is achieved by constructing a recursive calculation strategy for the distance metrics of kNN and a fast selection strategy for the kth smallest metric. Another advantage of the proposed method is that it is capable of tackling the electrical measurements with oscillatory trends. Case studies on simulation data from the reduced equivalent model of Great Britain (GB) power system and measurements from an actual power system in Europe (called European power system here) are used to demonstrate the effectiveness of the proposed method.

The paper is organized as follows. Section II gives a brief description of the kNN method. Section III presents the real-time detection method based on kNN. The application results and analysis on the two case studies are provided in Section IV, while our conclusions are drawn in Section V.

II. THE kNN BASICS

In this section, the basics of the kNN method are briefly introduced to lay the foundation for the subsequent presentation of the kNN-based real-time detection method.

The kNN method has been widely applied for anomalous window detection [10], [11], [13]–[15]. It adopts a similarity metric to measure the distance between each window in a time series and the other windows. Windows with similar sequences of samples are called near neighbors. Anomalous windows are those distinct from the underlying trend of the time series. The distance of a window to its kth nearest neighbor, known as anomaly index, is the key of kNN to detect anomaly.

Similarity metrics reported in the literature include the Euclidean distance (ED) [10], [11], the cosine similarity (CS) [16], and the dynamic time warp (DTW) [17]. Among them, the ED is more commonly used because of its simplicity and good geometrical interpretation. It is defined as the 2-norm of the displacement vector between two points $\boldsymbol{p}$ and $\boldsymbol{q}$ in a L-dimensional space, which can be written as follows:

$$d(\boldsymbol{p},\boldsymbol{q}) \triangleq \left\|\boldsymbol{p}^{\mathrm{T}}-\boldsymbol{q}^{\mathrm{T}}\right\|_2 = \sqrt{\sum_{j=1}^{L}\left(p_j-q_j\right)^2} \tag{1}$$

where $\boldsymbol{p}^{\mathrm{T}} = [p_1 p_2 \cdots p_L]$, $\boldsymbol{q}^{\mathrm{T}} = [q_1 q_2 \cdots q_L]$, "T" denotes the transpose operator, $\|\bullet\|_2$ denotes the 2-norm of a vector, $d(\boldsymbol{p},\boldsymbol{q}) \geq 0$, and $d(\boldsymbol{p},\boldsymbol{q}) = 0$ indicates the maximum similarity occurring only when two windows are equal in all L samples, i.e., $\boldsymbol{p} = \boldsymbol{q}$.

Using (1) as the foundation, for a window in a time series, its anomaly index value can be calculated as the ED between this window and its kth nearest neighbor. The anomaly index value of an anomalous window will be significantly higher than that of any normal window, which is the reason why the kNN method can be used for anomaly detection in a time series.

The above is the brief description of the kNN method. Details about this technique can be further found in [10].

III. REAL-TIME DETECTION BASED ON kNN

If the measurements of an electrical variable are viewed as a time series, the detection of power system disturbances can be achieved by detecting anomalous windows in this time series. Thus, a real-time detection method based on kNN, referred to as RD-kNN, is proposed in this section. The RD-kNN method mainly includes two parts: (1) the off-line modelling; (2) the on-line detection. In the following, the RD-kNN method is presented in detail.

A. OFF-LINE MODELLING

The off-line modelling step calculates a sequence of anomaly index values by applying kNN on the measurements historically recorded under the ambient condition with no disturbance occurring. It then determines a detection threshold for online monitoring whether a power system disturbance affects an electrical variable or not.

More specifically, the symbol x_i denotes the ith measured electrical variable for monitoring (e.g., frequency, voltage, current, or power) and $x_{i,j}$ denotes the jth measurement of x_i at the jth sampling time point. For the variable x_i, the following matrix can be built using the dataset $\{x_{i,j}\}_{j=1}^{N}$ with N measurements:

$$\begin{aligned}\boldsymbol{X}_i &= \begin{bmatrix} \boldsymbol{x}_{i,1}^{\mathrm{T}} \\ \boldsymbol{x}_{i,2}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_{i,N-L+1}^{\mathrm{T}} \end{bmatrix} \\ &= \begin{bmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,L} \\ x_{i,2} & x_{i,3} & \cdots & x_{i,L+1} \\ \vdots & \vdots & \cdots & \vdots \\ x_{i,N-L+1} & x_{i,N-L+2} & \cdots & x_{i,N} \end{bmatrix}\end{aligned} \tag{2}$$

where $\boldsymbol{X}_i$ is called the embedding matrix of x_i in kNN, and its row $\boldsymbol{x}_{i,r}^{\mathrm{T}}$ denotes the rth window of the recorded dataset, while L denotes the number of measurements in each window.

Each row of $\boldsymbol{X}_i$ is then compared with the other rows, using the square of the ED (SED) as follows:

$$\begin{aligned} d_i^2\left(\boldsymbol{x}_{i,e},\boldsymbol{x}_{i,r}\right) &\triangleq \left\|\boldsymbol{x}_{i,e}^{\mathrm{T}}-\boldsymbol{x}_{i,r}^{\mathrm{T}}\right\|_2^2 \\ &= \sum_{j=1}^{L}\left(x_{i,e-j+L}-x_{i,r-j+L}\right)^2 \end{aligned} \tag{3}$$

The reason of using the SED rather than directly using the ED is for the convenience of the real-time on-line calculation, which can be obviously observed later in Section III-B. Accordingly, an anomaly index value $AI_{i,r}$ for the rth row $\boldsymbol{x}_{i,r}^{\mathrm{T}}$ is determined as given in [10]. It is the kth smallest SED between $\boldsymbol{x}_{i,r}^{\mathrm{T}}$ and all other rows except the near-in-time rows of $\boldsymbol{x}_{i,r}^{\mathrm{T}}$, where the near-in-time rows of $\boldsymbol{x}_{i,r}^{\mathrm{T}}$ are those having at least one sample in common with $\boldsymbol{x}_{i,r}^{\mathrm{T}}$, e.g., $\boldsymbol{x}_{i,L}^{\mathrm{T}}$ is the last near-in-time row of $\boldsymbol{x}_{i,1}^{\mathrm{T}}$. Here, the exclusion of the SEDs between $\boldsymbol{x}_{i,r}^{\mathrm{T}}$ and its near-in-time rows during the determination of $AI_{i,r}$ is to avoid treating such near-in-time rows as the near neighbors of $\boldsymbol{x}_{i,r}^{\mathrm{T}}$, as suggested in [11].

After each row of $\boldsymbol{X}_i$ in (2) gains its corresponding anomaly index value, it is necessary to determine a threshold for anomaly detection based on the obtained sequence of anomaly index values $\left\{AI_{i,r}\right\}_{r=1}^{N-L+1}$. As no prior knowledge is available with regard to the distribution of $\left\{AI_{i,r}\right\}_{r=1}^{N-L+1}$, the detection threshold AI_i^{α} with the confidence level α can be determined according to the strategy in [5], i.e., $(1-\alpha)\,(N-L+1)$ is rounded towards the nearest integer δ and the δth highest value of $\left\{AI_{i,r}\right\}_{r=1}^{N-L+1}$ is taken as AI_i^{α}.

In addition to the calculation of the above anomaly index values $\left\{AI_{i,r}\right\}_{r=1}^{N-L+1}$ and the related detection threshold AI_i^{α} for the individual electrical variable x_i, inspired by [11], the system-wide anomaly index values providing a global characterization of the group of variables can be calculated as:

$$AI_r = \frac{1}{m}\sum_{i=1}^{m}\left|AI_{i,r}\right|, \quad 1 <= r <= N-L+1 \tag{4}$$

where m denotes the total number of the measured electrical variables, and the system-wide detection threshold AI^{α} can be determined as the δth highest value of the obtained $\{AI_r\}_{r=1}^{N-L+1}$.

B. ON-LINE DETECTION

On completion of off-line modelling, on-line detection should be considered. Real-time calculation of the anomaly index value is of prime importance for on-line detection. It requires a strategy for recursively calculating the SED metric of kNN and another strategy for fast selection of the kth smallest SED. The specific details are as follows.

The symbol $\boldsymbol{x}_{i,p}^{\mathrm{T}} = \left[x_{i,p-L+1}\ x_{i,p-L+2}\ \cdots\ x_{i,p}\right]$ denotes the vector of the L continuous measurements newly collected from the variable x_i, where p represents the present sampling time point. In order to determine whether $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ is anomalous or not, the present anomaly index value $AI_{i,p}$ for $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ is defined as the kth smallest SED between $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ and all rows of $\boldsymbol{X}_i$ in (2). The reason for this definition is that all rows of $\boldsymbol{X}_i$ are normal windows with the ambient characteristic which can be used as the foundation for evaluating whether or not the newly obtained $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ deviates from normal. If $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ is anomalous, the SEDs between it and the rows of $\boldsymbol{X}_i$ will be large and the corresponding anomaly index value $AI_{i,p}$ will also be large and exceed the related detection threshold AI_i^{α}. For the rth row $\boldsymbol{x}_{i,r}^{\mathrm{T}}$ of $\boldsymbol{X}_i$, the SED between it and $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ can be expressed as:

$$\begin{aligned} d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right) &\triangleq \left\|\boldsymbol{x}_{i,p}^{\mathrm{T}}-\boldsymbol{x}_{i,r}^{\mathrm{T}}\right\|_2^2 \\ &= \sum_{j=1}^{L}\left(x_{i,p-j+1}-x_{i,r-j+L}\right)^2 \end{aligned} \tag{5}$$

From (5), it can be seen that the calculation of $d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)$ requires $2L-1$ additions and L multiplications. So, for r from 1 to $N-L+1$, the total number of additions is $(N-L+1)\,(2L-1)$ and the total number of multiplications is $(N-L+1)\,L$. Usually, this is not a problem since high performance processors are widely used in modern monitoring systems. However, if the size N of the dataset and the window length L are large, the on-line computation load should be taken into consideration. To better meet the real-time requirement, a recursive calculation strategy for the SED metric which can significantly reduce the number of operations needed in (5), called *Strategy* Γ here, is developed by making use of previously-calculated results.

1) *Strategy* Γ for recursively calculating the SED metric

For the vector $\boldsymbol{x}_{i,p-1}^{\mathrm{T}} = \left[x_{i,p-L}\ x_{i,p-L+1}\ \cdots\ x_{i,p-1}\right]$ that is obtained a sampling time point earlier than the vector $\boldsymbol{x}_{i,p}^{\mathrm{T}}$, the SED between it and the $(r-1)$th row $\boldsymbol{x}_{i,r-1}^{\mathrm{T}}$ of $\boldsymbol{X}_i$ can be written as follows:

$$\begin{aligned} d_i^2\left(\boldsymbol{x}_{i,p-1},\boldsymbol{x}_{i,r-1}\right) &\triangleq \left\|\boldsymbol{x}_{i,p-1}^{\mathrm{T}}-\boldsymbol{x}_{i,r-1}^{\mathrm{T}}\right\|_2^2 \\ &= \sum_{j=1}^{L}\left(x_{i,p-j}-x_{i,r-j+L-1}\right)^2 \end{aligned} \tag{6}$$

Combining (5) and (6), the following expression can be obtained:

$$\begin{aligned} &d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)-d_i^2\left(\boldsymbol{x}_{i,p-1},\boldsymbol{x}_{i,r-1}\right) \\ &= \left\|\boldsymbol{x}_{i,p}^{\mathrm{T}}-\boldsymbol{x}_{i,r}^{\mathrm{T}}\right\|_2^2-\left\|\boldsymbol{x}_{i,p-1}^{\mathrm{T}}-\boldsymbol{x}_{i,r-1}^{\mathrm{T}}\right\|_2^2 \\ &= \left(x_{i,p}-x_{i,r-1+L}\right)^2-\left(x_{i,p-L}-x_{i,r-1}\right)^2 \end{aligned} \tag{7}$$

From (7), it can be seen that $d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)$ can be recursively calculated from $d_i^2\left(\boldsymbol{x}_{i,p-1},\boldsymbol{x}_{i,r-1}\right)$ by using the term $\left(x_{i,p}-x_{i,r-1+L}\right)^2-\left(x_{i,p-L}-x_{i,r-1}\right)^2$. The recursive calculation of $d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)$ requires four addition and two multiplication operations. Since the window length L in (5) is usually much larger than four, the on-line computational load of (5) can be reduced significantly by this recursive calculation, which is beneficial to the real-time implementation efficiency. Besides, it should be noted that (7) is subject to the condition $r>=2$. For the case $r=1$, $d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)$ cannot be recursively calculated and can only be calculated by (5). Thus, the strategy for recursively calculating $d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)$ can be finally expressed as the following formula:

$$d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)=\begin{cases} d_i^2\left(\boldsymbol{x}_{i,p-1},\boldsymbol{x}_{i,r-1}\right)+\left(x_{i,p}-x_{i,r-1+L}\right)^2 \\ \qquad -\left(x_{i,p-L}-x_{i,r-1}\right)^2, & r>=2 \\ \sum_{j=1}^{L}\left(x_{i,p-j+1}-x_{i,r-j+L}\right)^2, & r=1 \end{cases} \tag{8}$$

From (8), the reason why SED rather than ED is used can be easily understood, which is mainly due to the consideration of the convenience of the on-line recursive calculation.

Using (8), the SED sequence $\left\{d_i^2\left(\boldsymbol{x}_{i,p},\boldsymbol{x}_{i,r}\right)\right\}_{r=1}^{N-L+1}$ can be obtained online with high efficiency. Then, the present anomaly index value $AI_{i,p}$ for $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ can be determined as the kth smallest element of this sequence. Because the selection of the kth smallest SED also has an effect on the real-time performance, a selection strategy, called *Strategy* ΓΓ here, is constructed for fast selecting the desired value.

2) *Strategy* ΓΓ for fast selection of the kth smallest SED

An intuitive way of selecting the kth smallest element in a sequence is to firstly sort this sequence into ascending order and then select the kth element from the sorted sequence. This can be easily and rapidly attained through the existing software, such as the built-in function 'sort' in MATLAB developed by the MathWorks company. However, since only the kth smallest element of a sequence is required, it is not necessary to sort the whole sequence. That is, if k elements are obtained from a sequence which are smaller than the rest of this sequence, only the order of these k elements needs to be concerned about and the maximum one of these k elements is exactly the kth smallest element of the entire sequence.

Strategy ΓΓ is developed directly from the above consideration. Specifically, an array which can hold k ordered elements is set up and the first k elements of the SED sequence are put into this array after they are sorted in ascending order. Then, the remaining $N-L+1-k$ elements of the SED sequence are fetched one by one and compared to the elements of the ordered array. If the fetched element is larger than the maximum element of the ordered array, the fetched element is removed and the ordered array remains unchanged; otherwise, the maximum element of the ordered array is removed and the fetched element is inserted into the array ensuring that all k elements in the updated array are still in the ascending order. After each of the remaining elements of the SED sequence is dealt with by this process, the maximum element of the ultimately obtained array is exactly the kth smallest SED. An illustrative description of *Strategy* ΓΓ is shown in Fig. 1, where the elements of the ordered array are denoted by the symbols $d_i^{2(1)}, d_i^{2(2)}, \cdots, d_i^{2(k)}$, and a fetched element from the remaining $N-L+1-k$ elements of the SED sequence is denoted by the symbol $d_i^{2(*)}$.

The comparison between a fetched element $d_i^{2(*)}$ and the elements $d_i^{2(1)}, d_i^{2(2)}, \cdots, d_i^{2(k)}$ of the ordered array is called a round of comparison here. Referring to Fig. 2, after a round of comparison, one of the following three scenarios will occur:

(1) $d_i^{2(*)} > d_i^{2(k)}$: the ordered array remains unchanged and $d_i^{2(*)}$ is removed.

(2) $d_i^{2(*)} < d_i^{2(1)}$: $d_i^{2(*)}$ is put in front of $d_i^{2(1)}$ and $d_i^{2(k)}$ is removed.

(3) $d_i^{2(j-2)} <= d_i^{2(*)} <= d_i^{2(j-1)}$, where j denotes an integer between 3 and $k+1$: $d_i^{2(*)}$ is inserted between $d_i^{2(j-2)}$ and $d_i^{2(j-1)}$, and $d_i^{2(k)}$ is removed.

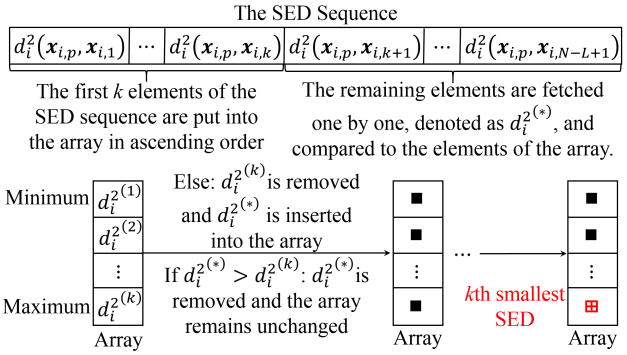

FIGURE 1. ***Strategy* ΓΓ for fast selection of the *k*th smallest SED.**

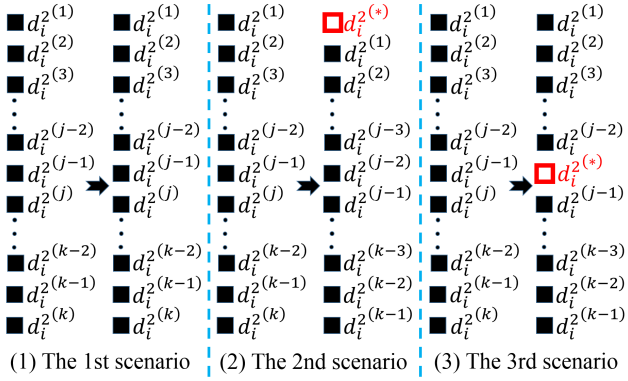

FIGURE 2. **An illustration of three different scenarios after a round of the comparison.**

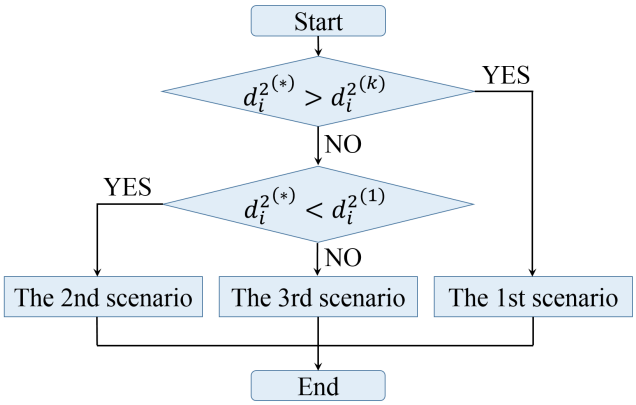

FIGURE 3. **The flow chart of a round of comparison.**

Whether a fetched element $d_i^{2(*)}$ is removed according to the 1st scenario or is inserted in the array according to the 2nd or the 3rd scenario, the elements in the updated array are still denoted as $d_i^{2(1)}, d_i^{2(2)}, \cdots, d_i^{2(k)}$. The flow chart of a round of comparison is shown in Fig. 3. It can be seen that, a fetched element only needs to be compared with one element of the ordered array in the 1st scenario or two elements of the ordered array in the 2nd scenario. However, in the 3rd scenario, a fetched element needs to be compared with all k

elements of the ordered array for the worst case, e.g., the case when $d_i^{2(k-1)} <= d_i^{2(*)} <= d_i^{2(k)}$ and $d_i^{2(2)}, d_i^{2(3)}, \cdots, d_i^{2(k-1)}$ are compared with $d_i^{2(*)}$ in turn. So, a relatively small value of the parameter k is beneficial to the real-time implementation.

To further reduce the number of comparisons for the 3rd scenario, the idea of binary search [18] is introduced to search the target position in the ordered array for $d_i^{2(*)}$ where $d_i^{2(j-2)} <= d_i^{2(*)} <= d_i^{2(j-1)}$ can be met with j denoting an integer between 3 and $k+1$. The binary search begins by comparing $d_i^{2(*)}$ to the middle element of the ordered array. If $d_i^{2(*)}$ is smaller than or equal to the middle element, then the search continues on the former half of the ordered array; otherwise, the search continues on the latter half of the ordered array. The search continues, eliminating half of the elements, and comparing $d_i^{2(*)}$ to the middle element of the remaining elements, until the target position in the array is found. Here, the number of comparisons in the 3rd scenario is $\log_2(k)$ at most, which is smaller than k. An illustrative example of this binary search process is shown in Fig. 4 for an intuitive observation, where the parameter k is equal to 5.

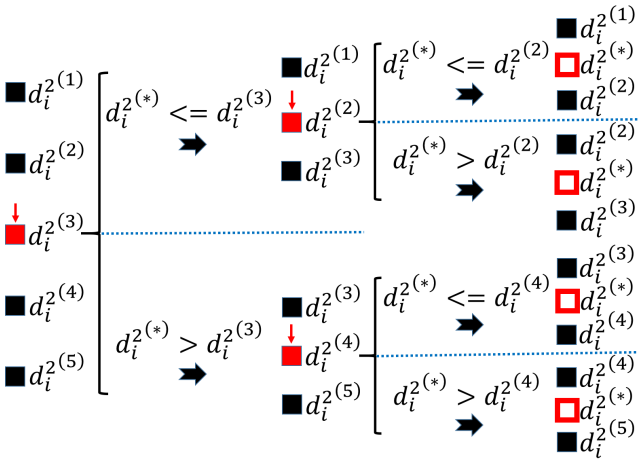

FIGURE 4. An illustrative example of the binary search in the 3rd scenario.

In addition to determining the target position in the ordered array for a fetched element when the 3rd scenario occurs, the above binary search process can also be used to help putting the first k elements of the SED sequence into the ordered array one by one in ascending order. The only difference is that, when one of the first k elements of the SED sequence is put into the target position in the ordered array, the maximum element in the ordered array need not to be removed. The number of comparisons is $\log_2(k!)$ at most.

Thus, *Strategy* $\Gamma\Gamma$ for rapid selection of the kth smallest SED has been constructed. The total number of comparisons for this strategy is $\log_2(k!) + \log_2(k) \cdot (N-L+1-k)$ at most. After *Strategy* $\Gamma\Gamma$, the present anomaly index value $AI_{i,p}$ for $\boldsymbol{x}_{i,p}^{\mathrm{T}}$ can be determined as this selected SED. Furthermore, the present system-wide anomaly index value AI_p globally characterizing the group

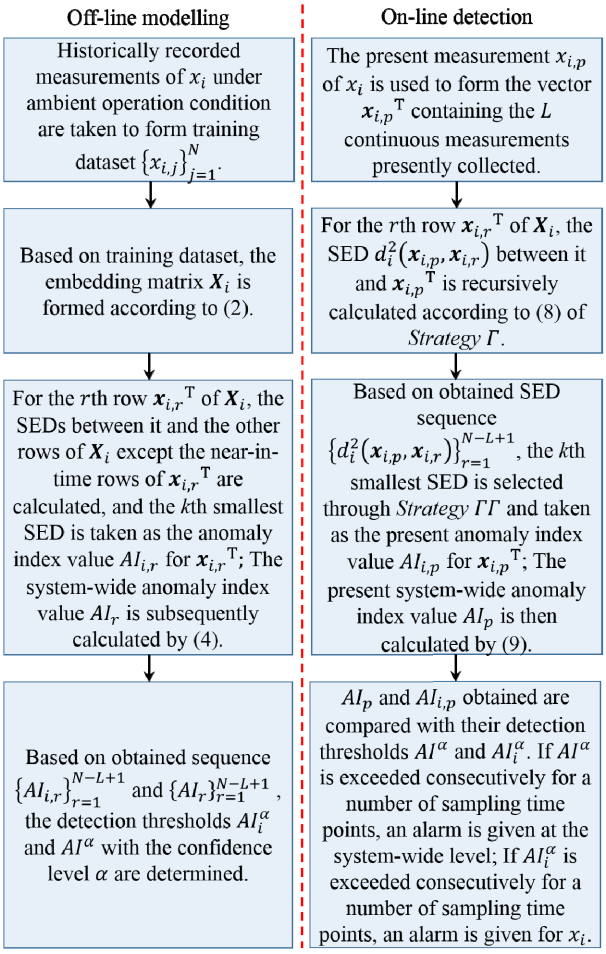

FIGURE 5. The procedure for the RD-kNN method.

of variables can be calculated as:

$$AI_p = \frac{1}{m}\sum_{i=1}^{m} |AI_{i,p}| \tag{9}$$

For the on-line disturbance detection, AI_p and $AI_{1,p}, AI_{2,p}, \cdots, AI_{m,p}$ obtained can be compared with their corresponding detection thresholds AI^{α} and $AI_1^{\alpha}, AI_2^{\alpha}, \cdots, AI_m^{\alpha}$, respectively.

C. PROCEDURE FOR THE RD-kNN METHOD

The RD-kNN method has been developed through *off-line modelling* and *on-line detection*. The procedure for the RD-kNN method is now summarized in Fig. 5, where the historical measurements $\{x_{i,j}\}_{j=1}^{N}$ recorded under the ambient operation condition and the present measurement $x_{i,p}$ collected online are all normalized with the mean and variance of the ambient measurements $\{x_{i,j}\}_{j=1}^{N}$.

D. PARAMETER SETTINGS FOR THE RD-kNN METHOD

The RD-kNN method involves the settings of the parameter k and the window length L. Presently, there is no standard and unified rule to optimally select the values for them. For the parameter k, it is better to assign it a relatively small value

from the perspective of the real-time requirement when online detection is implemented, because the total number of comparisons for selection of the kth SED in *strategy* ΓΓ increases with k increasing. As recommended in [10], a typical value of k is 3, which is also used in this paper. The reason why k is not set to be even smaller, e.g., $k = 1$, is from the consideration of avoiding false alarms during the normal operation condition. As for the window length L, it usually relates to the sampling interval and the duration of disturbances. According to [11], in order to be sufficient for characterizing disturbances, the measurements should be recorded with an appropriate sampling interval so that the duration of disturbances could be described by at least 40 measurements whenever possible. This requires having an idea of the duration of typical disturbances or anomalies in advance which can be judged from past experience with the system. For instance, if the experience of the site is that frequently occurring disturbances usually have a typical duration of 1 s, then the sampling interval should not be larger than 0.025 s to guarantee at least 40 measurements in disturbances. Under this circumstance, $L = 40$ could be considered for use.

IV. CASE STUDIES

In this section, the RD-kNN method is evaluated in two case studies, involving simulation data from the reduced equivalent model of GB power system [19] developed by the Power and Control Group at Imperial College London, and real measurements from the European power system.

A. REDUCED EQUIVALENT MODEL OF GB POWER SYSTEM

The reduced equivalent model of GB power system is developed using MATLAB/Simulink software and consists of 37 synchronous generators (10-coal, 8-nuclear, 5-hydro, and 14-combined cycle gas turbine), 11 wind generators, 48 generator buses, 33 load buses, and 151 transmission lines. The synchronous machines are represented using a transient model and recommended ranges of values for different generator parameters are obtained from [20]. The inertias for gas turbine generators, steam turbine generators, and hydro generators are selected between 4pu & 6pu, between 6pu & 10pu, and 3pu, respectively. The wind turbine generators are represented using the generic Type-4 WTG model [21]. The model has three main parts representing Scotland, England and Wales respectively. An inter-area mode with participation from synchronous generators located in both Scotland and England is present in the model. More details about this model can be found in [19].

The frequencies of 37 synchronous generators are monitored as the electrical variables to reflect the system condition and denoted as $x_1, x_2, \cdots, x_{37}$. Usually, during ambient operation, the frequencies of synchronous generators fluctuate around the steady-state value (50 Hz in the UK) and the fluctuation trends are mainly due to random events such as the normal variation of load demands. Besides, inter-area oscillations may also be observed in the frequency measurements of some generators when groups of synchronous machines in one part of the system oscillate with respect to groups in another part of the system. When loads significantly deviate from operating points, a resulting power mismatch is reflected in the frequencies of all synchronous generators. This larger change in operating points or disturbance is commonly seen in actual power systems. Possible causes for this type of disturbance are the intermittent operation of large load equipment or the synchronized surges in electricity consumption, known as TV pickup. However due to the fluctuating and oscillatory characteristics of the frequency measurements, the detection of such disturbance is a challenging task. Here, the RD-kNN method is applied, with the aim of providing increased situational awareness of the frequencies to power system operators.

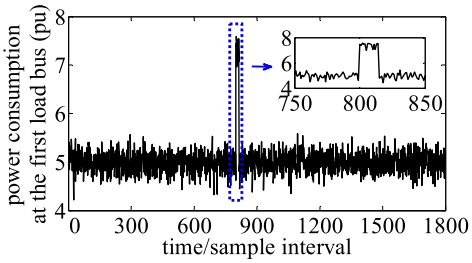

FIGURE 6. The step power change at the first load bus in the first case study.

The total simulation time is 180 s and the sampling interval is 0.1 s, generating 1800 frequency measurements. At the 800th sampling time point, a step power change occurs at the first load bus and lasts for 15 sampling time points, as shown in Fig. 6. As a result, a disturbance takes place in the frequencies $x_1, x_2, \cdots, x_{37}$. The first 500 frequency measurements with no disturbance are used as training data for the off-line modelling of the RD-kNN method, while the remaining 1300 frequency measurements containing the described disturbance are used as testing data for investigating the detection performance of the RD-kNN method. All the 1800 frequency measurements are normalized with the means and the variances of the first 500 frequency measurements. To provide a compact demonstration, the normalized measurements of $x_1, x_2, \cdots, x_5$ are shown in Fig. 7, with the disturbance highlighted in the rectangle.

Referring to Fig. 7, the number of measurements in disturbance event is more than 40. Thus, the 0.1 s sampling interval meets the recommendation in [11] that the duration of disturbances be described by at least 40 measurements. Accordingly, $L = 40$ is taken for use and the detection chart of the system-wide anomaly index AI on the testing data of the frequencies $x_1, x_2, \cdots, x_{37}$ is shown in Fig. 8. The detection threshold is determined with the confidence level $\alpha = 99\%$. Besides, to investigate the effect of different choices of L on the detection performance, $L = 41, 42, 43, 44$ are also taken for use, respectively. For each choice of L, the detection time point of AI is 813. It suggests that the detection performance of AI is not sensitive to variations in L around 40.

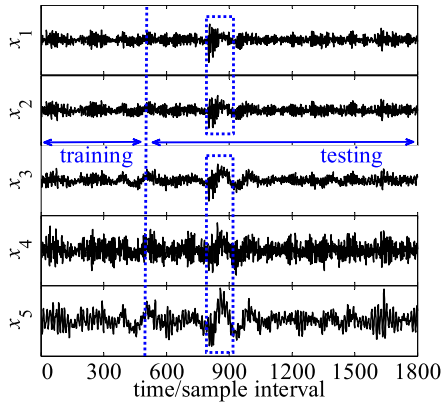

FIGURE 7. The normalized measurements of $x_1, x_2, \cdots, x_5$ in the first case study.

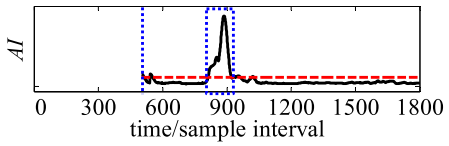

FIGURE 8. The detection chart of AI on the testing data of $x_1, x_2, \cdots, x_{37}$ for $L = 40$ in the first case study, showing the time-series values of AI (solid line) and the detection threshold (dashed line).

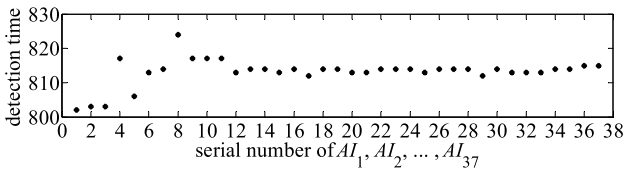

FIGURE 9. The detection times of $AI_1, AI_2, \cdots, AI_{37}$ for $L = 40$ in the first case study.

After the disturbance effect on the frequencies $x_1, x_2, \cdots, x_{37}$ is globally characterized by AI, the disturbance effect on each of $x_1, x_2, \cdots, x_{37}$ can be further checked by the anomaly indices $AI_1, AI_2, \cdots, AI_{37}$. Fig. 9 shows the detection times of $AI_1, AI_2, \cdots, AI_{37}$ for $L = 40$. Observing from Fig. 9, the detection time points of $AI_1, AI_2, \cdots, AI_{37}$ are different, ranging from 800 to about 820. The reason for the difference is that the frequencies $x_1, x_2, \cdots, x_{37}$ suffer from the disturbance effect with different degrees, as exemplified in Fig. 7. To provide a compact demonstration, the detection charts of $AI_1, AI_2, \cdots, AI_5$ on the testing data of $x_1, x_2, \cdots, x_5$ for $L = 40$ are shown in Fig. 10. As expected, $AI_1, AI_2, \cdots, AI_5$ provide different indications of the disturbance, coinciding with the observation from Fig. 7 that $x_1, x_2, \cdots, x_5$ are affected differently by the disturbance. On the other hand, all of $AI_1, AI_2, \cdots, AI_5$ detect the disturbance after it occurs because their time-series values obviously exceed the related detection thresholds, illustrating their satisfactory detection performance.

In order to evaluate the real-time efficiency of the RD-kNN method when on-line detection is conducted on the testing data of $x_1, x_2, \cdots, x_{37}$, the time of calculating AI for each data point is stored. The computations are carried out on an Intel(R) Core(TM) i7-4770 (3.40 GHz) with 16.0 GB RAM, and with Windows 7 Enterprise and MATLAB version R2014a. The maximum calculation time is in the order of 0.012 s and is much smaller than the 0.1 s sampling interval, meaning that the RD-kNN method meets the real-time requirement.

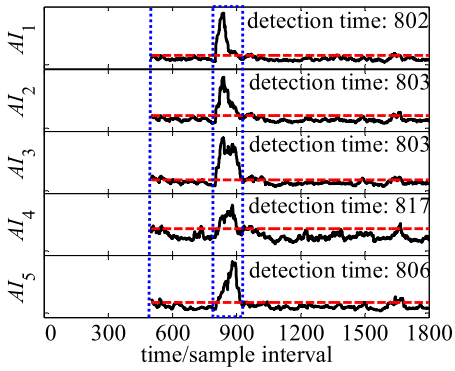

FIGURE 10. The detection charts of $AI_1, AI_2, \cdots, AI_5$ on the testing data of $x_1, x_2, \cdots, x_5$ for $L = 40$ in the first case study, showing the time-series values of $AI_1, AI_2, \cdots, AI_5$ (solid lines) and the detection thresholds (dashed lines).

TABLE 1. The electrical variables for monitoring in the second case study.

Variable	Description
x_1	Voltage amplitude in substation 1
x_2	Current amplitude in substation 1
x_3	Active power in substation 1
x_4	Apparent power in substation 1
x_5	Reactive power in substation 1
x_6	Voltage amplitude in substation 2
x_7	Current amplitude in substation 2
x_8	Active power in substation 2
x_9	Apparent power in substation 2
x_{10}	Reactive power in substation 2

B. EUROPEAN POWER SYSTEM

The RD-kNN method is next applied to the European power system data. The electrical variables for monitoring are described in Table I, and 3000 measurements are recorded for each variable with a 0.1 s sampling interval. The first 1000 measurements are taken to form training dataset since there is nothing abnormal in them to worry an operator in the control room and they reflect the characteristic of ambient operation, whereas the remaining 2000 measurements are taken to form testing dataset since they contain a disturbance caused by a switching operation, according to information received from the supplier of the data. All the 3000 measurements are normalized with the means and the variances of the first 1000 measurements, and the normalized measurements

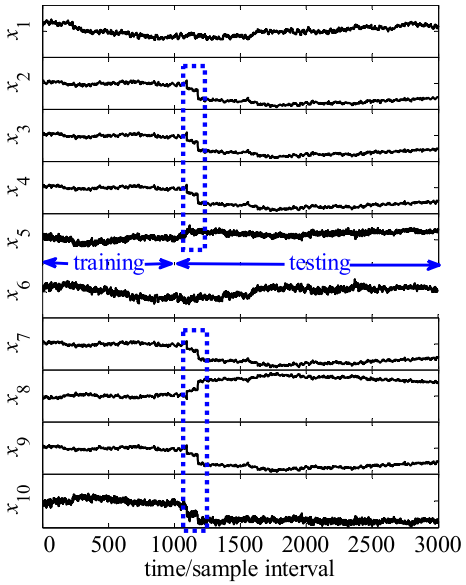

FIGURE 11. The normalized measurements in the second case study.

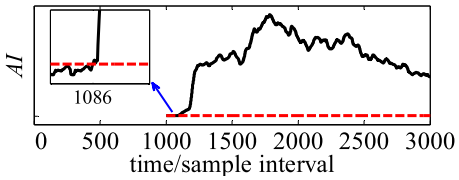

FIGURE 12. The detection chart of AI on the testing data of $x_1, x_2, \cdots, x_{10}$ for $L = 40$ in the second case study, showing the time-series values of AI (solid line) and the detection threshold (dashed line).

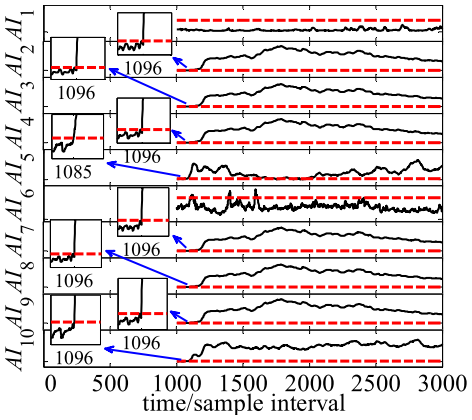

FIGURE 13. The detection charts of $AI_1, AI_2, \cdots, AI_{10}$ on the testing data of $x_1, x_2, \cdots, x_{10}$ for $L = 40$ in the second case study, showing the time-series values of $AI_1, AI_2, \cdots, AI_{10}$ (solid lines) and detection thresholds (dashed lines).

are shown in Fig. 11 with the disturbance highlighted in the rectangle.

Referring to Fig. 11, the disturbance inside the rectangle is characterized by a ramp lasting about 100 measurements. Thus, the 0.1 s sampling interval used in this real power system can meet the recommendation in [11] that disturbances be captured by at least 40 measurements. Accordingly, $L = 40$ is taken for use and the detection chart of AI on the testing data of the variables $x_1, x_2, \cdots, x_{10}$ is shown in Fig. 12. Again, the detection threshold is calculated with the confidence level $\alpha = 99\%$. It can be seen that AI reacts sharply and stays well above the corresponding detection threshold when the disturbance occurs, providing a definite indication of the disturbance. This shows the RD-kNN method has potential for practical application, because AI is system-wide anomaly index and the practical implementation of a detection method in a control room usually takes the form as a traffic light with green or red indicators for the overall state of the system. As in the first case study, $L = 41, 42, \cdots, 44$ are also taken to investigate the effect of different choices of L on the detection performance. The obtained detection results are almost the same as that in Fig. 12, meaning that the detection performance of AI is also not sensitive to variations in L around 40.

Furthermore, the detail about the reaction an individual variable exhibits to the disturbance can be checked by each of the anomaly indices $AI_1, AI_2, \cdots, AI_{10}$. As shown in Fig. 13, the disturbance is well detected in $x_2 \sim x_4, x_7 \sim x_{10}$ by $AI_2 \sim AI_4, AI_7 \sim AI_{10}$. These are the expected results, since the effects of the disturbance on $x_2 \sim x_4, x_7 \sim x_{10}$ can be seen in their measurements where a transient happens, as shown in the rectangle of Fig. 11. In contrast, the disturbance is not detected in x_1, x_6 by AI_1, AI_6, which is in line with the visual inspection that both x_1 and x_6 are not affected by the disturbance. It is also worth noting that the disturbance has certain effect on x_5 and this effect is not as obvious as those on $x_2 \sim x_4, x_7 \sim x_{10}$. However, AI_5 still detects the disturbance in x_5. It can be seen from the above results that the RD-kNN method is capable of detecting the disturbance effectively.

To assess the real-time performance of the RD-kNN method, the time of calculating AI for each testing data point is stored. The maximum calculation time is about 0.007 s, far smaller than the 0.1 s sampling interval. This means that the RD-kNN method is well qualified for the real-time implementation.

V. CONCLUSION

A new method called Real-time Detection based on k-Nearest Neighbor (RD-kNN) has been proposed to detect power system disturbances in real time. The contribution lies in the application extension of kNN from off-line detection to on-line detection by developing *Strategy* Γ for recursively calculating distance metrics and *Strategy* $\Gamma\Gamma$ for fast selection of the kth smallest metric. The application results on simulation data from the reduced equivalent model of GB power system and real data from the European power system have illustrated that the RD-kNN method can effectively detect disturbances.

Our future work will investigate the real-time multivariate detection of power system disturbances by integrating

multivariate statistical analysis with the present work. Further investigation on the real-time classification of power system disturbances will also be conducted based on the present work.

ACKNOWLEDGMENT

The authors would like to thank Dr. Mats Larsson of ABB Corporate Research Center, Baden-Dättwil, Switzerland for providing data to support this paper.

REFERENCES

[1] P. Anderson and K. Timko, "A probabilistic model of power system disturbances," *IEEE Trans. Circuits Syst.*, vol. 29, no. 11, pp. 789–796, Nov. 1982.
[2] J. Thambirajah, N. F. Thornhill, and B. C. Pal, "A multivariate approach towards interarea oscillation damping estimation under ambient conditions via independent component analysis and random decrement," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 315–322, Feb. 2011.
[3] E. Barocio, B. C. Pal, D. Fabozzi, and N. F. Thornhill, "Detection and visualization of power system disturbances using principal component analysis," in *Proc. IREP Symp.-Bulk Power Syst. Dyn. Control-IX (IREP)*, Rethymno, Greece, 2013, pp. 1–10.
[4] W. Zhu *et al.*, "A novel KICA–PCA fault detection model for condition process of hydroelectric generating unit," *Measurement*, vol. 58, pp. 197–206, Dec. 2014.
[5] L. Cai, X. Tian, and S. Chen, "A process monitoring method based on noisy independent component analysis," *Neurocomputing*, vol. 127, pp. 231–246, Mar. 2014.
[6] A. Ajami and M. Daneshvar, "Data driven approach for fault detection and diagnosis of turbine in thermal power plant using independent component analysis (ICA)," *Int. J. Elect. Power Energy Syst.*, vol. 43, no. 1, pp. 728–735, Dec. 2012.
[7] Z. Liu, Q. Hu, Y. Cui, and Q. Zhang, "A new detection approach of transient disturbances combining wavelet packet and Tsallis entropy," *Neurocomputing*, vol. 142, pp. 393–407, Oct. 2014.
[8] F. B. Costa, "Fault-induced transient detection based on real-time analysis of the wavelet coefficient energy," *IEEE Trans. Power Del.*, vol. 29, no. 1, pp. 140–153, Feb. 2014.
[9] S.-J. Huang, T.-M. Yang, and J.-T. Huang, "FPGA realization of wavelet transform for detection of electric power system disturbances," *IEEE Trans. Power Del.*, vol. 17, no. 2, pp. 388–394, Apr. 2002.
[10] I. M. Cecílio, J. R. Ottewill, J. Pretlove, and N. F. Thornhill, "Nearest neighbors method for detecting transient disturbances in process and electromechanical systems," *J. Process Control*, vol. 24, no. 9, pp. 1382–1393, Sep. 2014.
[11] I. M. Cecílio, J. R. Ottewill, H. Fretheim, and N. F. Thornhill, "Multivariate detection of transient disturbances for uni- and multirate systems," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 4, pp. 1477–1493, Jul. 2015.
[12] E. Barocio, B. C. Pal, N. F. Thornhill, and A. R. Messina, "A dynamic mode decomposition framework for global power system oscillation analysis," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 2902–2912, Nov. 2015.
[13] E. Keogh, J. Lin, and A. Fu, "HOT SAX: Efficiently finding the most unusual time series subsequence," in *Proc. 5th IEEE Int. Conf. Data Mining*, Houston, TX, USA, Nov. 2005, pp. 226–233.
[14] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.
[15] G. O. Campos *et al.*, "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," *Data Mining Knowl. Discovery*, vol. 30, no. 4, pp. 891–927, Jul. 2016.
[16] N. F. Thornhill, H. Melbø, and J. Wiik, "Multidimensional visualization and clustering of historical process data," *Ind. Eng. Chem. Res.*, vol. 45, no. 17, pp. 5971–5985, Jul. 2006.
[17] D. Fabozzi and T. Van Cutsem, "Assessing the proximity of time evolutions through dynamic time warping," *IET Generat., Transmiss. Distrib.*, vol. 5, no. 12, pp. 1268–1276, Dec. 2011.
[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2009, p. 39.
[19] L. P. Kunjumuhammed, B. C. Pal, and N. F. Thornhill, "A test system model for stability studies of UK power grid," in *Proc. IEEE PowerTech*, Grenoble, France, Jun. 2013, pp. 1–6.
[20] P. Kundur, *Power System Stability and Control*. New York, NY, USA: McGraw-Hill, 1994.
[21] WECC Renewable Energy Modeling Task Force, "WECC wind power plant dynamic modeling guide," WECC, Tech. Rep., Nov. 2010. [Online]. Available: http://renew-ne.org/wp-content/uploads/2012/05/WECCWindPlantDynamicModelingGuide.pdf

LIANFANG CAI received the B.Eng. and Ph.D. degrees from the China University of Petroleum, Qingdao, China, in 2009 and 2014, respectively. He is currently a Post-Doctoral Research Associate with the Imperial College London, U.K. His research interests include data-driven power system monitoring, modeling of power systems with energy storage, and multivariate statistics.

NINA F. THORNHILL (SM'93) received the B.A. degree in physics from Oxford University, Oxford, U.K., in 1976, the M.Sc. degree from the Imperial College London, London, U.K., and the Ph.D. degree from the University College London.

She is currently a Professor with the Department of Chemical Engineering, Imperial College London, where she holds the ABB Chair of Process Automation.

STEFANIE KUENZEL (GS'11–M'14) received the M.Eng. and Ph.D. degrees from the Imperial College London, London, U.K., in 2010 and 2014, respectively. She is currently a Lecturer with the Department of Electronic Engineering, Royal Holloway, University of London. Her current research interests include wind generator modeling and interaction studies.

BIKASH C. PAL (M'00–SM'02–F'13) is currently Professor of Power Systems with the Imperial College London. He is research active in power system stability, control, and computation. He has graduated 15 Ph.D. students and published 60 technical papers in the IEEE TRANSACTIONS and *IET journals*. He has co-authored two books and two awards receiving the IEEE Task Force/Working Group reports. He is currently the Editor-in-Chief of IEEE TRANSACTIONS ON SUSTAINABLE ENERGY. He has contributed to IEEE in power system stability and control.