IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 28, 2021, accepted April 17, 2021, date of publication April 20, 2021, date of current version May 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074562

Synergizing PMU Data From Multiple Locations in Indian Power Grid-Case Study

**MAKARAND SUDHAKAR BALLAL[1], (Senior Member, IEEE),
AND AMIT RAMCHANDRA KULKARNI[2], (Student Member, IEEE)**
[1]Department of Electrical Engineering, Visvesvaraya National Institute of Technology, Nagpur 440010, India
[2]Maharashtra State Electricity Transmission Company Ltd., Prakashganga, Mumbai 400051, India

Corresponding author: Amit Ramchandra Kulkarni (amitkul2019@rediffmail.com)

ABSTRACT Synchrophasor technology improves power grid visibility by installing phasor measurement units (PMUs) over a wide area in the power system. Big data received from PMUs contains important information about grid behavior. This information is useful in monitoring the safety and security of the grid. An extensive state-of-the-art review of big data analytics and its prime applications in power systems are expressed in this paper. It presents, a general background in data analysis techniques such as exploratory data analysis, statistical data analysis, and unsupervised data mining techniques like clustering. Two 400 kV transmission line tripping events are analyzed from the data recorded by the PMUs installed in the western part of the Indian grid i.e., Maharashtra State Electricity Transmission Company Limited (MSETCL) grid. The box plots, Correlogram, and the formation of clusters carried out for the PMU data recorded under ambient and disturbance events. This provides insights on how effective big data helps to make the right decision at right time for effective management of the power grid under normal and contingency conditions.

INDEX TERMS Phasor measurement unit (PMU), box plot, connectivity index, correlation technique, Dunn index, Hierarchical clustering, *K*-means clustering, partitioning around medoids (PAM), Silhouette width.

I. INTRODUCTION

Synchrophasor technology is a potential tool for diagnosing, preventing, and control for the grid system. Synchrophasor is a phasor measurement unit (PMU) used for finding the health of the power grid. The Indian power grid has also taken major steps to implement smart grid technology like a wide-area measurement system (WAMS) after blackouts in the northern and north-eastern parts of India in July-2012 by launching a pilot project. Power transmission utilities have installed PMUs for the reliable wide-area monitoring, protection, and control (WAMPAC) system. The national WAMS project called Unified Real-Time Dynamic State Measurement (URTDSM) project includes the furnishing of 1740 PMUs in the transmission terminals i.e. extra-high voltage (EHV) substations across India. Thirty-two Phasor Data Concentrators (PDCs) shall be installed at the State level, Regional level, and National level.

Super PDCs will collect data from PMUs at inter-State transmission stations and inter-State power plants as well as from Master PDCs. Super PDC shall be installed at each Regional Load Dispatch Center (RLDC). They will be equipped with data storage, analytical tools, and a graphical software package. For better management, the Indian power grid is geographically divided into five regions. These regions are Northern, Eastern, Western, North Eastern, and Southern grids. Therefore, the five super PDCs shall be placed at five RLDCs. At the end of 2020, PMU placement is greater than 1500 in India. After the installations of all these PMUs under the WAMS project, the data collected by PMUs and their applications shall increase manifold [1].

Research articles relevant to the application of PMU technology are rigorously surveyed in [2]. It represents a panorama of research progress lines. A discerned compression method that employs the ingrained correlation within PMU data by dimensional and temporal redundancies is described in [3]. In [4], the wavelet transform and matrix pencil (WTMP) approach is used to extract the dominant oscillations and power system parameters from wide-area measurements. The singular value decomposition (SVD) is applied to abate the covariance signals attained by the natural excitation technique. Only thereafter, the orders and range of

The associate editor coordinating the review of this manuscript and approving it for publication was Hazlie Mokhlis.

the corresponding frequency are estimated by SVD from the positive power spectrum matrix. The computation data size is also reduced.

In [5], an online data-driven technique is explained for the disclosure of substandard synchrophasor measurements. This method clouts the Spatio-temporal analogies with different-time-instant synchrophasor measurements and contrives the low-grade synchrophasor data as Spatio-temporal outliers. In [6], event detection and its characterization are explained according to physical disturbance. Some statistical features are extracted from the transient signals to demonstrate the associated practical phenomenon in every event type. In [7], the big data issues of the smart grid are supervised. The multisource of energy data and features of a smart grid are explained. Some theoretical and practical implemented applications for big data analysis are demonstrated. A short history of the PMU concept and applications of synchrophasor technology are discussed in [8]. In [9] feasible solution for wind power management using data mining algorithms for an enormous quantity of data accumulated from the PMU is described. However, the defined clusters are used as subservient variables using classification and regression tree algorithms. In [10], PMUs have been treated for detecting disruptions and deflation in the grid. However, this requires time-based clustering as it limits the PMU population to search in a small limited region.

WAMS-based coherency detection methodology for renewable power generation is explained in [11]. The Kernel principal component analysis, coherency detection method, and clustering analysis vested on intimacy propagation are applied. It is observed that the weakly damped oscillations induced by a large penetration level of renewable power resources can be detected correctly. Thus, it is possible to enhance the situational awareness of the anxious power system with momentous uncertainty. A bagged equating of multiple linear regressions model to recoup the uncounted synchronized frequency data is explained in [12]. This method is based on altogether learning by resetting and equating numerous linear regressions to estimate the missing data. An online event identification technique in a wide-area power system by PMU is described in [13]. In this, offline zonal analysis and an online event location estimation technique are combined. An assessment of the conspicuous aspects in large data analytics advancement in the power systems is given in [14]. Here, the classifications of the enduring and the faltering ingredient in the structures are executed. In [15], the state-of-the-art techniques corresponding to big data in the smart grid have been reviewed. Efficient data investigation for a big volume of data is challenging in power systems due to the incorporation of more advanced information and communication technology. A multi-scale PMU data compression technique by clustering inquiry of wide-area power systems is demonstrated in [16]. Spatial clustering (density-based) with noise is applied for the preconditioning of synchrophasor data.

PMU data corresponding to phase angle difference of bad quality is screen out in [17] by clustering technique. However, if the error in phase angle difference is less, it may not be possible to discriminate between the correct data and wrong data. A critical review of various methods carried out in [18] for the optimal placement of PMUs and their utilization. An adaptive Matrix Pencil algorithm entrenched on wavelet soft-threshold de-noising is explained in [19] to deal with the low-frequency oscillation (LFO) signals deduced from the WAMS in power systems. The identified LFO signal can be fitted only after calculating the appropriate modal parameters of the signal. The accuracy of the fitting index (AFI) is used to indicate the similarity between the fitting signal and the genuine signal. In general, efficient mode recognition can be acknowledged only if the value of AFI is more than 10 dB. Otherwise, it is imperative to endeavor for mode identification.

Gaussian assimilation clustering approach is explained in [20]. This is for the characterization of the errors noticed in PMU data. That might be due to the saturation of the current transformer and/or burden of the connecting cables. In [21], the data compression is performed on PMU data received from eight important substations of the power grid. It accommodates two major events in the Maharashtra State Electricity Transmission Company Limited (MSETCL) transmission network. The data compression is carried out by singular value and the eigenvalue decompositions. Effective grid management is possible if the correct and useful data is accessible by the system operator. The data useful for effective grid management consist of parameters, such as the magnitude and phase of the current and voltage phasors, frequency, rate of change of frequency (ROCOF), active and reactive power flow, the status of the circuit breakers, oscillations during ambient and contingency conditions, etc. This data is used to examine power system dynamics, such as the effect of a disturbance as it disseminates to nearby regions. Thus, it creates transient real-time swings in the system. The data applications cover wide-area monitoring, fault location, state estimation, islanding detection, protective relaying, etc. These applications help real-time grid operations by furnishing wide-area power system visualization and situational awareness.

Synergy manifests participation, association, or combine response of components connected in the system. The convention of synergy tells that the joint interaction of system components attends a result greater than the elemental sum of parameters. In a modern power system, PMUs are commissioned at important substations and at power plants for WAMS applications. PMUs collect the data all the time (24 hrs. × 7 days) and feed it to load dispatch centers. Whenever any disturbance appears in a power grid, this PMU data contains important information about the nature of the disturbance. The data of every PMU have certain co-relation or attachment pertain to that specific disturbance. The union or alliance or connection of the data among all the PMUs is termed Synergy. Having synergy means trust, collaboration, connectivity, and ultimately co-creation among the data received from PMUs, it helps to create prominent effects and results. It also propagates better solutions to problems and accomplishes the vision and mission of power utilities. The objective of this research is to identify the synergy among the data obtained from various locations in the powers system. This is achieved by applications of data analysis techniques.

The contributions of this research paper are summarized as follows:

1. The outlier detection using box plots analysis was performed on PMU data of MSETCL substations to determine the power system behavior under normal and abnormal conditions. In data mining, outlier detection is meant for the determination of patterns in data.

2. The correlation technique is applied to PMU measurements to quantify the degree of system parameters from different locations during events. This particular type of analysis is useful to the system operator for efficient grid management.

3. Thereafter, the clustering technique is applied to practical PMU data by which a given data set is divided into groups called clusters. This prevents the loss of valuable time and information if a server fails.

The remainder of this paper is harmonized as follows. Section II presents a general background in data analysis techniques. Section III of the paper describes line tripping events in the MSETCL grid under investigation. Section IV illustrates in detail the application of various data analysis techniques to synchrophasor data under investigation along with its interpretation. Conclusions are given in Section V.

II. PRELIMINARIES

Good quality data having hidden information about system behavior is required to analyze and also to upgrade the attainment of the power system. A brief overview of some of the important data analysis techniques applied to PMU data is given in this section. It includes exploratory data analysis techniques like box plots and statistical analysis techniques like correlation. It also covers brief information on popular unsupervised data mining techniques like clustering. These techniques are applied to PMU data available during disturbance to have better insight for power system management.

A. EXPLORATORY DATA ANALYSIS TECHNIQUE: BOX PLOTS

The box and whisker plot is represented by the box at the center, with three quartiles marked along-with two whiskers at both sides of the box. These two whiskers on both sides touch maximum and minimum points in the data. Figure 1 illustrates a typical box plot. It depicts data distribution information like Minimum; first quartile (Q1) is that tab under which 25% of data points lie, whereas 50% data is down (Q2) or Median. The third quartile (Q3) is that mark below which 75% of data points lie.

The difference between the Q1 and Q3 is termed as Inter-Quartile Range (IQR). The minimum value is given by Q1-(3/2)IQR, whereas the maximum value is given by Q3 + (3/2)IQR. For any arbitrary number "n", the data point n × (Q3- Q1) above Q3 and n × (Q3 - Q1) below Q1 gives outliers [22]. Thus information on outliers and related values is obtained. It also gives information related to data like its symmetry, whether it is tightly bounded or skewed in nature.

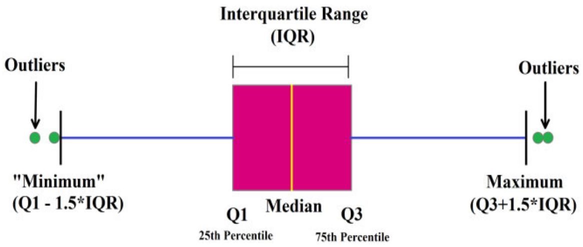


FIGURE 1. Typical box plot representation.

B. STATISTICAL DATA ANALYSIS TECHNIQUE: CORRELATION ANALYSIS

The correlation indicates a relationship among two or more variables. The correlation or interrelationship coefficient is an outcome of correlation analysis and is obtained in terms of a value establishing a bond within the variables. It denotes the strength and direction of association amongst variables. The Pearson's Correlation Coefficient (PCC) is one of the widely used interrelationship coefficients to understand correlation amongst variables. When applied to sample data, it is called as Sample Pearson Coefficient. PCC measures the statistical correlation among two reciprocated variables. It is recognized as an excellent method of aligning the alliance with variables of significance as it is dependent on the covariance method [23]. It provides intelligence about the amplitude and the direction of the association or relationship.

Let us consider, X and Y are the PMU measurements from two different substations. To investigate correlations among X and Y, the PPC formula is given as,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where x_i and y_i are the individual sample points indexed with i and the mean of sample x and y are given as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The correlation coefficient bounds between −1 and 1. A value 1 denotes a linear equation showing perfect consociation between X and Y. It means all data points on a line for which Y increases as X increases. Whereas, a value −1 indicates all data points on a line for which Y decreases as X increases. And a value of 0 depicts that there is no linear relationship within the variables [24]–[25].

The points describe the approved protocol for translating the correlation coefficient as 0 for no linear relationship,

< 0.9 for a very high relationship. Values of correlation coefficients in the ranges 0.7 to 0.9, 0.5 to 0.7, and 0.3 to 0.5 stand for a high, moderate, and low correlated relationship. The *corrplot()* function in "*R*", the software used for analysis conceives a graphic array of a correlation matrix [26]. It highlights the topmost connected variables in a data table. This is used for the analysis purpose in this paper to establish a correlation between PMU measurements obtained from various field locations.

C. DATA MINING TECHNIQUE: CLUSTERING

Data mining is specified as a process passed down to select usable data for valuable intelligence from a large volume of raw data. This relates to extrapolating patterns and additional information from collected data [27]. Though there are several data mining techniques, the important ones include; pattern recognition, allocation, organization, aberration, accumulation, retrogression, and forecasting. Out of these various data mining techniques, clustering which is the unsupervised learning technique having objects similar to one another put in the same cluster is focused for its use with smart-grid data. The three important clustering techniques used for PMU data analysis are as follows.

1) *K*-MEANS CLUSTERING

K-means is an elementary unsupervised training algorithm that deals with the clustering issue. The main idea is to designate *K* centers for every cluster. The key idea of *K*-Means clustering is that it tries to decrease the variation within-cluster. The determination of within-cluster variation is instinctive and there are numerous ways to clarify it.

Let the cluster is C_i, $i\epsilon\{1, 2, \ldots K\}$ and $W(C_i)$ is the cluster variation of C_i. *K*-Means algorithm lessens the sum of *K* within-cluster variations [28]. In analytical terms, it solves the ensuing optimization problem.

$$Minimize\{\sum_{i=1}^{K} W(C_i)\} \tag{2}$$

In this case, within-cluster variation is determined by Euclidean distance given as

$$W(C_k) = \frac{1}{|C_k|}\sum_{i,i'\in=C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 \tag{3}$$

In (3), $|C_k|$ is the number of elements in cluster C_k. The double sum determines the squared Euclidean distance. The computational cost is significantly reduced by considering the squared distance.

The inner sum in (3) is for *j* dimensions in the data. For all *j* dimensions, it is catching the deviation, and squaring it, and adding it all together. The internal sum, *i* and *i*′ are firmed and both correspond to the cluster C_k. The exterior sum is choosing all the accessible pairs of observations in a cluster. If there are n_k observations or data points in the clusters, then the inner sum of that cluster is given by

$$\binom{n_k}{2} = \frac{n_k(n_k-1)}{2} \tag{4}$$

Therefore, the ratio of summation of pairwise Euclidean distances to the number of elements in the particular cluster gives "within-cluster" variations. Following are the important steps in the *K*-Means clustering algorithm:

i. Randomly select *K* centroids in the data points or observations
ii. Determine the Euclidean distance of all the observations across the centroids.
iii. Ascribe each observation to a cluster, positioned on the minimum squared distance against the centroids.
iv. After step 2, absolute members in *K* clusters are acknowledged. Now from these members, calculate the centroids of the clusters.
v. Go back to step 2, re-appoint all the observations to clusters, from the recent centroids.
vi. If the cluster accreditation is similar to the previous, then conclude. This point is taken as the converging point otherwise restart from step ii.

In this way, a new compelling is carried out among the same data set points and the closest new center. A loop has been obtained. From this loop, it came to notice that the *K* centers revise their location gradually until centers do not move anymore. Finally, this algorithm minimizes an objective function given in (3).

2) HIERARCHICAL CLUSTERING TECHNIQUES

Hierarchical clustering is an algorithm that clubs identical objects into block called clusters. Therefore, this is a set of nested clusters that can be arranged as a tree. This clustering technique is branched into two types [29].

1. Agglomerative Hierarchical Clustering Technique: At the beginning, each data point or observation is considered as an independent cluster. It is assumed that every observation gives rise to an individual cluster. At each stage of implementation of this technique, a new bigger cluster is formed by joining the two most similar clusters. This technique follows the bottom-up approach. The matching clusters are iteratively blended with other clusters until *K* clusters are formed. The algorithm is as below.

i. In the beginning proximity of each observation/ data point is noted and computes the proximity matrix.
ii. Combine two similar data points into a single big cluster which is also called a node.
iii. This iterative process is continued, till all the observations are part of a single big cluster. In this way, all the clusters are combined and form one cluster. The refrain cluster package is a C++ library for hierarchical, agglomerative clustering. It caters to the fast execution of the algorithms when the input is an inconsistency index.

2. Divisive Hierarchical Clustering Technique: This is a top-bottom approach. All observations outset in one cluster

and divisions are discharged recursively as one reaches the bottom. It is opposite to the Agglomerative Hierarchical clustering. It is not much used in the real world. At the early stage, all the data points are treated as a single cluster. Those data points which are not identical are segregated iteratively in different clusters. Each data point that is separated is considered as a singular cluster. In the end, n clusters are available.

The Hierarchical Clustering Technique can be anticipated by a tree-like diagram that subscribes to the sequences of grouping or splits. This diagram is called a Dendrogram. To decide dissimilarity between two clusters of observations, Hierarchical clustering uses different linkage methods/agglomeration methods. The important ones are listed in [30] with related distance update formulae and information on cluster dissimilarity. Median is expressed as,

$$\sqrt{\frac{d(I,K)^2}{2}+\frac{d(J,K)^2}{2}-\frac{d(I,J)^2}{4}} \quad (5)$$

A new cluster is assumed to have been formulated by joining clusters I and J, whereas K is any other cluster. The size of clusters I, J, and K are is denoted by nI, nJ, and nK respectively.

3) PAM CLUSTERING ALGORITHM

PAM meant for Partition Around Medoids. The algorithm is used to determine a string of objects called medoids that are posted at the center in clusters. Medoids are indicative objects of a cluster with a data set. The average disparity of Medoids to all the objects in the cluster is nominal. Medoids are analogous in perception to means or centroids, but they are always restrained as members of the data set. This algorithm tries to reduce the overall dissimilarity of objects nearest to the selected object [31]. The algorithm works in the following two phases.

1. BUILD Phase: The BUILD phase entails the following steps

i. Initialize set S of the selected object and sum to it an object for which the total average distances to the rest of the objects is minimum.

ii. If O is taken as the set of objects then the set of unselected objects is $U = O - S$. Let an object $i \in U$ as a candidate for insertion into the set of preferred objects.

iii. For an object $j \in U - \{i\}$ estimate D_j, the dissimilarity in j and the nearest object in S.

iv. If $D_j > d(i, j)$ then only object j will tender to the judgment to choose object i as it benefits the quality of the clustering. Thus, $C_{ji} = \max\{D_j - d\ (j, i), 0\}$.

v. Calculate the total gain accessed by summing i to S as $G_i = \sum C_{ji}$.

vi. Select object i that maximizes G_i. Thus, $S = S\cup \{i\}$ and $U = U - \{i\}$.

These steps are continued until K objects have been confirmed.

2. SWAP Phase: The SWAP phase endeavor to upgrade the set of chosen objects to filter the ingredient of the clustering. This is made by taking all pairs $(i, h) \in S \times U$ and estimating the effect T_{ih} on the addition of heterogeneities between objects and the nearest preferred object caused by interchange i and h, that is, by placing i from S to U and placing h from U to S. The determination of T_{ih} involves the calculation of the contribution K_{jih} of each object $j \in U - \{h\}$ to interchange i and h. We have considered $d\ (j, i) \geq D_j$.

The software package "R" is used for clustering analysis and the *clValid* () function is utilized for validating the results [32]. It comprises clustering algorithms like Hierarchical, K-Means, and PAM. These are used in this paper to determine the goodness of clustering algorithms applied to the PMU dataset under investigation. While doing internal validation using this function, the quality of clustering is assessed by taking the dataset and its clustering partition as input. For the internal validation of clusters, measures like connectedness, closeness, and segregation of the cluster dissolutions are selected. The compactness is estimated by cluster homogeneity based on intra-cluster variance. The separation indicates the degree of dissociation between clusters.

The popular indices like Dunn index [33] and Silhouette width [34] combine both the measures compactness and separation into a single score. By using *clValid* () function the user can concurrently select different clustering algorithms, acceptance measures, and numbers of clusters in only one function call. It also determines the most appropriate clustering technique and an optimal number of clusters for the given dataset. The clustering validation for PMU data presented in this paper is done by using the following three popular indices.

a: CONNECTIVITY

The connectivity reveals to what grade data points are located in the exact cluster as their closest companions in the data space and is measured by the connectedness [35]. Let $nn_{i(j)}$ be the j^{th} closest companions of data point i. Also, consider x_i, $nn_{i(j)}$ equal to zero if i and $nn_{i(j)}$ are in a similar cluster and $1/j$ by any other way. Then, for a peculiar clustering partition, $\zeta = \{C_1, \ldots\ldots C_K\}$ of the N observations within K dissociate clusters, the connectivity is expressed as,

$$Conn(\zeta) = \sum_{i=1}^{N}\sum_{j=1}^{L} x_i, nn_{i(j)} \quad (6)$$

Here L is the parameter that determines connectivity measure as contributed by the number of neighbors. The value for connectivity lies between zero and ∞ is expected a minimum.

b: DUNN INDEX

The Dunn index is defined as the ratio of the littlest distance within observations not in a similar cluster to the greatest intra-cluster distance [31]. It is calculated as

$$D(\zeta) = \frac{\overset{\min}{C_k, C_l} \in \zeta(i \in \overset{\min}{C_k}, j \in C_l\, dist(i,j))}{\max_{C_{m\in\zeta}} diam(C_m)}, C_k \neq C_l \quad (7)$$

The maximum distance between data points in a cluster C_m is denoted by *diam* (C_m). The value for the Dunn index lies between zero and ∞ is supposed a maximum.

c: SILHOUETTE WIDTH

The Silhouette value of a particular data point represents the degree of certainty in the clustering. The average of each observation's Silhouette value is termed as Silhouette width [32]. When the data point is neatly clustered its value is close to 1, whereas when it is disorderly clustered it is close to-1. For observation i, it is defined as

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{8}$$

where a_i is the average distance between i and all other data points in the same cluster, and b_i is the average distance between i and the data points in the nearest neighboring cluster, i.e.

$$a_i = \frac{1}{n(C(i))} \sum\nolimits_{j \in C(i)} dist(i,j) \tag{9}$$

$$b_i = \overset{\min}{C_k} \in \zeta \backslash C(i) \frac{1}{n(C(i))} \sum\nolimits_{j \in Ck} \frac{dist(i,j)}{n(C_k)} \tag{10}$$

where $C(i)$ is the cluster that consists of observation i, dist(i, j) is the distance (e.g. Minkowski, Manhattan, Cosine, Hamming, Euclidean, etc.) between observations i and j. The cardinality $n(C)$ is for cluster C. The Silhouette width is in the interval $[-1, 1]$.

Figure 2 describes a flowchart indicating the application of clustering techniques and internal validation measures as applied to the PMU dataset under investigation. This gives information on the number of clusters formed. Applications of these clustering techniques to identify the synergy among the data received from PMUs placed at various locations in the power system are demonstrated in the following sections.

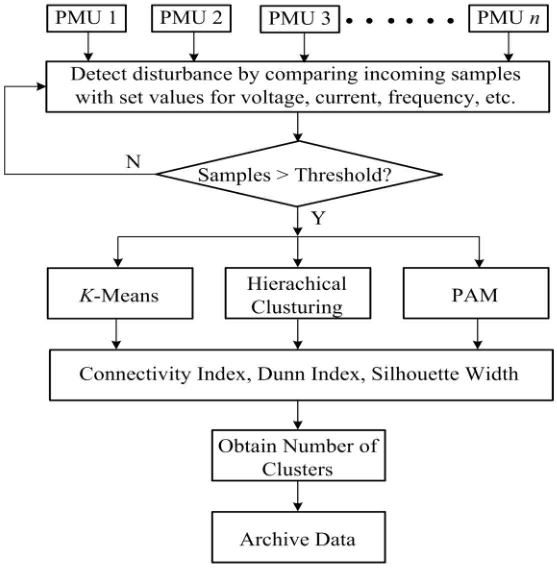

PMU 1
PMU 2
PMU 3
PMU n
Detect disturbance by comparing incoming samples with set values for voltage, current, frequency, etc.
N
Samples > Threshold?
Y
K-Means
Hierachical Clusturing
PAM
Connectivity Index, Dunn Index, Silhouette Width
Obtain Number of Clusters
Archive Data

FIGURE 2. **Flowchart indicating the application of clustering techniques and validation measures to PMU data.**

III. SYSTEM UNDER STUDY

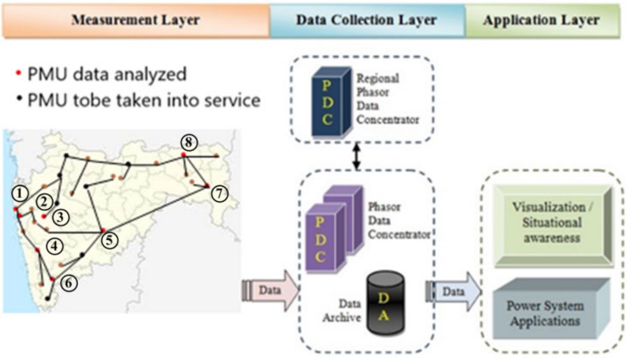
Maharashtra State Electricity Transmission Company Limited (MSETCL) is the largest electric power transmission utility in the State sector in India. It has 34 EHV substations of 400 kV and above level and 647 substations of 220 kV and below voltage level. Its transformation capacity is 128990 MVA and reactive Power Compensation is 5508 MVAR. This transmission system is capable to handle about 21000 MW of power. MSETCL comprises intra and inter-regional power transmission lines [21]. Important features of the MSETCL are given in Table 1.

TABLE 1. **Important features of the MSETCL [21].**

SN	Particulars	Details
1	EHV substations	681
2	HVDC Terminals	2
3	Transmission Line	48321 ckt. kms #
4	PMU Details	
	i. Number of PMU installed	15
	ii. IEEE standard	C37.118
	iii. Data rate	25 samples/second
	iv. Vector Error	> 1 percent
	v. Range	± 5% of the nominal frequency

1 km length of double circuit = 2 ckt. kms.

PMUs are connected at prime EHV substations in a URTDSM project begun by the Government of India. Commissioning of PMU in 15 EHV substations covers 67% of the MSETCL grid. The three-layered WAMS architecture of the MSETCL network is shown in Figure 3. It indicates PMUs commissioned in a few 220kV and 400kV substations. Some of them are connected to the major power plants and the ±500 kV HVDC system.



1. Padghe, 2. Kalwa, 3. Lonikand, 4. New Koyana, 5. Girawali, 6. Lamboti, 7. Chandrapur and 8. Koradi

FIGURE 3. **Three-layered WAMS architecture in MSETCL [21].**

The PMUs measure the data consist of eleven parameters of a power system. The general parameters bus voltages (V_a, V_b, and V_c) and line currents (I_a, I_b, and I_c), are used in this paper to check the synergy of data with each other. PMUs sent this data to the PDC placed at the State load dispatch

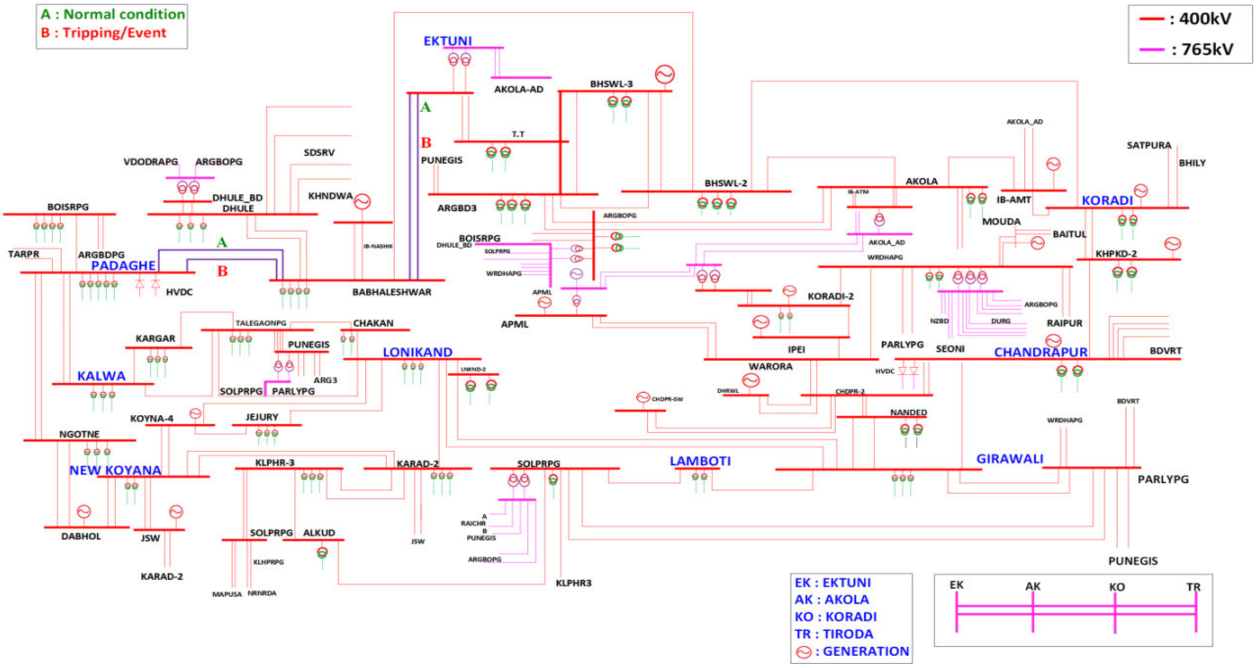


FIGURE 4. 400kV and 765 kV network overview of MSETCL.

center (SLDC) at Kalwa near Mumbai. All these measurements are synchronized with prevalent time. In the next layer, this data is used for numerous applications of WAMS like real-time visualization, state estimation, early warning, congestion management, oscillation monitoring, damping control, voltage stability, adaptive protection, etc.

Figure 4 depicts a 400kV and 765 kV network overview of the MSETCL-SCADA system. The 400kV Babhaleshwar-Ektuni circuit-1 and 400kV Babhaleshwar-Padghe twin circuit lines emanating from the 400kV Babhaleshwar substation experienced tripping cause this disturbance. These lines are indicated in violet color in Figure 4. The line-flows by the SCADA system during normal conditions are indicated by the letter 'A' in green color, whereas that during the tripping condition are shown by the letter 'B' in red color, and the important locations under discussion to describe this disturbance are indicated by blue color. Table 2 shows line power flows during the normal condition and after these two events. The restoration was made according to standard procedures set by the Indian Electricity Grid Code (IEGC) Regulation.

TABLE 2. Line flows preview during normal and line tripping conditions.

Grid Condition with Time	400 kV lines	Power Flows
Normal Condition(A) at 16:24:59 Hrs	Babhaleshwar-Padghe Babhaleshwar-Ektuni	1084 MW 719 MW
After Event-1&2 (B) at 16:30:06 Hrs	Babhaleshwar-Padghe Babhaleshwar-Ektuni	0 MW 0 MW

The PMU data received from eight 400kV substations (i) Padghe, (ii) Kalwa, (iii) Lonikand, (iv) New Koyana, (v) Girawali, (vi) Lamboti, (vii) Chandrapur, and (viii) Koradi is used for analysis. The PMU data compiled from these eight EHV substations, between the periods 16:22:30 to 16:32:30 on 25th May 2017. This practical field data registers two explicit line tripping events during this period. Event-1 gives 400 kV Babhaleshwar-Ektuni transmission line trip in zone-1 of distance relaying at 16:25:36:520 Hrs. Event-2 shows a 400 kV Babhaleshwar- Padghe transmission line trip at 16:30:38:160 Hrs. Both these events are recorded by the PMUs placed at eight substations. The cause of the tripping recorded is phase-b to ground fault. Figure 5 and Figure 6 depict voltage and current variation as revealed by the PMU system during these events [36].

The PMU data consist of eleven power system parameters received at SLDC Kalwa is classified into ambient data and event data. The one set of data contains; (V_a, V_b, and V_c), line currents (I_a, I_b, and I_c), active and reactive power (P, and Q), frequency (f), rate of change of frequency *(df/dt)*, and power angle (δ). The number of PMU measurements from a single PMU per day comes out approximately 23760000. This data is of big volume and it comprises hidden information about the power system dynamics. Therefore, after receiving the data at SLDC, it is required to explore such large data to extract the information hidden in it. This information is required for the improvement of system performance. The purpose of this research article is to identify the synergy among the data obtained from eight substations of the MSETCL grid. This is achieved by applications of data analysis techniques. The data analysis techniques are applied

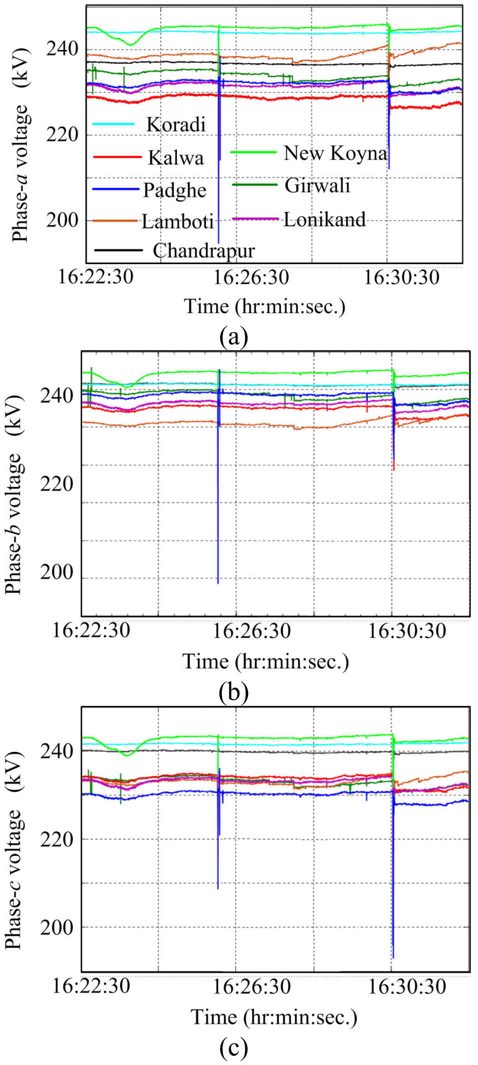


FIGURE 5. Practical results: Voltages indicated by the WAMS during 400kV Babhaleshwar occurrence; (a) phase-*a*, (b) phase-*b*, and (c) phase-*c*.

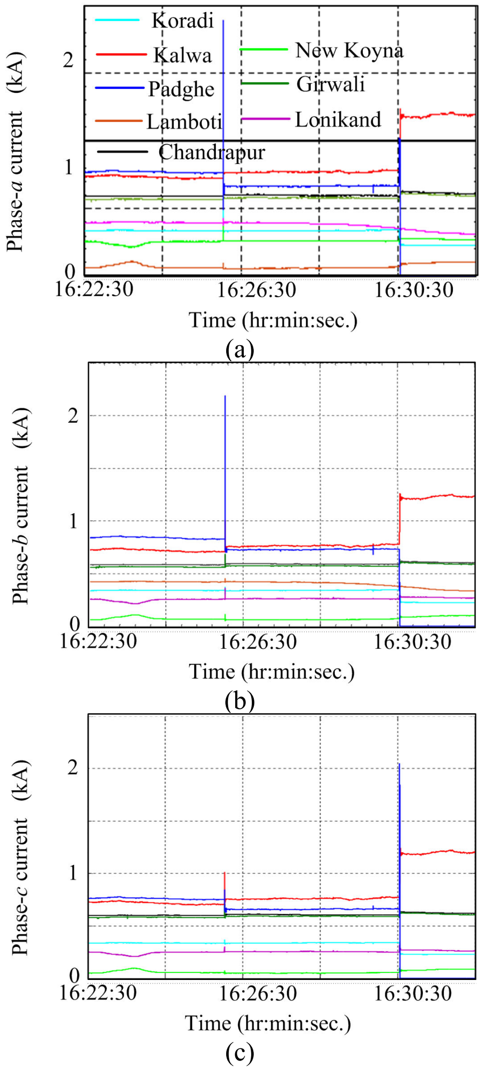


FIGURE 6. Practical results: Currents indicated by the WAMS during 400kV Babhaleshwar occurrence; (a) phase-*a*, (b) phase-*b*, and (c) phase-*c*.

to the voltages and current data recorded by PMUs during the event. This is used to understand the behavior of the grid.

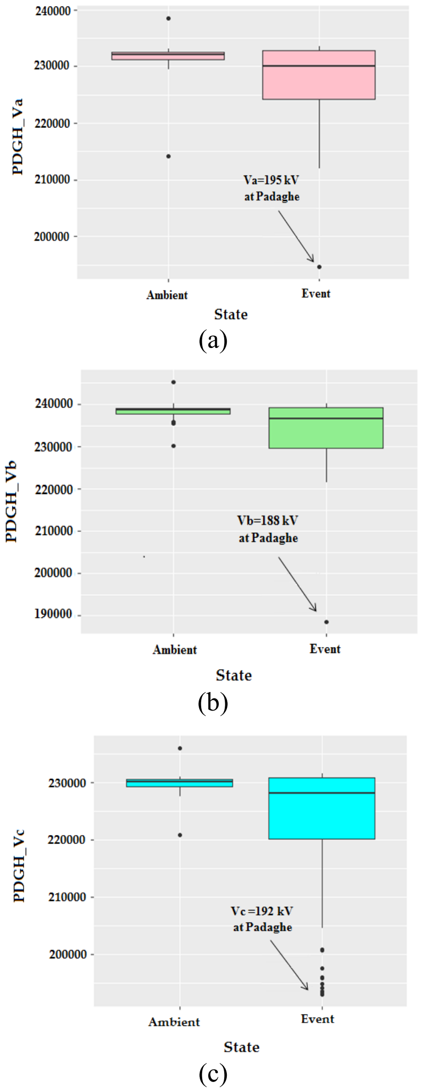
The event in the power system is nothing but a contingency. It is the breakdown or loss of a small part of the power system. It comprises loss of transmission line, failure on generator or transformer, etc. This results in an unplanned outage. In general, such outages may lead to overloads in other branches and/or abrupt changes i.e. rise or drop in system voltage. Contingency analysis is adopted to determine violations. The data analysis techniques extract the important information for contingency analysis which gives information of violations. This information assists the system operator to take remedial action for the removal violations and maintain system stability.

IV. RESULTS AND DISCUSSION

This section elaborates utilization of data analysis techniques described in section II with PMU data obtained from the MSETCL. Table 3 lists the abbreviations of 400kV substations. The PMU data from these locations are used for the analysis of disturbance under consideration. The pre-Event 1, ambient data of PMUs comprises 5288 samples, whereas PMU Event-1 data consists of 38 samples. This data is used for box plots and thereby identification of correlation

TABLE 3. Abbreviations of MSETCL substations.

400kV Substation	Abbreviation
Padaghe	PDGH
Lonikand	LNKD
Kalwa	KLW
Girawali (PARALI)	GRW
Chandrapur	CHNDR
Koradi	KORD
Lamboti(Solapur)	LMBT
NewKoyana	NEWKYN

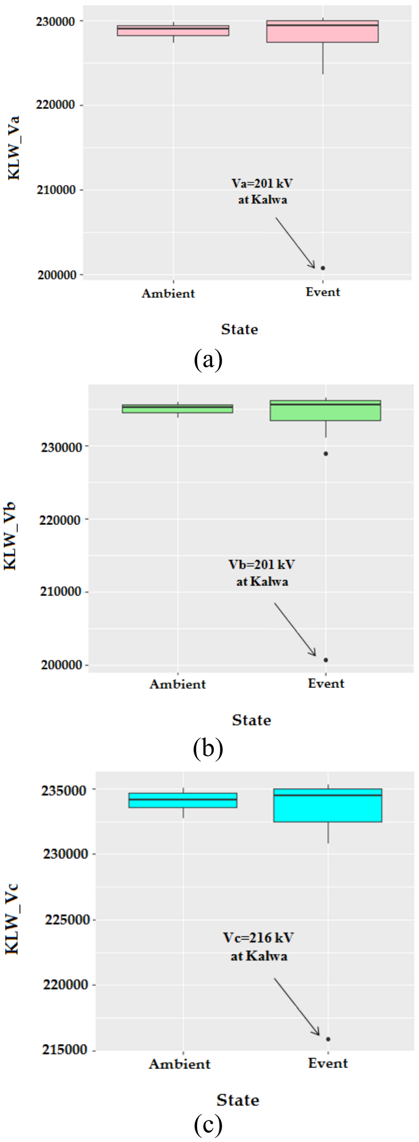


(a)

(b)

(c)

FIGURE 7. Box-plots of Voltages for 400kV Padghe substation; (a) phase-*a*, (b) phase-*b* and (c) phase-*c*.



(a)

(b)

(c)

FIGURE 8. Box-plots of Voltages for 400kV Kalwa substation; (a) phase-*a*, (b) phase-*b*, and (c) phase-*c*.

among various locations in the MSETCL grid. This study can assist system operators in improving decision-making during disturbances/contingencies.

A. APPLICATION OF BOX PLOTS

The utilization of box plots to understand the behavior of the synchrophasor data during ambient conditions and Event-1 is demonstrated. Figures 7-10 illustrate the behavior of voltages, and currents for 400kV Padghe, Kalwa, and Lonikand substations by their respective box plots. Figure 7 indicates the behavior of voltages at 400kV Padghe substation during the ambient and Event-1 period. Figures 8 and 9 depict the behavior of voltages at 400kV Kalwa and Lonikand substations respectively during ambient and Event-1.

Table 4 gives the comparison of voltage outliers as indicated by box plots and actual PMU measurements for Padghe, Kalwa, and Lonikand substations respectively. Figure 10 indicates the behavior of currents I_b and I_c at 400kV Padghe substations during the ambient and Event-1 period. The outlier for I_b and I_c indicates a current of approximately 2.2 kA, pointing it as the fault current. Three-phase voltages found

TABLE 4. Comparative chart of box plot indications and voltages measured by PMUs.

Substation Name	Boxplot indication for voltages based on PMU data during Event-1 in kV (L-G)			Actual voltages as indicated by PMU in kV		
	V_a	V_b	V_c	V_a	V_b	V_c
Padaghe	195	188	192	195	188	192
Kalwa	201	201	216	201	201	216
Lonikand	211	211	222	211	211	222

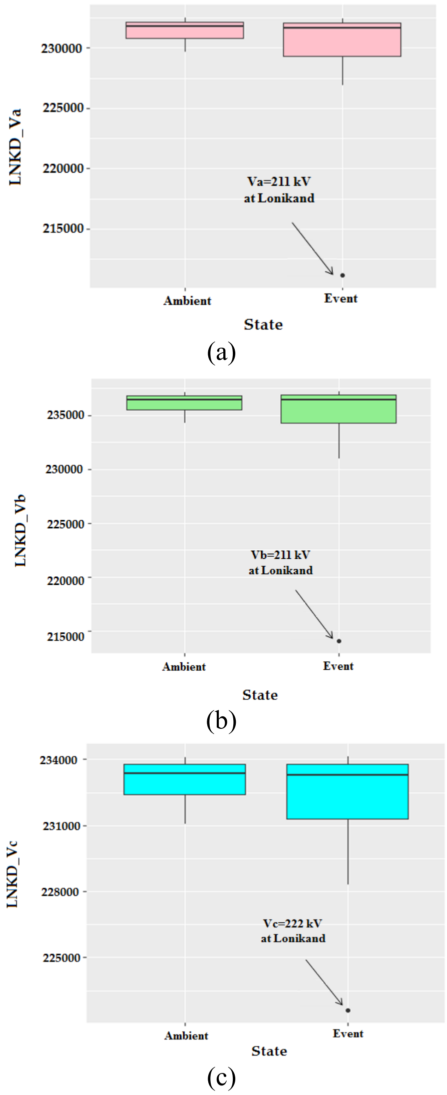


FIGURE 9. Box-plots of Voltages for 400kV Lonikand substation; (a) phase-*a*, (b) phase-*b*, and (c) phase-*c*.

normal (i.e. line to a ground voltage within 230-235kV) as indicated by the first quartile of the respective box plots during ambient conditions is shown in Figures 7-9 for Padghe, Kalwa, and Lonikand substations. The box plots during the event of line tripping condition show the first quartile appears bigger one compared to the third quartile. The voltages are varying to 220kV and the outlier of voltage touched to 188kV is shown in Figure 7(b). This indicates a drop in voltages during the event period. In India, voltages are said in normal condition for 400 kV systems, if they are in the range of 380-420kV (L-L). The voltage variation observed momentarily during the above disturbance by box plots indicates a violation of the normal condition. Similar changes in the rest of the parameters are also obtained with the box plot analysis. These box plots give valuable information on disturbance/event detection and changing dynamic conditions of the grid from normal to abnormal. Table 5 and 6 summarize

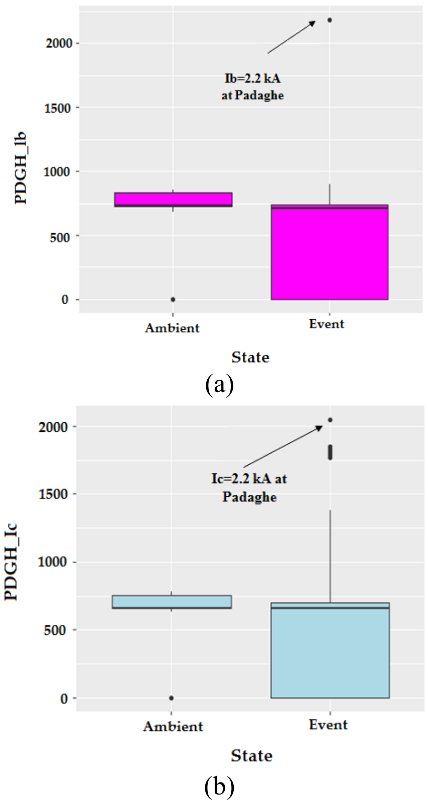


FIGURE 10. Box-plots of Currents for 400kV Padghe substation; (a) phase-*b*, and (b) phase-*c*.

box plot data analytics for Kalwa, Lonikand, and Padghe substations for the parameters V_b and I_b during Event-1 and post Event-1 (ambient condition).

B. UTILIZATION OF CORRELATION TECHNIQUE

This sub-section demonstrates how the line tripping event has an impact on changes in the dynamic operating conditions of the grid. One of the statistical analysis techniques known as Pearson's Correlation Coefficient (PCC) is utilized effectively in understanding the response of PMU data received during and after the events from various 400kV substations. It broadly depicts correlation matrices for important system parameters during ambient and Event-1 condition by applying this technique to PMU data for selected voltage parameters. Each coordinate (circle) of the correlation matrix represents the correlation coefficient of two PMU measurements from different locations with an exception of diagonal elements.

Any variable related to a particular location is self-correlated as shown in Figures 11-13. The circle color indicates the closeness of correlation to 1 or −1. The coordinate sign gives information about whether PMU measurements for the given location pair are positively correlated or negatively correlated. Thus any circle representing brown/red shade is termed as de-correlated. It is observed that buses with high proximity to disturbance location or having connectivity

TABLE 5. Summary of block plots analysis for Phase-*b* voltage.

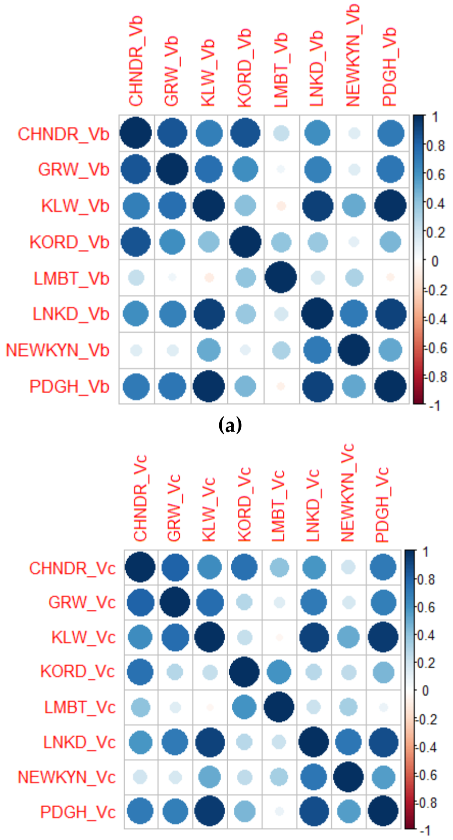
Substation	During Event-1, V_b in kV	Post Event-1, V_b in kV
Kalwa	Minimum : 200.778	Minimum : 233.846
	1st Quartile : 235.458	1stQuartile : 234.524
	Median : 236.182	Median : 235.245
	Mean : 235.953	Mean : 235.100
	3rd Quartile : 236.623	3rd Quartile : 235.639
	Maximum : 237.577	Maximum : 235.966
Lonikand	Minimum : 214.059	Minimum : 234.316
	1st Quartile : 235.458	1st Quartile : 235.541
	Median : 236.182	Median : 236.457
	Mean : 235.953	Mean : 236.175
	3rd Quartile : 236.623	3rd Quartile : 236.804
	Maximum : 237.577	Maximum : 237.171
Padghe	Minimum : 188.594	Minimum : 237.329
	1st Quartile : 237.751	1stQuartile : 238.013
	Median : 238.599	Median : 238.701
	Mean : 238.171	Mean : 238.583
	3rd Quartile : 238.993	3rd Quartile : 239.135
	Maximum : 245.298	Maximum : 239.449

TABLE 6. Summary of block plots analysis for Phase-*b* current.

Substation	During Event-1, I_b in Amp	PostEvent-1, I_b in Amp
Kalwa	Minimum : 700.30	Minimum : 700.30
	1st Quartile : 729.00	1st Quartile : 710.60
	Median : 761.50	Median :722.00
	Mean : 839.50	Mean : 721.50
	3rd Quartile : 773.70	3rd Quartile : 730.60
	Maximum : 1262.40	Maximum : 744.20
Lonikand	Minimum : 214.059	Minimum : 214.90
	1st Quartile : 257.40	1st Quartile : 238.40
	Median : 2361.70	Median : 256.30
	Mean : 259.90	Mean : 248.30
	3rd Quartile : 264.30	3rd Quartile : 257.60
	Maximum : 360.1	Maximum : 260.40
Padghe	Minimum : 0.0003	Minimum : 825.30
	1st Quartile : 725.6603	1st Quartile : 831.30
	Median : 731.1410	Median : 843.80
	Mean : 627.1874	Mean : 841.30
	3rd Quartile : 833.2473	3rd Quartile : 848.20
	Maximum : 2168.1128	Maximum : 858.60

to one another indicate a strong correlation within a given dataset.

The PCC and degree of correlation indication explained in Section II assist in establishing the correlation of data. The data when applied to the correlation matrix of Figure11 gives an idea about the correlation between voltages V_b and V_c as indicated by PMUs from various locations during the ambient and Event-1 period. Figure 11 (a)-(b) shows a graphical display of correlation matrix for V_b and V_c during ambient indicating the high correlation of PMU data, similarly



(a)

(b)

FIGURE 11. Correlogram for Voltages of 400kV substations during ambient condition; (a) phase-*b*, and (b) phase-*c*.

Figure 12 shows a graphical display of correlation matrix for V_b and V_c during Event-1 based on PMU data from substations listed in Table 3. The variable with the substation abbreviation indicates related system parameters based on PMU measurement of that substation using which correlation matrix is formed. For example, PDGH_Vb denotes the phase-*b* voltage of 400kV Padghe substation. The parameters about phase-*b* and phase-*c* have shown maximum changes during this disturbance. Therefore, these parameters are considered for analysis. This graphical display of a correlation matrix is termed a Correlogram. The color code scale given with the matrix illustrates the degree of correlation of PMU data for different 400kV substations from very high correlation to low correlation for the variable under consideration. Figures 11-13 indicate that correlation coefficients are proportional to the color intensity and size of the circle.

Figure 13 shows the correlation coefficient for voltages V_b and V_c during Event-1. The right side of the Correlogram shows the correlation coefficient and the respective colors. The PMU is installed on 400kV Padghe-Babhaleshwar circuit-1. This line connects to the 400kV Babhaleshwar substation where line tripping and related disturbance are noticed. Table 7 shows the correlation for voltages based on PMU data for Event-1.In this case, when the 400kV Padghe-Babhaleshwar line tripped, 400kV Padghe-Kalwa

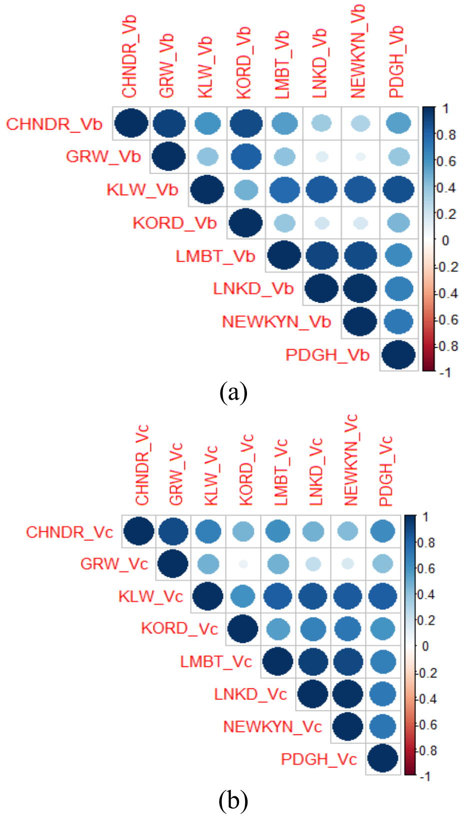
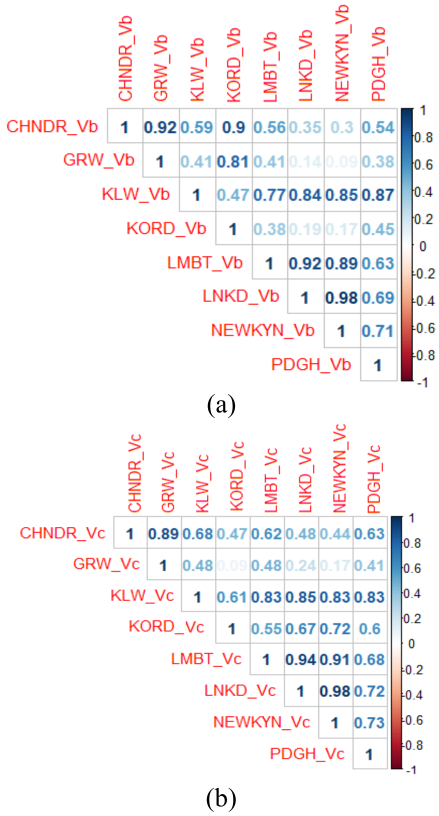


FIGURE 12. Correlogram- graphical display for Voltages of 400kV substations during Event-1; (a) phase-*b*, and (b) phase-*c*.



	CHNDR_Vb	GRW_Vb	KLW_Vb	KORD_Vb	LMBT_Vb	LNKD_Vb	NEWKYN_Vb	PDGH_Vb
CHNDR_Vb	1	0.92	0.59	0.9	0.56	0.35	0.3	0.54
GRW_Vb		1	0.41	0.81	0.41	0.14	0.09	0.38
KLW_Vb			1	0.47	0.77	0.84	0.85	0.87
KORD_Vb				1	0.38	0.19	0.17	0.45
LMBT_Vb					1	0.92	0.89	0.63
LNKD_Vb						1	0.98	0.69
NEWKYN_Vb							1	0.71
PDGH_Vb								1

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

(a)

	CHNDR_Vc	GRW_Vc	KLW_Vc	KORD_Vc	LMBT_Vc	LNKD_Vc	NEWKYN_Vc	PDGH_Vc
CHNDR_Vc	1	0.89	0.68	0.47	0.62	0.48	0.44	0.63
GRW_Vc		1	0.48	0.09	0.48	0.24	0.17	0.41
KLW_Vc			1	0.61	0.83	0.85	0.83	0.83
KORD_Vc				1	0.55	0.67	0.72	0.6
LMBT_Vc					1	0.94	0.91	0.68
LNKD_Vc						1	0.98	0.72
NEWKYN_Vc							1	0.73
PDGH_Vc								1

(b)

FIGURE 13. Correlogram-numeric display showing correlation coefficients during Event-1; (a) phase-*b*, and (b) phase-*c*.

TABLE 7. Correlation indication for voltages based on PMU data for event 1.

Substation Name	Boxplot indication for voltages based on PMU data during Event-1 in kV (L-G)		Actual voltages as indicated by PMU in kV	
	V_b	V_c	V_b	V_c
Lonikand-NewKoyna	0.98	0.98	Very High	Very High
Lamboti-Lonikand	0.92	0.94	Very High	Very High
Padaghe-Kalwa	0.87	0.83	High	High
Padaghe-NewKoyana	0.71	0.73	High	High
Padaghe-Lonikand	0.69	0.72	Moderate	High

line flows might be got affected seriously if other sources were absent. The inference can be seen from the Correlogram indicating the correlation coefficients. From Table 7 it is clear that the highest correlation of 0.98 is observed between Lonikand and New Koyana pair. The correlation coefficient for the Padghe-Kalwa pair for voltages V_b and V_c is 0.87 and 0.83 respectively indicating a high Correlation. The Correlogram of other power system parameters can also be obtained in a similar way to understand their behavior during various power system events.

Event-2 is related to 400kV Babhaleshwar-Ektuni line tripping. The effect of this tripping can be seen on 400kV Babhaleshwar-Walunj circuit-1, 400kV Taptitanda-Walunj double circuit, and also on 400kV Ektuni-Taptitanda double circuit.

Thus, the impact due to disturbance on different lines and substation locations can be understood from the Correlograms. Correlograms accessed using PMU measurements under different contingencies serves as an important decision-making tool for system operator while managing the grid.

C. APPLICATION OF CLUSTERING TECHNIQUES TO PMU DATA AND CLUSTER VALIDATION

This sub-section of the paper elaborates the application of some of the important clustering techniques like K-Means, Hierarchical, and PAM clustering to PMU data accessed during the disturbance. This dataset comprises three-phase voltages, currents, and frequency from eight substations. The data of 15000 samples each for these seven parameters from eight substations comes out to 840000 samples in total.

The K-means, PAM, and Hierarchical clustering techniques are applied to this dataset. This assists in understanding the characteristics of data from different field locations within a grid during a disturbance. The Connectivity, Dunn, and Silhouette indices against the cluster numbers from 2 to 6 are shown in Fig. 14, Fig. 15, and Fig. 16 respectively.

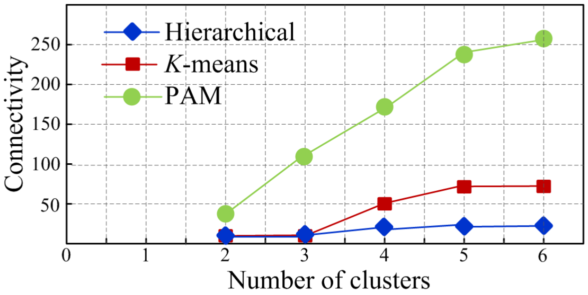


FIGURE 14. Connectivity index versus the number of clusters for different clustering methods.

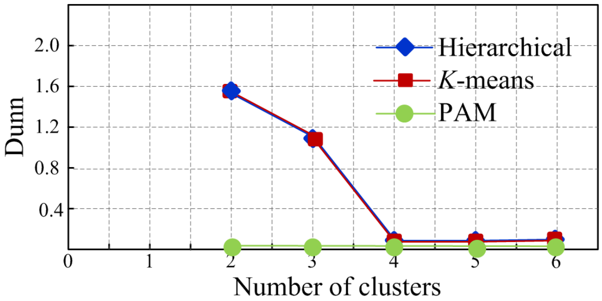


FIGURE 15. Dunn index versus the number of clusters for different clustering methods.

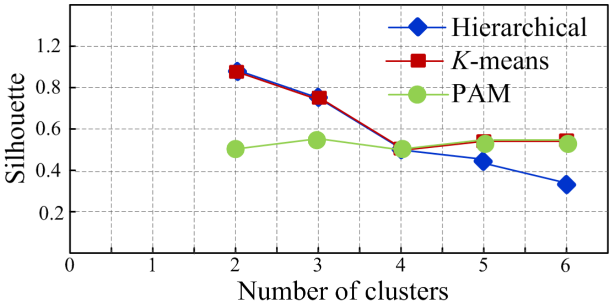


FIGURE 16. Silhouette index versus the number of clusters for different clustering methods.

The smallest Connectivity index asserts a reliable optimal partition with the optimal number of clusters. A minimum value of the Connectivity index gives the optimal number of clusters. Fig. 14 depicts variation in Connectivity index for three different clustering techniques. It is observed that the Connectivity index for Hierarchical clustering is better than the other two methods over the entire clustering range. Dunn index is defined as the ratio of the minimum distance between observations (not in the same cluster) to the maximum intra-cluster distance. Its value is between zero and infinity and should be maximal. A larger value of the Dunn index affirms good clustering. The number of clusters that maximize this index is considered the optimal number of clusters. Thus, from Fig. 15 and Table 8 it is observed that the Dunn Index is more for Hierarchical clustering.

TABLE 8. Summary of block plots analysis for phase-*b* current.

Method	Size of Clusters	Connectivity	Dunn	Silhouette
Hierarchical	2	2.9290	1.5279	0.8817
	3	5.8579	1.0247	0.7597
	4	14.4028	0.0277	0.4990
	5	17.3317	0.0277	0.4416
	6	18.5829	0.0277	0.3997
K-means	2	2.9290	1.5279	0.8817
	3	5.8579	1.0247	0.7597
	4	43.3151	0.0102	0.4990
	5	58.1667	0.0107	0.5534
	6	56.9290	0.0109	0.5542
PAM	2	30.2512	0.0033	0.4945
	3	93.0679	0.0017	0.5625
	4	145.577	0.0023	0.5001
	5	198.6210	0.0024	0.5308
	6	215.4373	0.0024	0.5329

Silhouette width specifies an approach for interpretation and validation of firmness within clusters of data. The method provides a concise graphical visualization of the classification of each object. The Silhouette width runs from -1 to $+1$. The high value indicates that the object is finely matched to its cluster and it is poorly matched to nearby clusters. The clustering configuration is appropriate only if most objects have a high value. In case, if various points have a negative or small value, then the clustering architecture may have either large or cramped clusters. Figure 16 depicts the internal validation result using the Silhouette value for three clustering methods. It illustrates the variation in Silhouette value for three different methods under varying conditions of the number of clusters from 2 to 6. It is seen that the Silhouette value for 2 clusters is around 0.9 (near to unity) in respect of Hierarchical and *K*-means clustering. For PAM clustering it is observed around 0.5.

Therefore, from Table 8, it is seen that for most of the clusters by the Hierarchical clustering method the connectivity index is minimum, Dunn index is maximum and the Silhouette value approaching 1.

Hierarchical clustering hardly brings the correct solution. It embroils lots of aberrant decisions. It does not perform with missing data, but it works poorly with mixed data types. Also, it does not perform well on big data sets. Thus, its main output i.e. Dendrogram, is frequently misinterpreted. However, *K*-means clustering is unsupervised learning and it is for unlabeled data i.e., data without defined groups.

The algorithm performs iteratively to designate every data point to one of the K groups based on the provided features. But, in this study, the Hierarchical clustering gave the same results as K-means clustering. It is possible to find the optimal number of clusters by observing the Dendrogram of a Hierarchical clustering. This task is easier compare to K-means where efforts are made to predict an optimal number in advance. Therefore, the Hierarchical clustering method was selected as more suitable for this PMU data set.

Computing the optimal number of clusters in a given data set is a vital issue in partitioning clustering. E.g. in K-means clustering, the user needs to specify the number of clusters K generated. But still, there is no definitive answer to this query. In some way, the optimal number of clusters is subjective. It depends on the approach used for measuring similarities and the terms used for partitioning. An easy and suitable solution contains the inspection of Dendrogram. It is originated by Hierarchical clustering to check if it proposes an exact number of clusters. Unfortunately, this is also a subjective approach. The package *NbClust* contains 26 indices used to find recommended number of clusters. It is observed that most indices in cluster 2 are shown in Fig.17. The optimal score for these three clustering algorithms based on the internal validation is given in Table 9.

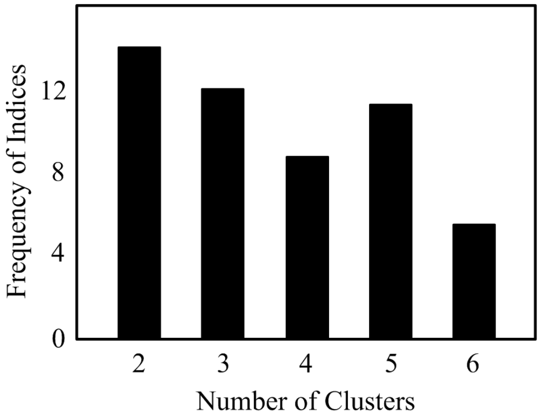


FIGURE 17. Frequency of indices in the number of clusters.

TABLE 9. Optimal score from internal validation.

Method	Type	Cluster	Score
Hierarchical	Connectivity	2	2.9290
Hierarchical	Dunn	2	1.5279
Hierarchical	Silhouette	2	0.8817

Thus such information on the formation of clusters using real-time PMU measurements under dynamically changing conditions of electrical power systems gives a better understanding of the behavior of the grid. This assists system operators in appropriately managing grid operations to keep the system stable at State, Regional, National levels.

V. CONCLUSION

PMUs are installed in India as part of a Central Government Unified Real-Time Dynamic State Measurement (URTDSM) project. The project has been incorporated to implement applications of synchrophasor technology in the Indian power system operation. This paper identifies the synergy among the PMU data obtained from various substations under the western part of the Indian power system i.e. MSETCL grid by applications of data analysis techniques. Data received from PMUs installed at 400 kV substations in the MSETCL grid is used to analyze 400kV line tripping events. It is found that the box plots clearly distinguish between ambient state and disturbance condition. The Pearson's Correlation Coefficient defined the correlation of voltages among different substations. It is observed that the correlations are firmed among Lonikand-Koyna, Lamboti-Lonikand, and Padghe-Kalwa substations. The Correlogram from the PMU data indicates the correlation between the substations during events.

This paper also demonstrates the application of three clustering algorithms viz. K-Means, Hierarchical, and Partitioning Around Medoids (PAM). The internal validation measures such as the Connectivity index, Dunn index, and Silhouette width are used to determine the best clustering algorithms. Thus, the information from box plots, Correlogram, and the formation of clusters under different operating conditions gives a better idea about the grid. The outcome of this analysis can serve as an important decision-making tool to system operators while managing the grid at State, Regional, National levels. The timely and accurate capture of data plays an important role in the efficient management of the power grid. Therefore, future research should focus on advanced techniques, such as artificial intelligence (AI), and machine learning (ML). This will not only exploit the practical PMUs data but also make the decision process less dependent on human interference.

REFERENCES

[1] *ww.cea.nic.in> reports>committee> scm> allindia> agenda_note*. Accessed: Jan. 1, 2021. [Online]. Available: https://cea.nic.in/pspc-region/all-india

[2] F. Aminifar, M. Fotuhi-Firuzabad, A. Safdarian, A. Davoudi, and M. Shahidehpour, "Synchrophasor measurement technology in power systems: Panorama and state-of-the-art," *IEEE Access*, vol. 2, pp. 1607–1628, 2014, doi: 10.1109/ACCESS.2015.2389659.

[3] P. H. Gadde, M. Biswal, S. Brahma, and H. Cao, "Efficient compression of PMU data in WAMS," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2406–2413, Sep. 2016, doi: 10.1109/TSG.2016.2536718.

[4] C. Liu, G. Cai, D. Yang, and Z. Sun, "Extraction and analysis of inter-area oscillation using improved multi-signal matrix pencil algorithm based on data reduction in power system," *Int. J. Emerg. Electr. Power Syst.*, vol. 17, no. 4, pp. 435–450, Aug. 2016, doi: 10.1515/ijeeps-2015-0160.

[5] M. Wu and L. Xie, "Online detection of low-quality synchrophasor measurements: A data-driven approach," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2817–2827, Jul. 2017, doi: 10.1109/TPWRS.2016.2633462.

[6] S. S. Negi, N. Kishor, K. Uhlen, and R. Negi, "Event detection and its signal characterization in PMU data stream," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3108–3118, Dec. 2017, doi: 10.1109/TII.2017.2731366.

[7] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid—A review," *Renew. Sustain. Energy Rev.*, vol. 79, pp. 1099–1107, Sep. 2017, doi: 10.1016/j.rser.2017.05.134.

[8] A. G. Phadke and T. Bi, ‘‘Phasor measurement units, WAMS, and their applications in protection and control of power systems,’’ *J. Mod. Power Syst. Clean Energy*, vol. 6, no. 4, pp. 619–629, Jul. 2018, doi: 10.1007/s40565-018-0423-3.

[9] M. Klari, I. Kuzle, and N. Holjevac, ‘‘Wind power monitoring and control based on synchrophasor measurement data mining,’’ *Energies*, vol. 11, no. 12, p. 3525, Dec. 2018, doi: 10.3390/en11123525.

[10] H. Gharavi and B. Hu, ‘‘Space-time approach for disturbance detection and classification,’’ *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5132–5140, Sep. 2018, doi: 10.1109/TSG.2017.2680742.

[11] Z. Lin, F. Wen, Y. Ding, Y. Xue, S. Liu, Y. Zhao, and S. Yi, ‘‘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, doi: 10.1109/TPWRS.2018.2820066.

[12] N. T. Le and W. Benjapolakul, ‘‘A data imputation model in phasor measurement units based on bagged averaging of multiple linear regression,’’ *IEEE Access*, vol. 6, pp. 39324–39333, 2018, doi: 10.1109/ACCESS.2018.2856768.

[13] D.-I. Kim, A. White, and Y.-J. Shin, ‘‘PMU-based event localization technique for wide-area power system,’’ *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 5875–5883, Nov. 2018, doi: 10.1109/TPWRS.2018.2824851.

[14] H. Akhavan-Hejazi and H. Mohsenian-Rad, ‘‘Power systems big data analytics: An assessment of paradigm shift barriers and prospects,’’ *Energy Rep.*, vol. 4, pp. 91–100, Nov. 2018, doi: 10.1016/j.egyr.2017.11.002.

[15] Y. Zhang, T. Huang, and E. F. Bompard, ‘‘Big data analytics in smart grids: A review,’’ *Energy Inf.*, vol. 1, no. 1, pp. 1–24, Dec. 2018, doi: 10.1186/s42162-018-0007-5.

[16] G. Lee, D.-I. Kim, S. Kim, and Y.-J. Shin, ‘‘Multiscale PMU data compression via density-based WAMS clustering analysis,’’ *Energies*, vol. 12, no. 4, p. 617, Feb. 2019, doi: 10.3390/en12040617.

[17] A. Xue, S. Leng, Y. Li, F. Xu, K. E. Martin, and J. Xu, ‘‘A novel method for screening the PMU phase angle difference data based on hyperplane clustering,’’ *IEEE Access*, vol. 7, pp. 97177–97186, 2019, doi: 10.1109/ACCESS.2019.2930094.

[18] T. Johnson and T. Moger, ‘‘A critical review of methods for optimal placement of phasor measurement units,’’ *Int. Trans. Electr. Energ Syst.*, vol. 2020, May 2020, Art. no. e12698, doi: 10.1002/2050-7038.12698.

[19] J. Chen, X. Li, M. A. Mohamed, and T. Jin, ‘‘An adaptive matrix pencil algorithm based-wavelet soft-threshold denoising for analysis of low frequency oscillation in power systems,’’ *IEEE Access*, vol. 8, pp. 7244–7255, 2020, doi: 10.1109/ACCESS.2020.2963953.

[20] T. Ahmad and N. Senroy, ‘‘Statistical characterization of PMU error for robust WAMS based analytics,’’ *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 920–928, Mar. 2020, doi: 10.1109/TPWRS.2019.2939098.

[21] M. Ballal, A. Kulkarni, and H. Suryawanshi, ‘‘Data compression techniques for phasor measurement unit (PMU) applications in smart transmission grid,’’ *Int. J. Emerg. Electr. Power Syst.*, vol. 21, no. 3, p. 266, Jul. 2020, doi: 10.1515/ijeeps-2019-0266.

[22] N. J. Carter, N. C. Schwertman, and T. L. Kiser, ‘‘A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry,’’ *Stat. Methodol.*, vol. 6, no. 6, pp. 604–621, Nov. 2009, doi: 10.1016/j.stamet.2009.07.001.

[23] *Pearson Correlation Coefficient*. Accessed: Dec. 15, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[24] *Correlation coefficient*. Accessed: Dec. 15, 2020. [Online]. Available: https://en.wikipedia.org/wiki/ Correlation_coefficient#cite_note-3

[25] J. R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Sausalito, CA, USA: University Science Books. 1997. Accessed: Dec. 15, 2020. [Online]. Available: http://hep.ucsb.edu/courses/ph128_18f/Taylor.pdf

[26] *Correlation Matrix: A Quick Start Guide to Analyze, Format and Visualize a Correlation Matrix Using R Software*. Accessed: Dec. 16, 2020. [Online]. Available: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-acorrelation-matrix-using-r-software

[27] *Data Mining Techniques*. Accessed: Dec. 18, 2020. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

[28] *K-Means Clustering: A gentle overview*. Accessed: Dec. 17, 2020. [Online]. Available: https://rpubs.com/riazakhan94/kmeans

[29] *Understanding the Concept of Hierarchical Clustering Technique*. Accessed: Dec. 18, 2020. [Online]. Available: https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

[30] D. Müllner, ‘‘Fastcluster: Fast hierarchical, agglomerative clustering routines forRandPython,’’ *J. Stat. Softw.*, vol. 53, no. 9, pp. 1–18, 2013, doi: 10.18637/jss.v053.i09.

[31] P. Yang and Q. Zhu, ‘‘Finding key attribute subset in dataset for outlier detection,’’ *Knowl.-Based Syst.*, vol. 24, no. 2, pp. 269–274, Mar. 2011, doi: 10.1016/j.knosys.2010.09.003.

[32] G. Brock, V. Pihur, S. Datta, and D. Datta, ‘‘Clvalid: An R package for cluster validation,’’ *J. Stat. Softw.*, vol. 25, no. 4, pp. 1–22, 2018, doi: 10.18637/jss.v025.i04.

[33] J. C. Bezdek and N. R. Pal, ‘‘Cluster validation with generalized Dunn's indices,’’ in *Proc. 2nd New Zealand Int. Two-Stream Conf. Artif. Neural Netw. Expert Syst.*, Dunedin, New Zealand, 1995, pp. 190–193, doi: 10.1109/ANNES.1995.499469.

[34] P. J. Rousseeuw, ‘‘Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,’’ *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[35] J. Handl, J. Knowles, and D. B. Kell, ‘‘Computational cluster validation in post-genomic data analysis,’’ *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, Aug. 2005, doi: 10.1093/bioinformatics/bti517.

[36] *PCM—Western Regional Power Committee Meeting Agenda*. Accessed: Jan. 15, 2021. [Online]. Available: http://www.wrpc.gov.in/pcm/139_PCM_Agenda.pdf

MAKARAND SUDHAKAR BALLAL (Senior Member, IEEE) received the B.E. degree in electrical engineering from the Government College of Engineering, Aurangabad, India, in 1993, the M.Tech. degree in integrated power system from the Visvesvaraya National Institute of Technology, Nagpur, India, in 1997, and the Ph.D. degree by Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India, in 2007. From 1997 to 2012, he was with Maharashtra State Electricity Transmission Company Ltd., Mumbai, India, where he worked on the commissioning, installation, testing, and maintenance of various HV and EHV electrical equipment and accessories. He is currently working as a Professor with the Department of Electrical Engineering, Visvesvaraya National Institute of Technology. He has 15 years of experience in the power sector. His research interests include condition monitoring of electrical machines, power system protections, power quality, and power electronics and drives.

AMIT RAMCHANDRA KULKARNI (Student Member, IEEE) received the master's degree in electrical power systems from the University of Pune, in 2005. He is currently pursuing the Ph.D. degree with the Visvesvaraya National Institute of Technology, Nagpur, India. He is currently working with Maharashtra State Electricity Transmission Company Ltd., as an Additional Executive Engineer. He also worked on project in system studies at IIT Bombay, India. He has worked in various areas, like power system studies and regulatory affairs, protection, testing and transmission project execution, smart grid, and automation project development. His research interests include power system stability, power system planning and studies, power system operation, control and grid management, power system protection, and smart grid, areas like WAMS, FACTS, and renewable energy management systems (REMS). He was instrumental in implementing India's first Wide Area Measurement System (WAMS) Project grid wise in the state of Maharashtra, India. He is a part of various state and regional level study groups in India working in above areas.