Power System Coherency Detection From Wide-Area Measurements by Typicality-Based Data Analysis

Lucas Lugnani, *Student Member, IEEE*, Mario R. Arrieta Paternina, *Member, IEEE*, Daniel Dotta, *Member, IEEE*, Joe H. Chow, *Life Fellow, IEEE*, and Yilu Liu, *Fellow, IEEE*

***Abstract*—This paper presents a new data-driven methodology for power system coherency identification of generator and non-generator buses. This methodology is exclusively based on intrinsic statistical properties extracted directly from observations, without any prior assumption of the probability distribution function (PDF) for the data. The main advances of this proposal are: (*i*) gathering of statistical information from the data itself despite scenarios where the PDF may change (different inverter-based load and generation scenarios, load levels of the system, and changes in topology); and (*ii*) assignment of buses into coherent areas without any tuning of parameters, nor manually labeling of huge amounts of training data. This new method, called typicality-based data analysis (TDA), is applied to the correlation metric of the distance between dynamic responses of buses, either voltage angles or frequencies. Simulated signals from a benchmark power system with cases considering the presence of non-synchronous generation and islanding conditions, and real data associated with generation trips in the U.S. Eastern Interconnection are used to corroborate the methodology effectiveness.**

***Index Terms*—Coherency, clustering, data-driven, WAMS, statistical typicality.**

I. Introduction

THE electric power system is currently experiencing significant changes in its physical structure, with increasing rate of renewable penetration, as well as increasing dimension and complexity of power systems models [1]. In this new environment, where the power system flexibility [2] is a valuable asset, the concepts of coherency and model reduction can be useful to understand power system dynamic behaviors and develop advanced applications such as controlled islanding [3]–[8], wide-area control and protection [9]–[17].

On the other hand, thanks to the advance of the wide-area monitoring systems (WAMS), power system operators can analyze large amounts of data, potentially a valuable source of information about the power system dynamics. Of this manner, the easy access to data is changing power system analysis and demanding more effective tools to extract information from them.

A. State-of-The-Art

The concept of coherency proposed in [18] develops the slow modes analysis (the ones concerning inter-area oscillations) of the linearized inertial aggregated model which groups generators by considering the equivalent machine angle, with machines internal nodes connected by infinite admittances. In [13], the method in [18] is further advanced by adjusting the inertia aggregated model of a cluster, since this corrects the admittance connecting the internal nodes of machines by the fast modes of the model (local oscillating modes), improving the representation of the system. These model-based approaches (MBAs) have been extensively explored in the literature and their recent advances are reported in [14]. Despite these strong advances, MBAs rely on the linearized model of a high-dimensional and complex nonlinear system, resulting that there are no guarantees for employing this concept in power system contingencies that may change the system structure and excite nonlinear dynamics. Thus, MBAs may not be useful for online application in modern power systems [19], [20].

Conversely, with the advent of WAMS, there is a clear need to explore the use of PMU measurements (voltage phase angle and frequency) to identify generator coherency. Where the new paradigm is not to rely on power system models (parameters and topology), but rather make use of the power system measured responses. These data-driven methods (DDMs) can be divided into three main approaches: temporal signal clustering [6], [19]–[28], oscillatory mode detection [29]–[34] and machine-learning (ML) techniques [35], [36].

Regarding the temporal signal clustering approaches, an independent component analysis method using the rotor speed and angle of synchronous generators to determine their clusters, is proposed in [22]. Meanwhile, the authors in [19], [23] employ frequency deviation signals within a two step method that identifies clusters of generators by a cosine correlation index and aggregates the remaining buses into the coherent clusters. In [24], authors present an average correlation coefficient for clustering

Manuscript received November 23, 2020; revised April 24, 2021; accepted June 6, 2021. Date of publication June 10, 2021; date of current version December 23, 2021. The authors would like to thank the Power Information Technology Lab and FNET/GridEye for the real network data provided. They would also like to thank the agencies CNPq (grant 170100/2018-9), São Paulo Research Foundation (FAPESP) (grants 2016/08645-9, 2018/07375-3 and 2019/10033-0) for the financial support and equipment. Paper no. TPWRS-01925-2020. *(Corresponding author: Lucas Lugnani.)*

Lucas Lugnani and Daniel Dotta are with the Department of Electrical Engineering, University of Campinas, Campinas, SP 13083-852, Brazil (e-mail: lugnani@dsee.fee.unicamp.br; dottad@unicamp.br).

Mario R. Arrieta Paternina is with the Department of Electrical Engineering, National Autonomous University of Mexico, Mexico City 04510, Mexico (e-mail: mra.paternina@fi-b.unam.mx).

Joe H. Chow is with the Department of Electrical Computer and Systems Engineering, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180 USA (e-mail: chowj@rpi.edu).

Yilu Liu is with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA (e-mail: Liu@utk.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2021.3088261.

Digital Object Identifier 10.1109/TPWRS.2021.3088261

which is focused on an improvement of the Euclidean norm in combination with a threshold-defined heuristic algorithm. The Pearson product-moment correlation coefficient (PPCC) is introduced in [25] by defining a distance metric among PMU voltage angles and applying a hierarchical density-based spatial clustering of applications with noise (HDBSCAN) to select the clusters. Other investigations explore several distance metrics obtained from a special device to estimate rotor speeds and internal angles of generators; such metrics are ranked and processed by means of the criteria importance through inter-criteria correlation (CRITIC) and the kernel principal component analysis (KPCA) [6], [26], [27]. These processed indexes are then clustered using agglomerative hierarchical clustering (AHC), spectral clustering and affinity propagation (AP) methods. In [28], the use of PMU measurements and dynamic time warping (DTW) method form a strategy to identify coherency online from the rotor angles' information. Likewise, the work in [20] tackles a new data-driven methodology for slow-coherency clustering of generators, hinged on heuristically determined composition of cosine dissimilarity and Minckowski distance among PMU measurements of frequencies at the generator terminal buses. The clustering process itself is performed using the affinity-propagation technique.

In the second approach, recent works propose the use of oscillatory mode extraction techniques, such as the Koopman method [29], [37], to identify the coherency of generators. In [30], a data-driven method that estimates the system's modes using angle measurements from all buses is proposed by performing a spectral analysis using the Koopman operator to identify the dominant modes and clustering with the K-means method. Also, the work in [31] extracts the modes of WAMS measurements applying the Koopman operator and then applies spectral analysis to identify coherent groups of generators with high penetration of non-synchronous generation. Meanwhile, the authors in [32] employ a linear quadratic regulator (LQR) and Kalman filtering applied to synchrophasor measurements to estimate space-state variables for determining oscillations among areas and apply to those clustered areas in controlled islanding schemes. Likewise, the authors from [33] extract the modes using Taylor-Fourier Transform and cluster the generators using hierarchical agglomerative technique, with Elbow's method to improve the initial guess for the number of clusters. In [34], a fast frequency domain decomposition (FFDD) modal analysis method is applied to real measurements of oscillation monitoring, from the Réseau de Transport d'Électricité (RTE) power system, and the clustering is processed by the DBSCAN method.

Finally, ML approaches rely on large data sets to train the classifiers. In [35], bus angle and frequency measurements from PMUs are used to determine a dissimilarity rms-coherency criterion index between buses, for each disturbance event, forming a matrix of dissimilarity indexes. Gathering matrices from several events, a probability of similarity among buses is constructed and applied to a fuzzy medoids algorithm (FCMdd) to perform the clustering. In [36], the coherency detection is applied to unstable simulated transient events, which are first classified using binary labeling. Once, a relative large number of cases is simulated, hierarchical clustering is applied to group formation. Then, different classification techniques (decision tree, ensemble decision tree and multi-class support vector machine) are explored to identify the unstable responses (unstable groups).

Despite all advantages enclose in the three aforementioned approaches, there are some gaps to be fulfilled. The main limitations of the temporal signal clustering methods are usually associated with an empirical threshold that must be tuned, which may have to be re-tuned for anomalous events by expert users with a previous knowledge of the system dynamic behavior. Regarding the oscillatory mode extraction methods, the authors in [20] claim that these techniques are often affected by the inaccuracies of the mode estimation and the high computational burden required to process long observation windows. Finally, the key requirement for the success of machine-learning approaches is to have a representative database used in the training process to prevent over-fitting problems. This accuracy crucially depends on the quantity and quality of the available data as well as the time consuming task of manually labeling a huge amount of events. The training process of machine-learning methods also involves a large computational burden and manual configuration of hyper-parameters that must be retrained after possible classification failures. Furthermore, the interpretability may also be a limitation for deep learning methods when they are applied to critical tasks.

B. Problem Statement

Nowadays, the power system industry has been experiencing a major challenge since synchronous machines and controllers are replaced by inverter-based sources (IBS). The effects of the integration of a large amount of IBS, whose regulation and interaction with the rest of the system is still to be fully understood, may impact the identification of groups depending on the state of the system and location of the disturbance [19], [20], [26], [31]. To develop fully data-driven applications capable to process large amount of collected PMU data, it is helpful to understand the effects of IBS on coherency identification, islanding detection and model reduction of power systems. Table I shows a comparison of the required information/assumptions by the methods, this is marked by **X** and additional information that some of the methods can provide, besides machine clustering, such as islanding detection, marked by checkmark ✓, when compared with the method proposed here. It is important to point out that, to the authors' best understanding, some of the methods may be able to provide additional information, but they do not present any comment or results to that regard.

In this work, we introduce the concept and explore the advantages of a non-parametric statistical method for coherency tracking. This method does not have the constraints that the parametric methods impose for their application, which requires a previous knowledge about the process and dataset (population). This clearly reduces the effort to apply and understand the proposed method, improving its use in real world applications.

C. Contributions

The main contribution in this work is related to the extraction of statistical characteristics exclusively from the data, without any assumption of the distribution of the data. This idea establishes a new paradigm for data handling, as the number of clusters is automatically found from each data-set, regardless of parameter tuning like most data-driven methods. Further, the statistical information extracted from the data is supported mathematically. The method, typicality-based data analysis (TDA),

TABLE I
COMPARISON OF REQUIREMENTS FOR COHERENCY METHODS

	# of clusters	Initial centers	User defined constants	Gen. parameters	Network parameters
[31]	**X**		**X**	**X**	
[19], [23]			**X**	**X**	
[33]	**X**	**X**		**X**	**X**
[24]			**X**	**X**	
[38]			**X**		
[35]	**X**	**X**	**X**	**X**	
[30]	**X**			**X**	**X**
[36]	**X**		**X**		
[6], [26], [27]			**X**	**X**	
[28]	**X**	**X**	**X**		**X**
[32]	**X**	**X**	**X**		
[20]	**X**	**X**			
TDA					
	Algorithm training stage	PMU measurements	Non-gen. buses clustering	Robust to Fast Dynamics	Islanding detection
[31]		**X**		✓	
[19], [23]		**X**	✓	✓	✓
[33]		**X**			
[24]		**X**			
[38]	**X**	**X**	✓	✓	✓
[35]	**X**	**X**	✓	✓	
[30]		**X**	✓		
[36]	**X**	**X**			✓
[6], [26], [27]		**X**			
[38]		**X**			✓
[32]		**X**		✓	
[20]		**X**		✓	
TDA		**X**	✓	✓	✓

is applied to distances between frequency dynamic responses by employing the methodology in [39]. It is also customized to be implemented along with synchrophasor measurements from dynamic transient responses for the detection of power system islands by performing a clustering process. The main contributions are stated as follows: (*i*) this is a fully data-driven method which means that there is no necessity to determine the optimal number of clusters or initial guesses of centroids to initialize the grouping algorithm; (*ii*) contrary to conventional parametric statistical methods that must rely on probability density functions (PDF), assuming a set of fixed parameters that determine a probability model, non-parametric methods do not require previous knowledge of the process and the dataset (population) being handled; (*iii*) there is no necessity to manually label all the huge amounts of training data to build a representative database to be used in a training process aiming to prevent over-fitting problems; (*iv*) the mathematical background of the proposed approach is clear, allowing understanding of the results; (*v*) the method is capable of detecting the islanding conditions of the system and is robust to noisy measurements; (*vi*) due to its low computational complexity, it is suitable for transitory period applications; and (*vii*) the method is tested and validated using real PMU measurements from a large power system.

II. FUNDAMENTALS

A. Coherency

Coherent trajectories are defined as machines with responses indistinguishable from each other, i.e., the difference between their angles (θ) or frequencies (f), remains very small [13]:

$$\theta_k(t) - \theta_j(t) \leq \gamma \tag{1}$$

where k and j are generator buses, γ is an arbitrarily user-defined value for the maximum divergence between any two responses within an area. This method can be applied to either f or θ, since the first is a derivation of the second one, as stated by

$$\Delta f_i|_{t+\Delta t} = \frac{1}{\omega_0} \frac{\theta_i|_{t+\Delta t} - \theta_i|_t}{\Delta t} \tag{2}$$

where $\Delta f_i|_{t+\Delta t}$ stands for the frequency deviation (in Hz) of the i-th bus at the time step Δt, $\omega_0 = 2\pi f_0$, f_0 is the system nominal frequency in Hz, and θ_i is the i-th bus voltage angle. Since Δt and ω_0 are constant in (2), we can regroup them as a constant η and considering $\theta_i|_{t+\Delta t} - \theta_i|_t = \Delta\theta_i|_{t+\Delta t}$, such that Δf_i becomes:

$$\Delta f_i|_{t+\Delta t} = \eta \Delta\theta_i|_{t+\Delta t} \tag{3}$$

where $\eta = 1/(\omega_0 \Delta t)$.

B. The Euclidean Norm

A norm maps vectors onto a scalar to represent the distance between two time-domain responses. The Euclidean norm is used since is considered stable, i.e., it is reliable to the adjustments of window lengths when it is compared with the absolute norm, which is considered more robust to outliers. Meanwhile, the outlier robustness can be readily overcome by filtering [40]. Given two points of measurement k and j for every time instant t, the distance $dd_{k,j}(t)$ between their frequencies is expressed by [41]

$$dd_{k,j}(t) = [f_k(t) - f_j(t)]^2 \tag{4}$$

where $dd_{k,j}(t)$ is squared since the value of $f_k(t) - f_j(t)$ may be negative. The Euclidean norm is a distance metric that satisfies all the following conditions [42]: *i)* non-negativity: $dd_{k,j} \geq 0$; *ii)* identity of indiscernible: $dd_{k,j} = 0 \iff k = j$; *iii)* symmetry: $dd_{k,j} = dd_{j,k}$; and *iv)* triangle inequality: $dd_{k,h} + dd_{j,h} \geq dd_{k,j}$. Metrics such as the cosine similarity do not attain such conditions. This is important because, with the Euclidean norm, we are able to represent the distance among two responses by a scalar and retain the signal mathematical properties.

Additionally, the nominal frequency f_0 is removed, so that $dd_{k,j}(t)$ is calculated as

$$dd_{k,j}(t) = [(f_k(t) - f_0) - (f_j(t) - f_0)]^2 = [\Delta f_k(t) - \Delta f_j(t)]^2 \tag{5}$$

Next, the square root of the sum of all values for a time window T is computed, $T = t_0, \ldots, t_f$, where t_0 is the moment of the disturbance, and t_f corresponds to the time window ending. The square root of the sum of $dd_{k,j}(t)$ is calculated projecting it onto a matrix of scalar quantities $\nu(k, j)$, with each entry representing the distance between points of measurements k and j, expressed as [41]

$$\nu(k, j) = \sqrt{\sum_{t=t_0}^{t_f} [\Delta f_k(t) - \Delta f_j(t)]^2} \tag{6}$$

For every bus k, ν_k is a $1 \times N$ vector, corresponding to the distances between the dynamic response from bus k to the other buses, where N is the total number of buses with available measurement.

C. The Distance Metric: Correlation

The vector of scalar quantities ν_k represents the norm of the distance from bus k to the other buses, making up a data point in the data-set to be used by the TDA method. The proposed method requires a measured quantity between the data points in the set, defined by the user [39]. Therefore, the correlation $\rho_{k,j}$ between two data points in the vector $\nu(k, j)$ is defined by

$$\rho_{k,j} = corr(\nu_k, \nu_j) = \frac{cov(\nu_k, \nu_j)}{\sigma_{\nu_k} \sigma_{\nu_j}} \tag{7}$$

where $corr(\nu_k, \nu_j)$ indicates the correlation between the distances of buses k and j, distributed in R^N, $cov(\nu_k, \nu_j)$ represents the covariance between the distances ν_k and ν_j, and σ is the standard deviation.

III. Typicality-Based Data Analysis

In this section, the fundamentals and definitions of a non-parametric statistical method [39] applied to the coherency tracking are presented. This is a distribution free method that is exclusively based on ensemble statistical properties of the data derived entirely from the experimental discrete observations. These properties are defined as follows [39].

A. Cumulative Proximity

In graph (networks) theory, a measure of *centrality* is defined as the inverse of the so-called *farness* which is a sum of distances from one point to all other points [43]. From [39], the cumulative proximity is defined as a squared form of the *farness:*

$$q_N(\nu_k) = \sum_{j=1}^{N} \rho_{k,j}; \qquad \nu_k \in \nu_N \tag{8}$$

Cumulative proximity is an important association measure that is empirically derived from the observed data without making any *prior* assumptions about their generation model, and plays a fundamental role in deriving other statistical properties for the TDA method [39].

B. Standardized Eccentricity

This quantity is defined within the TDA method as a normalized cumulative proximity by half of the average cumulative proximity:

$$\epsilon_N(\nu_k) = \frac{2 q_N(\nu_k)}{\frac{1}{N} \sum_{j=1}^{N} q_N q_N(\nu_j)} \tag{9}$$

where the coefficient two is included to compensate distance duplication in the denominator. If ϵ is divided by the amount of data N, then the non-standard eccentricity ξ becomes $\xi_N(\nu_k) = \frac{1}{N} \epsilon_N(\nu_k)$, leading to the following bounds for the eccentricity value:

$$0 \leq \xi_N(\nu_k) < 1 \tag{10}$$

This property makes up a significant measure of the ensemble property related to the distribution tail and it is also empirically derived from the observed data. It plays an important role in anomaly detection, analysis of rare events, as well as for the estimation of the typicality [39]. By considering the *Chebyshev inequality* [44] that indicates the probability of data being outlier (a data sample ν is more than $n\sigma$, where σ denotes the standard deviation, distance away from the *mean* in a given distribution) and applying the standard eccentricity to it, the TDA version of the *Chebyshev inequality* becomes [39]:

$$P(\epsilon_N(\nu_k) \leq n^2 + 1) \geq 1 - \frac{1}{n^2} \tag{11}$$

By expressing the *Chebyshev inequality* by means of the standard eccentricity, this allows detecting anomalies in data. For instance, if the standardized eccentricity $\epsilon_N(\nu) > 10$, then ν exceeds the 3σ limitation, this event can be categorized as an anomaly. This information is significant for boundary data, since it minimizes the probability of data miss-location in wrong clusters.

C. Discrete Local Density

This property is defined as the inverse of the *standardized eccentricity* [39], becoming

$$D_N(\nu_k) = \frac{\sum_{j=1}^{N} q_N(\nu_j)}{2N q_N(\nu_k)}, \quad \text{with} \quad i = 1, 2, \ldots, N. \tag{12}$$

D. Discrete Typicality

This quantity is established as the normalized *density*. It quantifies how common, or typical, a value is within the data set under study. As comparison, the typicality can be seen as a

probability equivalent of a given random variable that takes the value of the measured point. The typicality is obtained from the data set instead of being assigned by model fitting of a probability mass function (PMF), and it is given by [39]

$$\tau_k(\nu_k) = \frac{D_N(\nu_k)}{\sum_{j=1}^{N} D_N(\nu_j)} = \frac{q_N^{-1}(\nu_k)}{\sum_{j=1}^{N} q_N^{-1}(\nu_j)} \tag{13}$$

The *discrete typicality* resembles the traditional unimodal PMF, being constructed from the data set and excluding the possibility of non-feasible values that may result in consequence of fitting to PMF [39]. In the following section, the proposed method is applied to frequency measurements.

E. Proof of the Clustering Concept of TDA

Let's assume a system with B buses and their respective frequency measurements that are converted into a scalar by (6), where $N_B = \nu_1, \ldots, \nu_n, \ldots, \nu_B$, $\nu_1 = \nu(1,1), \nu(1,2), \ldots, \nu(1,B)$. The distance metric of correlation among points given by $\rho_{1,2}$. Once the TDA is applied, the initial set points N_B is divided into clusters of points $(\alpha, \beta, \ldots, c)$, where c is the number of clusters found by the TDA method. Let α be a cluster of buses, whose Euclidean norms from N_B are denoted by

$$N_\alpha = \{\nu_i, \nu_k, \ldots, \nu_a\}$$

and their distances with respect to all buses are given by

$$P_\alpha = \{\rho_i, \rho_k, \ldots, \rho_a\}$$

Then, their eccentricities are expressed by

$$E_\alpha = \{\epsilon_i, \epsilon_k, \ldots, \epsilon_a\}$$

And their typicalities are symbolized by

$$T_\alpha = \{\tau_i, \tau_k, \ldots, \tau_a\}$$

where T_α and P_α are used to determine the closest points in the data-set distribution, as shown in **Algorithm 1**. Once, a cluster is found, the mean $\mu_\alpha(\nu)$ and standard deviation $\sigma_\alpha(\nu)$ are calculated for the cluster. ρ_α^* is the minimal correlation, i.e., maximum distance in cluster α.

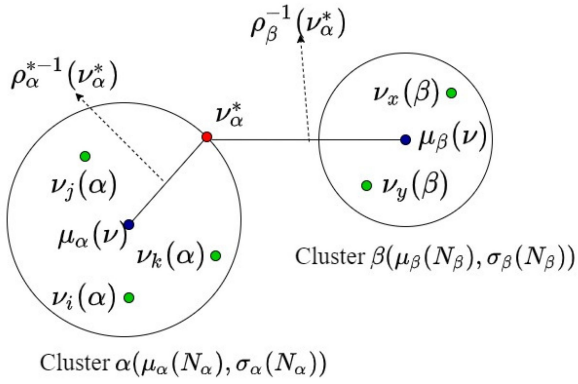
The probability of the point ν_α^* of α being less than $3\sigma_\alpha$ distant from μ_α of cluster α, i.e., bus α^* belonging to cluster α, can be seen by its eccentricity ϵ_α^*, when we apply ϵ_α^* to (11), with respect to the cluster standard deviation

$$P\left(\epsilon *_\alpha (\nu *_\alpha) \leq 3\sigma_\alpha^2 + 1\right) \geq 1 - \frac{1}{3\sigma_\alpha^2} \tag{14}$$

If we assume a normalized standard deviation of $\sigma_\alpha = 1$, then we get

$$P\left(\epsilon *_\alpha (\nu *_\alpha) \leq 10\right) \geq \frac{8}{9} \tag{15}$$

which is a conservative estimate since the Chebyshev inequality does not assume any prior information about the distribution of the data. For instance, the usual assumption for normal distribution in our case, the probability of being under 3σ of the mean is 99.7%. As it will be shown in the next section, the construction of the algorithm allocates each point ν to the cluster whose highest typicality point ν_τ has the closest distance ρ to ν. This in turns means that ν_α^* has the highest probability of belonging to cluster α, of all clusters. This is further exemplified in Fig. 1, where a visual representation of the statistical proof and algorithmic construction of the clusters is depicted. This will be also discussed in detail in the next section. Such construction and mathematical proof indicate the meaningfulness of the clustering produced by the TDA method.



$$\rho_\alpha^{*-1}(\nu_\alpha^*) < \rho_\beta^{-1}(\nu_\alpha)$$

&

$$P(\epsilon_\alpha^*(\nu_\alpha^*) \geq 9\sigma_\alpha^2 + 1) \leq \frac{1}{9} \leq P(\epsilon_\alpha^*(\nu_\alpha^*) \geq 9\sigma_\beta^2 + 1)$$

Fig. 1. Clustering validity example.

IV. TDA Application for Coherency Detection

A. Stage I. Pre-Processing

Due to non-electromechanical phenomena, the voltage angles may present spikes known as phase-shifts [45], which are unrealistic for machines rotor dynamics and bus frequencies overall. For this reason, a *moving median* filter is the first step in the pre-processing stage. Additionally, any PMU that reports data quality issues per flags STAT [46] (bits 6 to 15), is discarded. Errors in measurement that bring bias to the reported synchrophasors must be addressed by the state estimation and are out of the scope of this work. However, it is noteworthy to mention that a constant bias in the angle measurements would not impact the frequency since this is estimated regarding the angle variation. This stage comprises: (*i*) outlier removal with the *movmedian* Matlab function (this is applied using a 5-sample window); (*ii*) DC offset removal, i.e., difference from 60 Hz Δf is computed; the resulting signal is detrended with the dynamics separation algorithm [47], which is of great importance particularly in events such as generation trips, where the steady state component of the signal changes; (*iii*) computation of the Euclidean norm using (6), that is, $\nu(k,j) = \sqrt{\sum[dd_{k,j}^2]}$ [41]. This norm maps vectors onto scalars in order to represent time-domain responses in a scalar space distribution (reducing the dimension of the data-set). At the end of the pre-processing stage, a data set is generated in R^N, that is, the dimensional space of the data set is equal to the number of measurement points (ideally, equal

to the number of buses), with points $\nu_i = [\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,K}]^T$, $i = 1, 2, \ldots, N$, where each value in vector ν_i is a norm of bus i to another bus, and ν_i denotes the coordinates of bus i in such space.

B. Stage II. Metric (Correlation) Computation

In this investigation, the correlation ρ is adopted as a distance metric, being implemented as exhibits lines 4 to 8 of **Algorithm 1**.

C. Stage III. Properties Calculation

The TDA method clusters data using the typicality of each data point in the data set, using ρ as a metric. To reach the typicality value, the properties provided in the previous section are calculated for a given set of data points ν_k: the cumulative proximity $q_N(\nu_K)$ is computed using (8); the standardized eccentricity $\epsilon_N(\nu_k)$ is quantified using (9) (which is an important measure for data-handling correction, as $\epsilon_N(\nu_k)$ must be a value between 0 and 1); the discrete local density $D_N(\nu_k)$ is obtained from (12).

Finally, the typicality $\tau_k(\nu_k)$ of ν_k is calculated using (13) and taking into account the following properties: (*i*) the sum of the typicalities for all data points $\tau_k(\nu_k)$ is 1; (*ii*) all values of τ_k are between 0 and 1; and (*iii*) no prior assumptions of the data model are gathered. This is indicated through lines 10 to 12 of **Algorithm 1**.

D. Stage IV. Typicality Ranking

Once all τ_k are computed, the one with the maximum value is tagged as the global typicality τ_N^{D*} and placed in the first element of the vector $ranked_\tau$. A ranking of typicalities is accomplished as follows: the data point ν^2, where the superscript 2 indicates the position in the ranking of typicalities with the highest metric ρ to the data point of the global typicality τ_N^{D*}, and its τ_k^2 is assigned next in the vector $ranked_\tau$. Then, the data point ν^3 with the highest metric ρ to the data point ν^2 of the typicality τ_N^2, and its τ_k^3 arrayed next in the vector $ranked_\tau$. This is recursively performed until all typicalities are ranked, as pointed out in lines 15 to 19 of **Algorithm 1**.

E. Stage V. Cluster Formation and Filtering

The typicalities' peaks are found locating the points ν_k^* as initial cloud centers. This is carried out employing lines 20-26 of **Algorithm 1**. Once all cloud centers are located, the remaining data points ν_k are assigned to the center's cluster, in which it has the highest correlation ρ. This is conveyed in **Algorithm 1** from lines 28 to 30. For all clusters, the mean ($Cluster_\mu$) and deviation ($Cluster_\sigma$) of data points are computed by lines 34 to 37 in **Algorithm 1**. Finally, the clusters are filtered by clustering all clouds that are close together and recalculating their statistical properties. This is performed by lines 39 to 42 of **Algorithm 1** until the number of clusters remains unchanged. The final clusters correspond to the areas found using the TDA method. Here, it is important to point out that, by using the Euclidean distance among the measured frequency deviations, the TDA method implicitly takes into account the inertia of the generation units in the system as the typicality property of the method. All the process performed by **Algorithm 1** takes place in a single step manner, unlike the approach in [23], where the constant of neighborhood defined by the user must be changed for non-generator buses and supposes uniform inertia distribution. Other methods assume that the center of the inertia frequency deviation vector considers equal weights to all generators, unlike the TDA method that implicitly regards the inertia of each generator, since the Euclidean distance of frequencies is greatly influenced by the inertia of the areas.

V. Performance of the Typicality-Based Data Analysis

The TDA method is now applied to the New England 68-bus and 16-machine test system (S1) [48] and to real measurements from FNET/GridEye WAMS [49] for the Eastern Interconnection (S2).

The nonlinear simulations that provide the input data for the TDA method obtained from the power system toolbox (PST) [50], [51], assuming the availability of voltage angle/frequency responses at all buses. All simulations are carried out for 20 s with a time-step of 1 ms. The time window considered for calculation of ν_k in all cases is of 10s after disturbance takes place, as in [23]. The responses are decimated to 120 Hz, complying with the IEEE synchrophasor standard [46], to the simulated system *S1*. The transitory period of the response is useful for the detection of islanding condition; meanwhile, the transient period allows the correct slow-coherency detection. Additionally, the method was explored in *S1* for measurements with rates of 60 and 30 Hz, displaying similar results. The measurements from *S2* are by default 10 Hz.

A. 68-Bus System (S1) - Comparison to DCD

The 68-bus and 16-machine system *S1* is a reduced order equivalent of the interconnected New England transmission system (NETS) and New York power system (NYPS). All generators are represented by a sixth order model equipped with automatic voltage regulators (AVRs), and all loads are assumed as constant impedance [52]. *Cases S1.C1* and *S1.C2* intend to compare the areas found with those ones in [23], and illustrate the advantages against MBAs, since it detects islands and areas not connected, which is of great interest for wide-area control purposes. The noise tolerance is assessed including tests with noisy signals up to 30 dB of signal-to-noise ratio (SNR) for *Cases S1.C1* and *S1.C2*, but they are not displayed for the sake of brevity, since the TDA method is able to find the same areas.

1) Application on Case S1.C1: the first case is a three-phase fault at bus 27 in Fig. 4, at $t = 0.5\ s$, lasting 5 cycles.

The result of Stage I is a data set of the Euclidean norms ν_k in R^{68}, with number of points $N = 68$. In Stage II, each point ν_k have its correlation metric ρ_k to every other point ν_j, forming a metric vector with the same dimension. The correlations among the norms of all signals from *S1.C1* are projected onto the heat map in Fig. 2, where the strong correlations are represented in brown color. The main challenge now is to compute how the groups of high correlation buses can be formed into clusters.

In the proposed method, the clusters are obtained without any arbitrary cutoff constant using the TDA properties. The main result is the vector of typicalities for every point ν_k, which is depicted in Fig. 3(a) (before ranking). The high values of

Algorithm 1 TDA implementation for PMU dynamic response

Input: Let $\boldsymbol{\nu}_k$, $k = 1, \ldots N$ (data points) vector of scalar Euclidean norms between frequencies responses, with N being the number of PMUs.
Output: A set of coherent areas with generators and non-generator buses (**Clusters**).
Initialization: $\mathbf{t}_0$, $\mathbf{t}_f$, set of correlation metrics $\rho_{k,j}$
for k=1,k++ **do**
  for j=1,j++ **do**
    $\rho_{k,j} \leftarrow \frac{cov(\nu_k,\nu_j)}{\sigma_k \sigma_j} \quad \nu_k, \nu_j \in \nu_N$
  end for
end for
TDA properties computation
Cumulative proximity: $q_N(\nu_k) \leftarrow \sum_{j=1}^{N} \rho_{k,j}$;
Discrete local density: $D_N(\nu_k) \leftarrow \frac{\sum_{j=1}^{N} q_N(\nu_j)}{2Nq_N(\nu_k)}$
Discrete typicality: $\tau_k(\nu_k) \leftarrow \frac{D_N(\nu_k)}{\sum_{j=1}^{N} D_N(\nu_j)}$
Global typicality: $\boldsymbol{\tau}_N^{D*} \leftarrow max(\boldsymbol{\tau}_i^D) \quad i = 1, \ldots, N$
Starting from data point ($\nu(\tau_N^{D*})$), rank of typicalities (τ_k^D) for all data points (ν_k) based on the correlation metric ($\rho_{k,j}$):
for k=2, k++ **do**
  for j=1, j++ **do**
    $ranked_\tau(k) \leftarrow \tau_j(max(\rho_{\nu_{k-1},j}))$
  end for
end for
Finding data centers and data clouds: Find peaks of $ranked_\tau(k)$:
for k=1, k++ **do**
  if $[\tau^D(\nu(k-1)) < \tau^D(\nu(k))]$ & $[\tau^D(\nu(k)) > \tau^D(\nu(k+1))]$ **then**
    ν_k is a local maximum
    $\nu_k* \leftarrow \nu_k$ cloud center vector
  end if
end for
Forming data clouds around ν_k*, considering ρ:
for k=1, k++, $k \neq \nu_k*$ **do**
  $Cluster(k) \leftarrow argmax_k(\rho(\nu*, \nu_k))$
end for
Filtering data clouds:
while $size(Cluster)$ is unchangeable **do**
  Computing statistical of clouds:
  for k=1, k++ **do**
    $Cluster_\mu(k) \leftarrow \mu(\rho_{\nu,\nu*})$
    $Cluster_\sigma(k) \leftarrow \sigma(\rho_{\nu,\nu*})$
  end for
  Filtering the data clouds using $Cluster_\mu$ and $Cluster_\sigma$ and τ:
  if $[||\mu_N^i - \mu_N^j|| \leq 2\sigma_N^i]$& $[\tau_N^D(\mu_N^i) < \tau_N^D(\mu_N^j)]$ **then**
    $Cluster(j) \leftarrow [Cluster(j); Cluster(i)]$
  end if
end while
return Clusters

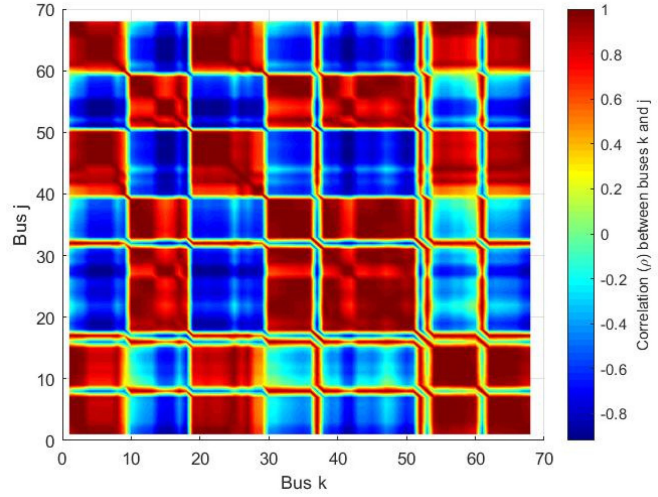


Fig. 2. Correlation map for *Case S1.C1*, for **TDA - Stage II**.

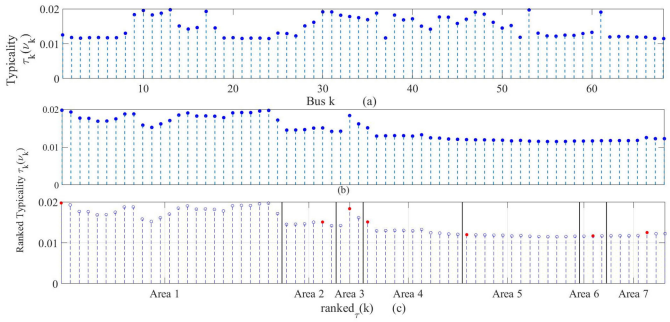


Fig. 3. *Case S1.C1:* Stages III and IV of the **TDA Algorithm 1**.

typicalities indicate that these buses consist of representative frequency responses. Otherwise, these buses are the ones with minor deviations when compared to the other ones in the same data set. It is also a clear indication that these buses have a strong connection within the measured buses. For example, Buses 10–13, 30, 31, 36, 48, 49, 53 and 61, with high typicality values, are part of the meshed area (NYPS area).

Next, in Stage IV, these typicalities values must be ranked starting from the global typicality τ_N^{D*} (maximum typicality value) according to the correlation illustrated in the Fig. 3(b). Where the peaks are the initial centers for each cluster that must be processed using **Algorithm 1 (line 24)**. In Fig. 3(b), the $x - axis$ refers to the position of τ in the ranked vector $ranked_\tau(k)$. At Stage V, the TDA algorithm detects the peaks in the ranked vector to form the initial clusters around those peaks. A filtering process is carried out regarding the mean and standard deviation from the clusters around the peaks. This filtering process takes place until the numbers of clusters does not change. In this case, the algorithm found the solution in three iterations. The final clusters of typicalities are depicted in Fig. 3(c), where the $x - axis$ still displays the buses ranked by the correlation metric.

The resulting seven clusters (areas) are presented in Table II and illustrated in Fig. 4. For comparison purposes with both DDMs and MBAs, Table II summarizes the areas found

TABLE II
AREAS IDENTIFIED BY THE TDA METHOD FOR *CASE S1.C1*

Area - TDA	Coherent Generators	Associated Non-Generator Buses
1	10,11,12,13	17,30-36,38-40,43-51,53,61
2	14,15,16	18,41,42
3	9	28,29
4	1,8	25-27,54-57,59,60
5	2,3	37,52,58,62-67
6	–	21,24,68
7	4,5,6,7	19,20,22,23
Area - DCD [24]	**Coherent Generators**	**Associated Non-Generator Buses**
1	2,3,4,5,6,7	19-24,37,52,56-60,62-68
2	1,8	25-27,55
3	9	28,29
4	12,13	17,34-36,39,43-45
5	10,11	30-33,38,40,46-49,51,53,54,61
6	15,16	18,42,50
7	14	41
Area - SlowCoh. [13]	**Coherent Generators**	**Associated Non-Generator Buses**
1	1,2,3,4,5,6,7	19-29,37,52,55-60,62-68
2	10,11,12,13	17,30-36,38,39,43-51,53,54,61
3	14	41
4	15	42
5	16	18
Area - AP [54]	**Coherent Generators**	**Associated Non-Generator Buses**
1	4,5,6,7	19-24,67,68
2	9	26,28,29
3	10,11,12,13	17,30-36,38-40,43-51,53,61
4	14,15,16	18,41,42
5	1,2,3,8	25,27,37,52,54-60,62-66

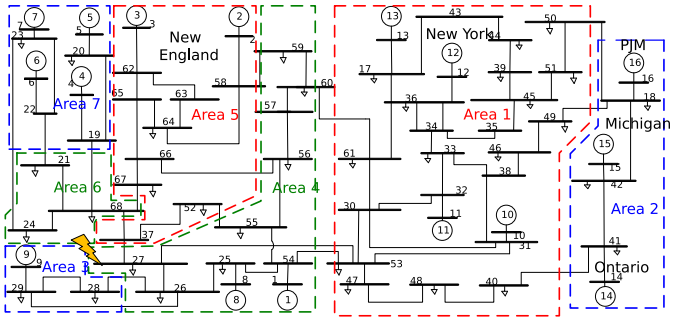


Fig. 4. Areas by the TDA method for *Case S1.C1*.

by [23], [13], and applying the AP algorithm [53] for the same correlation metric, ρ. The results from [13] are the same for all case, since it does not consider events, so it will be shown only in Table II.

Fig. 3(c) illustrates the ranked typicalities, demonstrating that the TDA method exhibits a fine definition of clusters, separating Area 1 from [23] into Areas 5, 6 and 7, shown in Fig. 4. This shows a stronger effect of local modes in the NETS system, which can only be captured if the window length considers the initial transitory period of the frequency response. This effect is not captured by traditional slow-coherency methods. The TDA approach also includes tie-line buses from NETS to NYPS into Area 4, which has the generators electrically closer to NYPS, whereas in [23] those buses get separated into Areas 1 and 5.

However, looking at the closeness in the responses of buses from Areas 6 and 7 in Fig. 5, we can see that TDA is sensitive to very small variations.

This additional information may be used for islanding control schemes purposes, as Area 6 is only comprised of load buses. Such information would not be achievable with MBAs since they construct areas with the consideration that every area has at least one generator. Also, DCD method from [23] would also

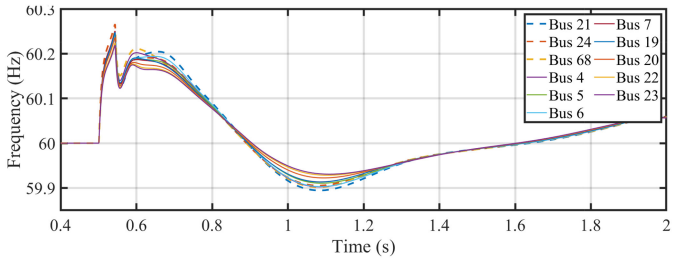


Fig. 5. Frequency signals of Areas 6 and 7 during transitory period.

not be able to detect such an area as it starts its construction of areas by the generators.

It is very interesting to remark that the TDA eccentricity ϵ is calculated using only the data and a distance metric, in this case, the correlation ρ. However, with this value (ϵ) and the first two moments, $\mu(\nu)$ and $\sigma(\nu)$, calculated once TDA clusters the buses, we can address how likely a point in the data-set is of belonging to the cluster, using the proof in Section III.E. In other words, we can attest that the selection of points, i.e., the area, by the TDA method from the data-set distribution is valid, using only the data information and the data distribution information, without the definition of any constant or limit.

To explore the meaning of the areas (clusters) provided by TDA, we show in Table III a summary of the distribution and distance metrics where the mean is adopted as the center of the cluster, and the typicality $\tau^D(\mu)$ of the center of the area calculated using (13), where Area 2 and Area 7 exhibit the lower and higher peak of the local distributions. Notice that these values, i.e., μ and $\tau^D(\mu)$, are equivalent to the mean of a PDF distribution and its peak. Notice that this is extracted exclusively from the data and the distance metric, without any a priori assumption of the PDF. The cluster average $\rho(\nu)$ shows that Area 7 is the most tightly coherent group, since it has the highest correlation average between buses. The largest value of the maximum deviation $\Delta\rho_{\max}$ is found in Area 1 and the smallest one is located in Area 7, showing that these are respectively the least and most coherent areas, in accordance with $\rho(\nu)$. The bus associated with the maximum deviation is displayed in column 6. To prove the correct grouping, the eccentricity ϵ and standard deviation $\sigma(\nu)$, measures are calculated in columns 7 and 8, resulting in the ratio between eccentricity and standard deviation, which shows the conservative probability of the least coherent buses being inside the clusters, due to this ratio being less than $3\sigma(\nu)+1$. This probability is higher, in reality. The overall average correlation ρ_{all} indicates that Areas 1 and 2 are the least coherent with the system. Summarizing the results in Table III, these statistical measures represent a proper clustering pattern, confirming that the clusters provided by the method are correct.

2) Application on Case S1.C2: In this case, a three-phase fault is applied at bus 33 at $t = 0.5\,s$ and cleared after 5 cycles. The TDA method finds 7 areas which are displayed in Table V and illustrated in Fig. 6. Table V also shows the areas for this case using the DCD method from [23] for comparison. We can see that TDA is able to find additional important local oscillations when the compared method fails to do so.

TABLE III
CORRELATION STATISTICS FOR CASE 1 (*S1.C1*)

Area	$\mu(\nu)$	$\tau^D(\mu)$	Avg. $\rho(\nu)$	$\Delta\rho_{max}$	Bus($\Delta\rho_{max}$)	$\epsilon(\nu(\Delta\rho_{max}))$	$\sigma(\nu)$	$\epsilon\backslash\sigma$ ratio	Avg. ρ_{all}
1	1.7946	0.2585	0.9324	0.1118	50	1.2138	0.7689	1.5786	-0.1716
2	2.7639	0.0899	0.9522	0.0152	18	2.1766	7.4491	0.2922	-0.2684
3	1.9645	0.1669	0.9952	0.0022	29	2.4451	2.1275	1.1493	0.1176
4	2.3639	0.1399	0.9552	0.0639	26	2.2177	1.3381	1.6573	0.2587
5	1.9669	0.2141	0.9883	0.0114	67	1.6388	2.1445	0.7642	0.2167
6	1.7422	0.1010	0.9986	0.0006	24	2.1875	1.4153	1.5456	0.2222
7	3.1269	0.3828	0.9990	0.0005	5	2.3993	1.3876	1.7290	0.1087

TABLE IV
CLUSTERING RESULTS WITH DIFFERENT TIME WINDOWS

Sampling rate	0.25s	1s	2s	3s	5s	10s
120 Hz	**X**	**X**	**X**	✓	✓	✓
60 Hz	**X**	**X**	**X**	✓	✓	✓
30 Hz	**X**	**X**	**X**	✓	✓	✓

TABLE V
AREAS IDENTIFIED BY THE TDA METHOD, *CASE S1.C2*

Area - TDA	**Coherent Generators**	**Associated Non-Generator Buses**
1	10,14,15,16	18,30,31,34,35,38,40-42,45-51,53
2	12,13	17,36,39,43,44,61
3	1,8	25,54,57,59,60
4	2	26,27,37,52,55,56,58,62,66
5	4,5,6,7	19-23
6	3,9	24,28,29,67,68
7	11	32,33
Area - DCD [24]	**Coherent Generators**	**Associated Non-Generator Buses**
1	9	28,29
2	10	–
3	11	32,33
4	14	41
5	15	42
6	16	18,50
7	1-8,12,13	Remaining Buses
Area - AP [54]	**Coherent Generators**	**Associated Non-Generator Buses**
1	11	–
2	3,4,5,6,7,9	19-24,29,67,68
3	–	32
4	–	33
5	14,15,16	18,41,42
6	12,13	17,36,39,43,44
7	10	30,31,34,35,38,40,45-51,53,61
8	1,2,8	25-28,37,52,54-60,62-66

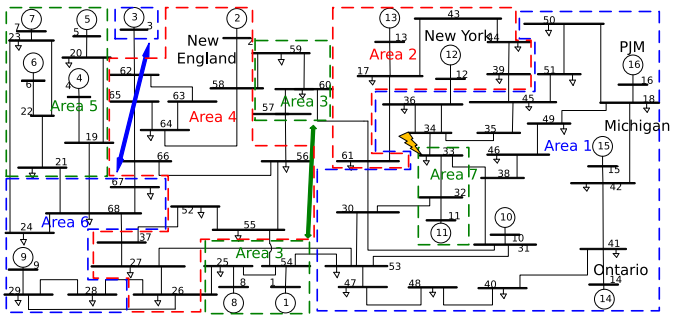


Fig. 6. Areas by the TDA method for *Case S1.C2*.

It is noteworthy to remark that the severity of the fault caused the isolation of the closest generator, i.e., Generator 11, and its closest load bus, bus 33, showing the method captures local modes whereas slow-coherency methods would not, as can be seen in Table V areas provided by TDA, slow-coherency, DCD

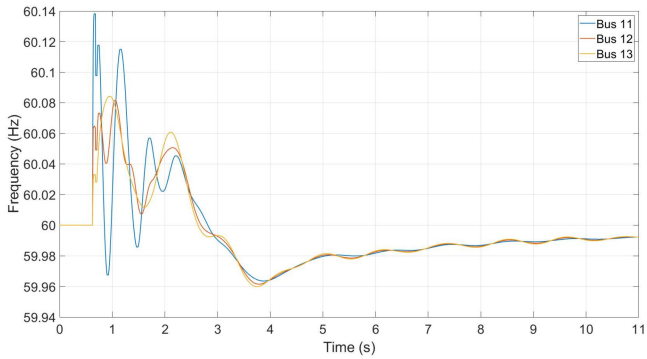


Fig. 7. Areas 2 and 7 generators responses for *Case S1.C2*.

from [23] and the AP algorithm from [53] for the correlation metrics ρ. This separation can also be seen in Fig. 7, in the frequency response of generators 11, 12 and 13, which are traditionally clustered together. Fig. 7 shows that the TDA method makes the appropriate separation, where Generator 11 gets isolated. This fact points out a great advantage of the proposal when compared with MBAs, since these methods would not be able to detect the isolation of this generator.

With this information, operators are able to detect islands in the system that can be intentionally produced aiming to prevent cascading events leading up to blackouts. From Tables II and V and Figs. 4 and 6, we can see the potential of the method for detecting islanding conditions, despite no lines are tripped in those simulations. This information is very valuable for the operator since it can be useful for determining possible parts of the system that can get isolated, without generation (like Area 6 in *Case 1* and part of Area 3 in *Case 2*). It can also be used as an indication of suitable intentional islanding schemes, where Areas 3 and 7 in *Case 1*, and Areas 5 and 7 in *Case 2* could become self-sustained in case of islanding, which can be required for preventing cascading events.

Since the islanding and protection phenomena require responses in faster times, the TDA method is also examined with smaller time windows. It is important to emphasize that a minimal window of 10 cycles must be observed considering the length of the fault (5 cycles) and initial transients. The method is able to detect the isolation of Generator 11 with only 15 cycles, with an average processing time of 24.6 ms, providing the detection of separation in less than a second after the fault. For the base window length of 10 s, we note that Areas 3 and 6 in Fig. 6 are not consecutive, which is also valuable information for deciding islanding control schemes.

TABLE VI
WPPS FOR *S1.C3W*

New Bus	Connection	P_{gen}[MW]
121	121-21	630
127	127-27	630
144	144-44	630
153	153-53	630

Table IV depicts that the method is able to address the same areas for window lengths starting at $3s$, for *Case 2*. For smaller time windows, since the response is dominated by faster modes and more damped modes, the number of areas is greater, indicating mostly local phenomena, such as islanding. It is important to note that for longer windows, the more important fast modes still show in the Areas, such as the isolation of Generator 11. Table IV also shows that the TDA method finds the same results to the three considered sampling rates, which is the case for all simulations, despite of the window length.

B. Application on Case S1.C1, With Presence of Wind Generation (S1.C3w)

The fault in *Case S1.C1* is now applied to a wind generation scenario (*S1.C3w*), enabling the investigation of how non-synchronous generation affects the coherence of the system and the TDA method. Thus, the total load is increased by 20%, homogeneously at all load buses as in [19]. A 13.3% of the additional load is supplied by wind power plants (WPPs), located in four new buses, shown in Table VI. An additional 6.7% generation is distributed by the synchronous generators, from New England system. The choice for installation of the WPPs takes into consideration the concentration of generators (for WPP at bus 121), load buses (for the WPPs at buses 127 and 144), and tie-lines in the case of the third WPP, at bus 153. The new generators are comprised of doubly-fed induction generators (Type-3 wind generators) equivalent models, with 30% of power output injected via inverters with voltage regulation and unitary power factor. The 10% additional generation is equally shared by all four WPPs.

It is noteworthy to remark that, while the increase in load is met by the non-synchronous generation, the transmission system remains unchanged, with the exception of the lines connecting the WPPs to the system. This alters the stability of the system, as the transmission lines may become overloaded, i.e., the poles of the system can come closer to the $j\omega$-axis in the s plane.

Once the initial condition is calculated for the new configuration of the system, the same three-phase fault is applied at bus 27. The frequency responses of the four WPPs is shown in Fig. 8(a), for WPP 127 with zoom after the initial frequency dip for better observation of oscillations, and 8(b) for the three remaining WPPs.

Note that the response from the WPP closer to the fault has a severe frequency dip at the moment of disturbance, due to its closeness to the fault. This behavior is also observed in WPP 121. Also note that the main difference between the responses of WPP 144 and 153 is in the transitory period, which appears in the resulting areas shown next.

The TDA method is applied following the same configurations, i.e., window length, sampling frequency, etc. The areas

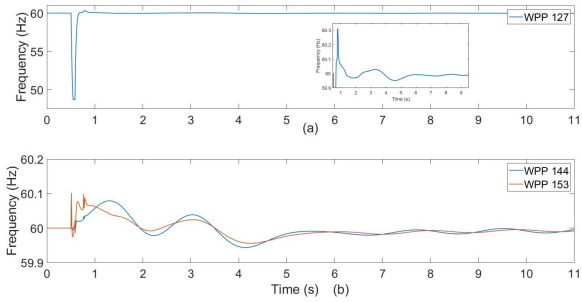


Fig. 8. WPPs frequency responses to fault at bus 27.

TABLE VII
AREAS IDENTIFIED BY THE TDA METHOD, *CASE S1.C3W*

Area - TDA - *S1.C3w*	Coherent Generators	Associated Non-Generator Buses
1	12,13,144	17,36,39,40,43-45,50,51
2	10,11,153	30-35,38,46-49,53,61
3	9	26-29
4	1,8	25,54,57-60
5	127	—
6	2,3	52,55,56,62-66
7	121	21,24,37,67,68
8	5,4	19,20
9	6,7	22,23
10	14	41
11	15	42
12	16	18
Area - TDA - *S1.C1*	**Coherent Generators**	**Associated Non-Generator Buses**
1	10,11,12,13	17,30-36,38-40,43-51,53,61
2	14,15,16	18,41,42
3	9	28,29
4	1,8	25-27,54-57,59,60
5	2,3	37,52,58,62-67
6	–	21,24,68
7	4,5,6,7	19,20,22,23

found by the method are displayed in Table VII, where the areas for the base case are reproduced for better visualization.

As mentioned before, the addition of non-synchronous generation influences the coherency of areas. 12 areas are found, that is, five additional areas, where it can be seen that Area 1 from *Case S1.C1* is split into Areas 2 and 3 in *Case S1.C3w.* Each of these areas has the addition of a WPP, i.e., the WPPs contribute to the coherency of these groups. Additionally, Area 7 from *Case S1.C1* is also split into Areas 8 and 9 for *Case S1.C3w*. The last area from *Case S1.C1* that was split was Area 2, which got separated into Areas 10, 11 and 12, each with a single equivalent generator. The impact of the presence of WPPs can clearly be seen in Fig. 9, where the frequency responses of the generators from Areas 1, 2 and 7 in *Case S1.C1* are plotted as if being grouped in the original case. It is clear that these generators no longer oscillate together, in the new configuration of the system, where WPPs are introduced.

The final additional area, however, is composed only of the new WPP at bus 127. This is reasonable since the non-synchronous generator is close to the fault, and has a degree of isolation from the system through its inverter. The area from Generator 9 gained one bus (the fault bus 27), due to its interaction with the WPP. Also, Area 6 from *Case S1.C1*, which did not possess any generator, gained two new buses (37 and 67) and WPP 121, all being close to the fault. Particularly, these two last effects, i.e., the isolation of WPP 127 and enlargement of the Area 6 show the effect of WPP in the power system, and also, the ability of the proposed method in capturing such events.

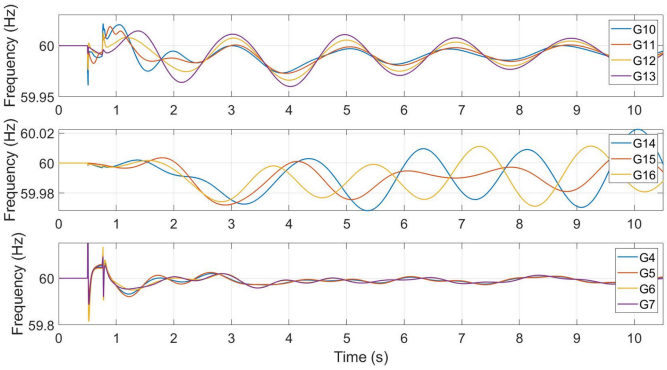


Fig. 9. Generators from Areas 1, 2 and 7 of *Case S1C1* in *Case S1.C3w.*

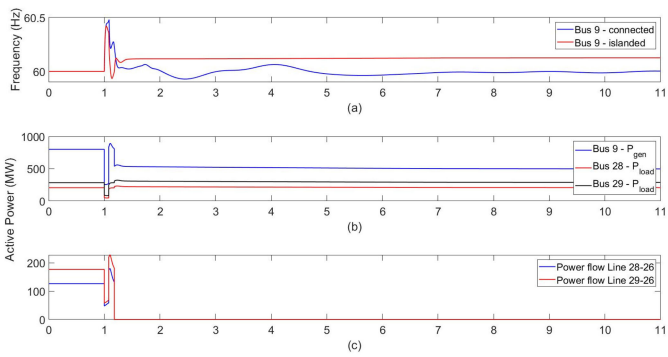


Fig. 10. Generator 9 frequency response: connected and islanded with speed regulator.

We can observe that, as seen in [19], [26], the addition of renewable generation to the system may reduce the damping of oscillations in the system, separating further the areas. This separation is also a consequence of the faster modes added by these plants, which cannot be observed if the transitory response is omitted. However, our estimation of the areas is done without user defined threshold constant γ for clustering as in [19], or complex algorithms like in [26].

C. Application on Case S1.C4i - Islanding Detection

To emphasize the capability of the method in detecting islanding conditions, an additional test is made. *Case S1.C1* is run again, with the lines between buses 28 and 26 and between buses 29 and 26 open, 100 ms after the fault is cleared. This effectively islands Area 3, as those are the only connections of this area to the rest of the system.

In the pre-fault condition of the system, Generator 9 is injecting 800 MW in the system, and the loads at buses 28 and 29 are consuming 206 and 284 MW, respectively. Thus, the lines 28-26 and 29-26 are exporting the remaining 310 MW, minus losses. Due to such pre-fault condition, when the lines are opened Generator 9 would accelerate indefinitely, as it does not contain a speed regulator. A speed regulator is added to Generator 9, as can be seen in Fig. 10(a) and the oscillations in that area cease since there is only one Generator supporting the loads.

The disturbance in generation at Bus 9 and loads at Buses 28 and 29 is shown in Fig. 10(b), and the interruption of power

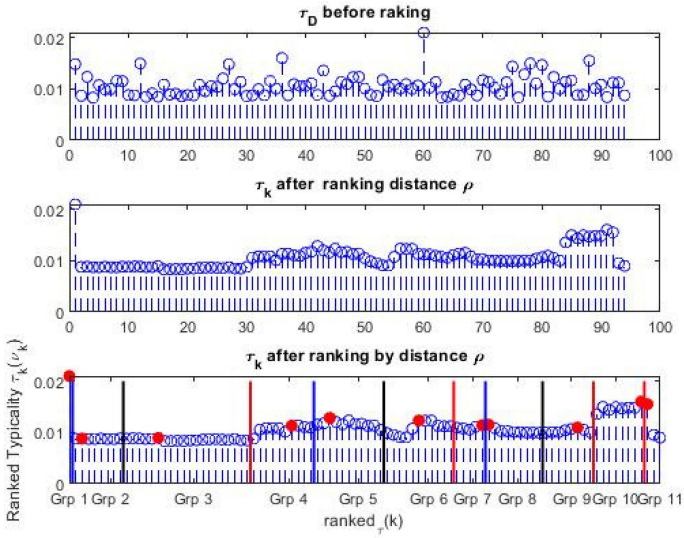


Fig. 11. Stages III and IV of the TDA for the sixth case in the EI system.

transfer from Area 3 to the rest of the system is shown in Fig. 10(c). The TDA method is run again for the PMU data-set of all buses in this case, without the knowledge of separation of those buses. The TDA method properly finds the same areas as the ones showed in Table II, that is, all the coherency detection features, with additional evaluation of Area 3 islanding.

D. Eastern Interconnection (S2)

Next, the TDA is applied to 10 real events recorded in the Eastern Interconnection (EI) by the FNET/GridEye project (a low-voltage WAMS synchronized via GPS [49]). It is noteworthy to remark that the EI system has non-synchronous generators connected and operating [54] which is expected to be handled by the TDA method. All events consisting of generation trips that taken place during the summer season in 2020, from July until September, are considered in this investigation. The number of frequency measurements per event varies from 92 up to 102. For instance, a generator trip occurring on September 12 (the sixth case) is depicted after the filtering process in the first plot in Fig. 12. The filter reduces the noise interference demonstrating the TDA robustness to the minimal remaining noise.

After clustering all above-mentioned events employing the **Algorithm 1**, an average of 7 groups per event are found, with a minimum of 2 and a maximum of 15 groups.

For the sixth case, the detrended frequency responses of the PMUs are displayed in Fig. 12, exhibiting the concept of coherency from (1) in the buses grouping (groups 1 to 11); i.e., the frequency measurements for buses in the same electrical region behave similarly. It is important to point out that the clustering process carried out by the TDA does not impose any user-defined parameter as γ in (1). All buses are clustered using the typicalities and correlation metrics, as shown in Fig. 11. Note that the first group is shaped by a single bus located at the edge of the EI (Bus 3001 in the Sasketchwan Province).

Considering all 10 events, the size of groups varies from 46 buses to single bus; particularly, some buses located at the edge of the EI. The most commonly clustered together buses are

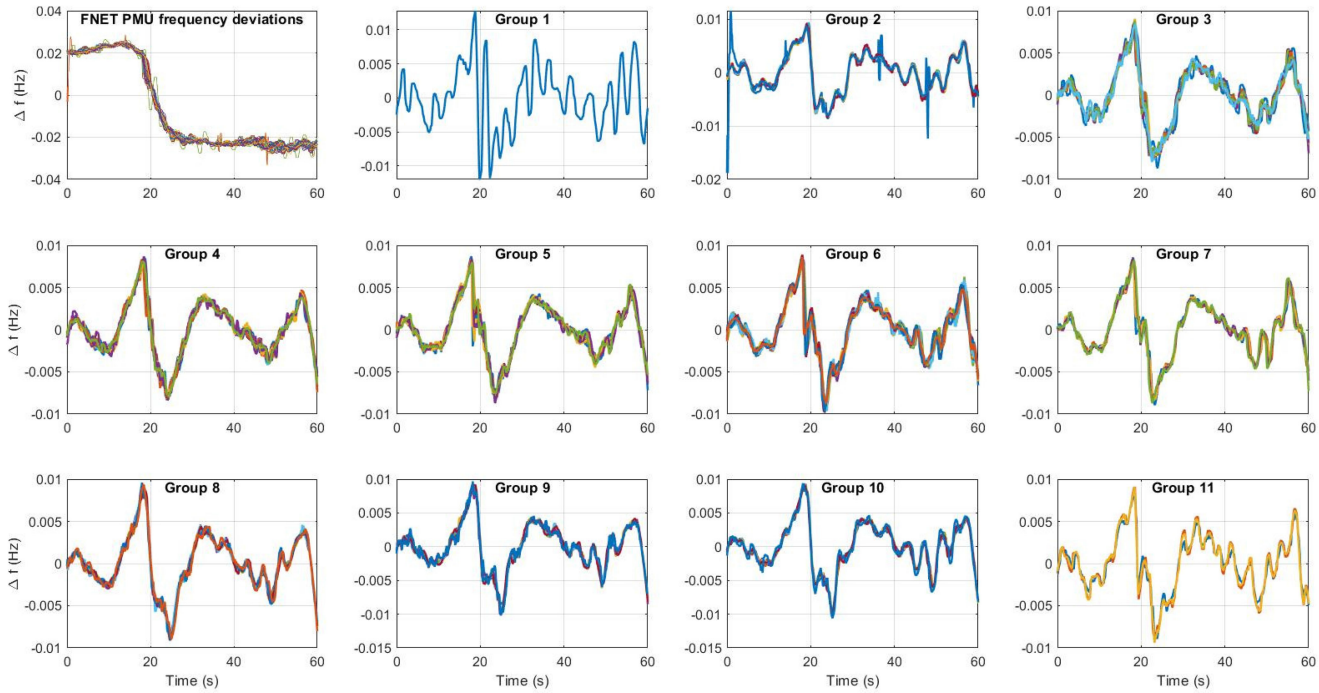


Fig. 12. FNET/GridEye groups for the sixth event.

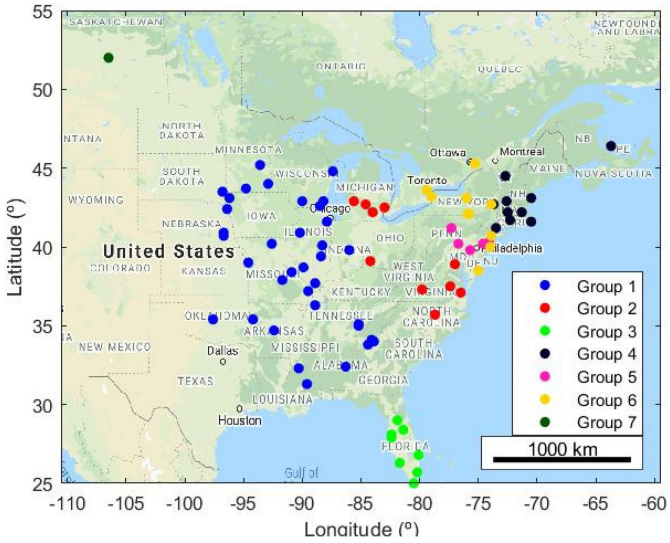


Fig. 13. Geographical distribution of groups in the EI.

depicted in their geographical distribution in Fig. 13, where the proximity of the groups indicates that the method can successfully identify coherent groups. As most of the events consisted of generator trips on the eastern side of the EI, the coherent groups closer to the Atlantic coast are smaller in geographic extent, as such disturbances excited more modes in these areas.

E. Processing Time

Since the method was not implemented on a specific purpose hardware which impacts the time duration. Each case was simulated 10 times to acquire the average time of CPU processing. The TDA method is implemented with MATLAB R2018a on an Intel Core i7-8850 U 2.00 GHz processor with 8 GB of memory, resulting in the average times presented in Table VIII, confirming the computational efficiency of the method to deal with hundredth of measurements in less than 82 ms. For comparison purposes, the work in [23] presents an average of 72.3 ms for 68 measurements. It can be seen that the TDA method is a faster method than DCD and slow-coherency, without requiring load flow results, the number of clusters like slow-coherency, and cutoff coherency constant in both methods, except for frequency measurements. It must also be pointed out that the TDA's execution time also includes the pre-processing stage time.

TABLE VIII
EXECUTION TIME FOR *CASES S1.C1, S1.C2*, AND *S2* AND COMPARISON METHODS

	S1.C1	S1.C2	S2	DCD avg.	Slow-coh.
Run Time (s)	0.04568	0.04213	0.08143	0.0723	0.0692

F. Discussion

For all three cases, the number of iterations, until the final number of clusters is reached, is maximum of 3. It is interesting to note that the correlation metric inherently takes into account the electrical proximity of buses. This is specially important for the clustering process to prevent miss-clustering. The number of areas with the TDA method partially depends on the initial behavior of bus responses, suggesting the importance of local characteristics like weak connections that indicate electrical islands.

TABLE IX
SLOW MODES COMPARISON - S1.C1

Reference	TDA	Error (%)	Slow-coh.	Error (%)	DCD [24]	Error (%)
0.3976	0.4050	1.86	0.3809	4.2	0.3566	10.31
0.6888	0.6365	7.59	0.6790	1.42	0.7861	14.12
1.0433	0.9754	6.50	0.7784	25.39	1.0143	2.77
Avg. Error		**4.39**		10.33		9.07

The islanding condition is evidenced in both cases for the NETS-NYPS grid, where *Case 1* exhibits Area 6 only composed of non-generator buses and *Case 2* has an exclusive area defined by Generator 11 and its closest load bus (bus 33). This result is also confirmed for *Case 2* in Fig. 7. This information, along with additional PMU coordinates, may support operators for islanding detection and define better islanding schemes.

VI. TDA VALIDATION

It is worth noting that DDMs find areas for the event under study, whereas MBAs find areas for all small-disturbance around the equilibrium point [13]. Thus, by comparing the reduced model from the TDA method with MBAs, we quantify how efficient the method is in producing a reduced order model while showing possible islands in the system, for controlled islanding and WAMS monitoring purposes.

Besides the dynamic simulations accomplished for the system *S1* and the TDA application, the PST is also used for its model reduction and linearization, being both applied to the areas found by the TDA and slow-coherency methods, and the ones provided in [23]. The comparison is done by using the slow-coherency [13] aggregation method whose implementation is available in the PST function s_coh3 [51]. This function uses as inputs: the areas found by the clustering method, the number of areas, the data of the system. The areas provided by the three clustering methods (TDA, slow-coherency and DCD) are used as input by the PST slow-coherency aggregation algorithm.

After gaining all three reduced models, their modes are compared with the ones of the complete model for *Case S1.C1*, using the svm_mgen function from PST to linearize the reduced models and extract the modes. This validation is achieved by contrasting the modal information derived with the TDA method against the one resulting from the DCD and slow-coherency techniques. For the sake of brevity, we only show *Case S1.C1*, however the same comparison is accomplished for *Case S1.C2* and the third case in [23]. There is a consistency in the error of the modes throughout the cases and adherence to the real value of the modes. This is of great importance for the validation of the method as a clustering method.

A. Validation Against Existing Methods for Case S1.C1

Table IX depicts the slowest modes obtained from the reduced model provided by the TDA areas compared with the full model and the areas from the slow-coherency and DCD techniques for *Case S1.C1*. The TDA clustering attains the best modes approximation. As mentioned in the previous section, the method performs without arbitrary tuning of coherency parameter γ as required by the compared methods. The proposed method also suggests the islands' detection and it is able to achieve all this within transitory speed conditions.

VII. CONCLUSION

In this paper, a new data-driven method was proposed to track changes of coherent measurements belonging to generator and non-generator buses in large-scale interconnected power systems. This is a non-parametric statistical method that does not require any previous knowledge of the power system dynamics or collected data. As a result, it is not necessary to use any parameter of the system, to specify and tune empirical thresholds or to check if statistical premises necessary to build a formal probability density functions for the data are met. The TDA method was first applied to an equivalent 68-bus test system and the results were compared against slow-coherency (model-based) and DCD (data-driven) methods, exhibiting improvements in terms of modal frequency approximation of reduced order models provided by each method using [18] and [23], respectively. Additionally, test results and validation were carried out using real measurement collected (FNET/GridEye project) from a large interconnected power system (Noth-America Eeastern Interconnection). The application of the method in a real system shown that the approach is robust to real noise and outliers, being capable to present high accuracy and consistent results. From the practical perspective, the method is also capable to detect local areas for islanding and accurate develop reduced order models with low computational burden.

Future work efforts will be to improve potential use of the TDA method for: fault location; area's detection of large frequency variations and consequently estimate the inertia distribution, data-driven center of inertia (COI) estimation, and designing advanced special protection schemes.

REFERENCES

[1] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbič, "Foundations and challenges of low-inertia systems," in *Proc. Power Syst. Comput. Conf.*, 2018, pp. 1–25.
[2] B. Mohandes, M. S. ElN. MoursiHatziargyriou, and S. El Khatib, "A review of power system flexibility with high penetration of renewables," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 3140–3155, Jul. 2019.
[3] H. You, V. Vittal, and X. Wang, "Slow coherency-based islanding," *IEEE Trans. power Syst.*, vol. 19, no. 1, pp. 483–491, Feb. 2004.
[4] C. G. Wang, B. H. Zhang, Z. G. Hao, J. Shu, P. Li, and Z. Q. Bo, "A novel real-time searching method for power system splitting boundary," *IEEE Trans. Power Syst.*, vol. 25, no. 4, pp. 1902–1909, Nov. 2010.
[5] O. Gomez and M. A. Rios, "Real time identification of coherent groups for controlled islanding based on graph theory," *IET Gener., Transmiss. Distrib.*, vol. 9, no. 8, pp. 748–758, 2015.
[6] Z. Lin, F. Wen, Y. Ding, and Y. Xue, "Data-driven coherency identification for generators based on spectral clustering," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1275–1285, Mar. 2018.
[7] S. A. Siddiqui, K. Verma, K. Niazi, and M. Fozdar, "Real-time monitoring of post-fault scenario for determining generator coherency and transient stability through ANN," *IEEE Trans. Ind. Appl.*, vol. 54, no. 1, pp. 685–692, Jan./Feb. 2018.
[8] S. Kamali, T. Amraee, and F. Capitanescu, "Controlled network splitting considering transient stability constraints," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 21, pp. 5639–5648, 2018.
[9] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparsity-promoting optimal wide-area control of power networks," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2281–2291, Sep. 2014.
[10] X. Wu, F. Dörfler, and M. R. Jovanović, "Input-output analysis and decentralized optimal control of inter-area oscillations in power systems," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 2434–2444, May 2016.
[11] G. Y. Babu and V. Sarkar, "Transient instability mitigation via repetitive corrective actions based upon the real-time macrocoherency evaluation," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5084–5095, Dec. 2020.

[12] S. Wang, S. Lu, N. Zhou, G. Lin, M. Elizondo, and M. Pai, "Dynamic-feature extraction, attribution, and reconstruction (UDEAR) method for power system model reduction," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2049–2059, Sep. 2014.

[13] J. H. Chow, *Power System Coherency and Model Reduction*. New York, NY, USA: Springer, 2013.

[14] I. Tyuryukanov, M. Popov, M. A. Van Der Meijden, and V. Terzija, "Slow coherency identification and power system dynamic model reduction by using orthogonal structure of electromechanical eigenvectors," *IEEE Trans. Power Syst.*, vol. 36, no. 2, pp. 1482–1492, Mar. 2021.

[15] E. E. Abraham, H. Marzooghi, J. Yu, and V. Terzija, "A novel adaptive supervisory controller for optimized voltage controlled demand response," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4201–4210, Jul. 2019.

[16] P. Henneaux *et al.*, "Benchmarking quasi-steady state cascading outage analysis methodologies," in *Proc. IEEE Int. Conf. Probabilistic Methods Appl. Power Syst.*, 2018, pp. 1–6.

[17] M. Papic, S. Ekisheva, and E. Cotilla-Sanchez, "A risk-based approach to assess the operational resilience of transmission grids," *Appl. Sci.*, vol. 10, no. 14, 2020, Art. no. 4761.

[18] R. Podmore, "Identification of coherent generators for dynamic equivalents," *IEEE Trans. Power App. Syst.*, vol. PAS-97, no. 4, pp. 1344–1354, Jul. 1978.

[19] A. M. Khalil and R. Iravani, "Power system coherency identification under high depth of penetration of wind power," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5401–5409, Sep. 2018.

[20] M. Naglic, M. Popov, M. A. van der Meijden, and V. Terzija, "Synchronized measurement technology supported online generator slow coherency identification and adaptive tracking," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3405–3417, Jul. 2020.

[21] H. A. Alsafih and R. Dunn, "Determination of coherent clusters in a multi-machine power system based on wide-area signal measurements," in *Proc. IEEE PES Gen. Meeting*, 2010, pp. 1–8.

[22] M. Ariff and B. C. Pal, "Coherency identification in interconnected power system - an independent component analysis approach," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1747–1755, May 2013.

[23] A. M. Khalil and R. Iravani, "A dynamic coherency identification method based on frequency deviation signals," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1779–1787, May 2016.

[24] M. Aghamohammadi and S. Tabandeh, "A new approach for online coherency identification in power systems based on correlation characteristics of generators rotor oscillations," *Int. J. Elect. Power Energy Syst.*, vol. 83, pp. 470–484, 2016.

[25] F. Znidi, H. Davarikia, and K. Iqbal, "Modularity clustering based detection of coherent groups of generators with generator integrity indices," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2017, pp. 1–5.

[26] Z. Lin, F. Wen, Y. Ding, and Y. Xue, "Wide-area coherency identification of generators in interconnected power systems with renewables," *IET Gener., Transmiss. Distrib.*, vol. 11, no. 18, pp. 4444–4455, 2017.

[27] Z. Lin *et al.*, "WAMS-based coherency detection for situational awareness in power systems with renewables," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5410–5426, Sep. 2018.

[28] H. U. Banna *et al.*, "Online coherence identification using dynamic time warping for controlled islanding," *J. Modern Power Syst. Clean Energy*, vol. 7, no. 1, pp. 38–54, 2019.

[29] Y. Susuki and I. Mezic, "Nonlinear Koopman modes and power system stability assessment without models," *IEEE Trans. Power Syst.*, vol. 29, no. 2, pp. 899–907, Mar. 2014.

[30] F. Raak, Y. Susuki, and T. Hikihara, "Data-driven partitioning of power networks via Koopman mode analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2799–2808, Jul. 2016.

[31] H. R. Chamorro, M. Ghandhari, and R. Eriksson, "Coherent groups identification under high penetration of non-synchronous generation," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2016, pp. 1–5.

[32] A. Thakallapelli, S. J. Hossain, and S. Kamalasadan, "Coherency and online signal selection based wide area control of wind integrated power grid," *IEEE Trans. Ind. Appl.*, vol. 54, no. 4, pp. 3712–3722, Jul./Aug. 2018.

[33] M. R. A. Paternina, A. Zamora-Mendez, J. Ortiz-Bejar, J. H. Chow, and J. M. Ramirez, "Identification of coherent trajectories by modal characteristics and hierarchical agglomerative clustering," *Elect. Power Syst. Res.*, vol. 158, pp. 170–183, 2018.

[34] M. M. Farrokhifard, M. Hatami, V. M. Venkatasubramanian, G. Torresan, P. Panciatici, and F. Xavier, "Clustering of power system oscillatory modes using dbscan technique," in *Proc. North Amer. Power Symp.*, 2019, pp. 1–6.

[35] I. Kamwa, A. K. Pradhan, and G. Joós, "Automatic segmentation of large power systems into fuzzy coherent areas for dynamic vulnerability assessment," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1974–1985, Nov. 2007.

[36] T. Guo and J. V. Milanovic, "Online identification of power system dynamic signature using PMU measurements and data mining," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1760–1768, May 2016.

[37] H. M. Al-Masri and M. Ehsani, "Impact of wind turbine modeling on a hybrid renewable energy system," in *Proc. Ind. Appl. Soc. Annu. Meeting*, 2016, pp. 1–8.

[38] F. Znidi, H. Davarikia, M. Arani, and M. Barati, "Coherency detection and network partitioning based on hierarchical dbscan," in *Proc. IEEE Texas Power Energy Conf.*, 2020, pp. 1–5.

[39] P. P. Angelov, X. Gu, and J. C. Príncipe, "A generalized methodology for data analysis," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2981–2993, Oct. 2018.

[40] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112, New York, NY, USA: Springer, 2013.

[41] G. E. Batista, E. J. Keogh, O. M. Tataw, and V. M. DeSouza, "Cid: An efficient complexity-invariant distance for time series," *Data Mining Knowl. Discov.*, vol. 28, no. 3, pp. 634–669, 2014.

[42] P. P. Angelov and X. Gu, *Empirical Approach to Machine Learning*. Cham, Switzerland: Springer, 2019.

[43] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Netw.*, vol. 1, no. 3, pp. 215–239, 1978.

[44] J. G. Saw, M. C. Yang, and T. C. Mo, "Chebyshev inequality with estimated mean and variance," *Amer. Statistician*, vol. 38, no. 2, pp. 130–132, 1984.

[45] P. W. Sauer, M. Pai, and J. Chow, *Power System Dynamics and Stability*. John Wiley Sons, New Jersey, USA, 2016.

[46] I. S. Association *et al.* "C37. 118.1-2011 IEEE Standard for Synchrophasor Measurements for Power Systems," Tech. Rep., 2011.

[47] C. Lackner, J. Chow, F. Wilches-Bernal, and A. Darvishi, "Voltage control performance evaluation using synchrophasor data," in *Proc. 53rd Hawaii Int. Conf. Syst. Sci.*, 2020, pp. 1–10.

[48] B. Pal and B. Chaudhuri, *Robust Control in Power Systems*. Springer Sci. Bus. Media, New York, USA2006.

[49] Y. Zhang *et al.*, "Wide-area frequency monitoring network (FNET) architecture and applications," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 159–167, Sep. 2010.

[50] J. H. Chow and K. W. Cheung, "A toolbox for power system dynamics and control engineering education and research," *IEEE Trans. Power Syst.*, vol. 7, no. 4, pp. 1559–1564, Nov. 1992.

[51] J. Chow, "Power system toolbox," 2020. Accessed: Aug. 2019, [Online]. Available: https://www.ecse.rpi.edu/ chowj/PST_2020_Aug_10.zip

[52] C. Canizares *et al.*, "Benchmark models for the analysis and control of small-signal oscillatory dynamics in power systems," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 715–722, Jan. 2017.

[53] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[54] A. Bloom *et al.*, "Eastern renewable generation integration study," Nat. Renewable Energy Lab.(NREL), Golden, CO, USA, Tech. Rep. NREL/TP-6A20-64472, 2016.