DATA ANALYTICS AND WIDE-AREA VISUALIZATION ASSOCIATED WITH POWER SYSTEMS USING PHASOR MEASUREMENTS

A Dissertation

by

IKPONMWOSA IDEHEN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Thomas J. Overbye
Committee Members,	Le Xie
	Timothy A. Davis
	Nicholas Duffield
Head of Department,	Miroslav M. Begovic

May 2019

Major Subject: Electrical Engineering

ABSTRACT

As power system research becomes more data-driven, this study presents a framework for the analysis and visualization of phasor measurement unit (PMU) data obtained from large, interconnected systems. The proposed framework has been implemented in three steps: (a) *large-scale, synthetic PMU data generation*: conducted to generate research-based measurements with the inclusion of features associated with industry-grade PMU data; (b) *error and event detection*: conducted to assess risk levels and data accuracy of phasor measurements, and furthermore search for system events or disturbances; (c) *oscillation mode visualization*: conducted to present wide-area, modal information associated with large-scale power grids.

To address the challenges due to real data confidentiality, the creation of realistic, synthetic PMU measurements is proposed for research use. First, data error propagation models are generated after a study of some of the issues associated with the unique time-synchronization feature of PMUs. An analysis of some of the features of real PMU data is performed to extract some of the statistics associated with data errors. Afterwards, an approach which leverages on existing, large-scale, synthetic networks to model the constantly-changing dynamics often observed in real measurements is used to generate an initial synthetic dataset. Further inclusion of PMU-related data anomalies ensures the production of realistic, synthetic measurements fit for research purposes.

An application of different techniques based on a moving-window approach is suggested for use in the detection of events in real and synthetic PMU measurements. These fast methods rely on smaller time-windows to assess fewer measurement samples for events, classify disturbances into global or local events, and detect unreliable measurement sources. For large-

scale power grids with complex dynamics, a distributed error analysis is proposed for the isolation of local dynamics prior any reliability assessment of PMU-obtained measurements.

Finally, fundamental system dynamics which are inherent in complex, interconnected power systems are made apparent through a wide-area visualization of large-scale, electric grid oscillation modes. The approach ensures a holistic interpretation of modal information given that large amounts of modal data are often generated in these complex systems irrespective of the technique that is used.

DEDICATION

To God and Family

ACKNOWLEDGEMENTS

The journey to attaining a doctoral degree in Electrical Engineering made me realize how much I have been blessed, and how I would not have successfully completed it without people.

Firstly, I would like to appreciate my advisor and committee chair, Prof. Tom Overbye for his support. I owe my knowledge of power systems to him, as without his invaluable directions and input, I would not have reached my professional goal. It is a lifetime honor knowing he was my advisor.

My gratitude goes to my doctoral committee whose insights and contributions have been very helpful. In spite of their different time constraints, their availability in my different PhD examinations is greatly appreciated. I am also grateful for the knowledge I acquired in the classes which they taught.

I would like to acknowledge the support of my friends, colleagues and staff in the research group, both past and present in the University of Illinois at Urbana-Champaign (UIUC) and Texas A & M University (TAMU). Their support and constructive criticisms were invaluable. My appreciation goes to the UIUC staff of Robin Smith, Joyce Mast and Prosper Panumpabi. In TAMU, my appreciation goes to Komal, Won, Adam, Bin, Tamara, Ceci, Ogbonnaya, Zeyu, Wei, Alex and the entire research group. I remember a fellow colleague, Ti, who unfortunately passed away, and am glad to have known and worked with him. I am grateful to Prof. Jose Silva-Martinez, Melissa Sheldon and Katie Bryan who were a great blessing in helping me transition to Texas A&M University when I had recently just transferred to the department.

This acknowledgement is incomplete without the mention of my family and friends. The parental love, discipline and sacrifice provided to me in my early years has made possible my journey through this life. I am eternally grateful to my late, loving mom, who passed away many

years ago. I am forever grateful to my hard-working dad who toiled, sacrificed all his comfort and supported my decision to go back to graduate school. To Vicky whose love, patience, support and prayers guided me through this journey, *I cannot wait to experience the bright future God has in store for us*. To my siblings, Osas, Egbe and Abebe, and brother-in-law Ndubuisi, *thanks for all your prayers, thoughts and love*. I also extend my appreciations to my cousin, Eseosa and Aunty Stella, who both would not allow me drop the ball at a time when I was in desperation; and Uyi, my cousin who always helped out with errands to get my documents mailed to me when I needed them. To an Illinois church family (Kings Assembly Church) and my friends, Olaolu Aj, Olaolu Davies, Tani Davies, Ladi, Oki, Emeka PE, Emeka Okekeocha, Olaniyi, Tola, Pat, Dinah, Uwa, Eddie and Dimeji, I say a big *thank you*!

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Thomas Overbye, Le Xie, Nicholas Duffield of the Department of Electrical Engineering, and Professor Timothy Davis of the Department of Computer Science. The PowerWorld simulation software used for the different simulation studies and visualization efforts were provided by PowerWorld Incorporation through TAMU licensing. The synthetic power grids used in this work were developed by colleagues in the research group.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Initial graduate study was supported by funding from the Illinois Center for a Smarter Electric Grid (ICSEG), and a 50% UIUC graduate teaching assistantship in the Fall Semester of 2016. Subsequent research funds were provided by a Power Systems Engineering Research Center (PSERC) high impact project T-57 which was titled, 'Life-Cycle Management of Mission-Critical Systems through Certification, Commissioning, In-Service Maintenance, Remote Testing and Risk Assessment. These funds covered research studies partly at UIUC and TAMU. Support from Bonneville Power Administration (BPA) project TIP-353 is also acknowledged.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1 INTRODUCTION

1.1 Motivation

The 21st century, modern power grid is a large and complex interconnected system generating bulk amounts of electricity, which are transmitted and distributed to load centers where they are then harnessed for the health and prosperity of its population. For the grid to effectively carry out its function, it is essential that grid resiliency and reliability, through improved grid monitoring capabilities, be maintained at all times.

A common post-disturbance recommendation following the occurrence of major power grid outages in the United States, and other parts of the world has been that the grid, by use of sensor measurements, be closely monitored and managed [1-6]. This would ensure that operating personnel remained in control of the system in the event of any grid disturbance. For example, operators would be able to: better locate disturbance sources in need of increased damping levels when low-frequency oscillations threaten the operational state of the system, identify coherent areas in the system prior to effecting controlled islanding to minimize wide-spread outage, and better coordinate tripping actions of generators when high power capacity transmission lines are lost.

After the 2003 U.S Northeastern blackout, further recommendations for an improved wide-area situational awareness of the grid finally led to the development and deployment of high-reporting sensor devices—the phasor measurement units (PMUs) and other similar synchrophasor devices equipped to measure power system quantities at rates as high as 120 samples per second in 60 Hz operating systems [7-9]. PMUs provide time-synchronized phasor measurements,

otherwise known as synchrophasors, through the aid of a time reference that is provided by a global positioning system (GPS) signal. This enables the wide-area monitoring and control of a grid by making use of measurements obtained from remote locations.

As of September 2017, it was reported that over 2,500 networked PMU devices had been installed on the North American grid [10]. Consequently, large amounts of data can now be generated, and this presents several opportunities for monitoring personnel to have a higher level of visibility of the system. However, the sudden data deluge also presents the operator with a dilemma of fully uncovering and interpreting the operational state of the system. Nonetheless, following from research efforts like those in [11-17], visual displays have become welcomed tools for presenting power system data in formats that are intuitive, thus reducing the challenges faced by operators in interpreting reported grid data.

Power grids are evolving in scale, and improved methods for presenting system information have been suggested. As one of its overarching objectives, [18] emphasizes the need for intelligent data analytics and improved visualization for presenting power system data in a manner that comprehensively describes the state of the grid. The goal is to support informed and coordinated decisions taken by control center operators to ensure that the grid remains in a good operational state.

1.2 Current Technologies and Challenges

Inaccessibility to real synchrophasor or PMU data causes several data-driven power system research activities to make use of laboratory, simulation data. Several power systems simulation software, such as PowerWorld, Power System Simulator for Engineering (PSS/E), Power Systems Computer-Aided Design (PSCAD), and a host of others have been known to generate

experimental, research-based data for analysis. Because they do not capture the normal variations of real grids, simulation data is often devoid of the normal trending features in real PMU data. Also, they do not take into consideration the system and measurement errors that make these real data unique. The use of error-free, simulation data without the features contained in industry-grade data may result in extreme predictions which do not take into account the true variability features in PMU measurements [19].

Like most measurement devices, a failure in the operational mode of a PMU device introduces data errors in reported measurements. In addition, considering the unique mode of operation (such as time synchronization and time stamping [3, 20]), data measurements reported from this device are now exposed to a new paradigm of time-based errors. High quality of reported data can no longer be guaranteed when errors are significantly present in data measurements.

The ability to report phasor angle measurements makes the PMU a critical device in various monitoring aspects of the grid. For example, the level of grid stress is monitored by checking the phase angle differences between nodes on the system, and which could not be done through the use of conventional devices like the remote terminal units (RTUs) and supervisory control and data acquisition (SCADA) devices [3, 8] However, it is important that PMU devices be accurately synchronized to an external reference, otherwise device time errors, which causes mis-synchronization and angle measurement errors, could cause Engineers to lose sight of the true stress levels on the grid.

To aid the identification of time errors, the IEEE documentation on the standard for synchrophasor data transfer, IEEE C37.118.2 [21], has incorporated time and message quality flag bits in every data frame of the C37.118.2 data transfer protocol used by PMU devices for transmitting measured data. These flag bits provide information about the status of time

synchronization and data quality of data recorded by the device. The authors of [22] mention some of the different means by which a dataset can lose its attribute of logical consistency through data mislabeling, duplication, erroneous time stamps and wrong identifiers. These have the potential of rendering measurement data unintelligible, and making power engineers lose sight of the presence and source of data anomalies. A practical example of an instance of data inconsistency is the real-time issue of mislabeled C37.118 flag bits reported by [23]. Also, [22] notes the possible transformations which occur during the data archival process, such that a chain of data-handling procedures exposes PMU data to instances of possible loss or corruption. Finally, [24] reported on data inconsistencies which could arise because of possible data packing issues during data transmission from a data concentrator to a data archive.

The large amounts of PMU grid data available to control centers improves the ability to track the health of the system. As one of the critical monitoring tasks, oscillation monitoring and control plays an important role in ensuring a safe and secure operation of the grid [25-27]. Current methods used for the visualization of power system data [14, 15, 28-30] often present oscillation information on a bus node or limited area basis. However, as mentioned by [18], it is more critical for engineers to have a comprehensive picture of system states when monitoring the evolution of grid trends and dynamics. Moreover, as the grid becomes more interconnected and larger in scale, it becomes more important for data presentation methods to provide holistic perspectives of the grid to system operators.

Synthetic PMU Data

An existing challenge for studying PMU data is the sourcing of actual industry PMU data due to several confidentiality issues. When available, they are often devoid of system event signatures, such as geomagnetic disturbances (GMDs), power flow oscillations and data

measurement errors. These challenges often prompt the use of artificially-generated data for study and research purposes.

1.3 Organization

The major contribution of this work is the implementation of data analytic methods and visualization of large-scale grids through the use of phasor measurements. It utilizes techniques to carry out data error detection in a large-scale system, formulates methods to present wide-area PMU error information, proposes a similarity-based technique to detect time-based errors in phasor measurements, and implements a wide-area visualization platform for large-scale oscillation results. These contributions have been hinged on the generation of realistic, research-grade synthetic data due to the confidentiality issues associate with the use of real PMU measurements.

A literature review is carried out in chapter 2. It summarizes data quality issues associated with PMU measurements, data error analytic techniques and oscillation monitoring, and the contributions of this work in these areas are also highlighted. In chapter 3, different mechanisms leading to PMU data errors are discussed. The purpose is to generate error propagation models, which are used to generate synthetic data errors for use in subsequent chapters. In chapter 4, a framework for the creation of realistic, synthetic PMU data is proposed. It leverages existing synthetic networks to model power system interactions that result in the variations observed in real data. Chapter 5 discusses a distributed method for data error detection in a large-scale system, and presents a similarity-based technique for assessing time errors in PMU measurements. In chapter 6, a multidimensional scaling technique is used to present the different aspects of PMU data errors, which are then displayed on a visualization dashboard interface. Wide-area visualizations of power system oscillations in large-scale electric grids, with a focus on modal estimation quality, modal

interactions and oscillation source detection are presented in chapter 7. Here, the goal is to provide a wide-area assessment of the system dynamics to an observer. A summary of the achievements of this work and future directions are discussed in chapter 8.

2 PRELIMINARY STUDIES AND CONTRIBUTIONS

An overview of the phasor measurement unit (PMU) is presented. Data quality issues associated with measurements obtained from the device is discussed in the first section. An assessment of PMU data for errors and oscillation disturbances are discussed in subsequent sections. The chapter is concluded with a summary of synthetic networks, which are being used for research purposes.

Overview of phasor measurement unit

A PMU is a time-synchronization device which can be used to measure electrical quantities, such as voltage, current, frequency or rate-of-frequency, on the power grid [7, 20, 31, 32]. One of its technological advantage lies in its ability to capture measurement samples at rates much faster than other grid sensors, such as SCADA devices. In a 60 Hz operating grid, PMUs can report measurements at 10, 12, 15, 20, 30 and 60 samples per second, while there have been instances of a report rate of 120 samples per second [24]. In comparison with SCADA devices, which sample data measurements once in 2 to 4 seconds, the large amounts of data generated by PMU devices enable a higher resolution and visibility of the grid. With the aid of an external time reference, via a timing pulse from a global positioning system (GPS) signal, PMUs are able to provide accurate, time-stamped and time-synchronized measurements of more than one microsecond accuracy. This enables wide-area, time-synchronized measurements for the purpose of estimation and analysis of power system states.

A basic, functional block diagram showing the mechanism of a PMU is shown in Fig. 2.1 [24].

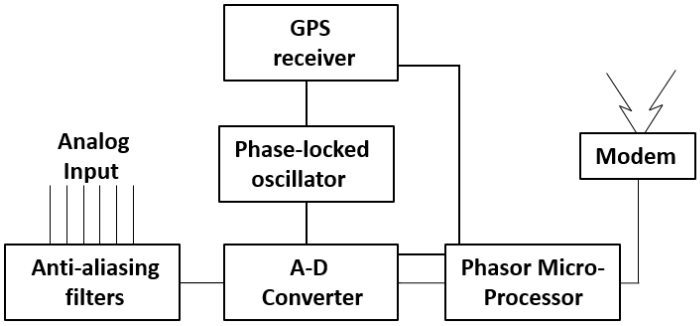


Figure 2.1 Functional block diagram of the elements in a PMU device

External, one pulse per second (PPS) time reference obtained through a GPS receiver is used in a phase-locked loop to create sampling clock pulses, which are used for sampling analog signals (e.g. current and voltage). The quantity phasor, which consists of a magnitude and angle component, is then computed using any of the available phasor estimation techniques.

2.1 Data Errors – A PMU Data Quality Issue

According to [22], data quality encompasses the aspects of accuracy, timeliness and availability. Any activity which reduces the high-quality level of any of the aspects can be defined to be a source of error. The synchrophasor network, comprising of PMU devices, data concentrators, communication links and the phasor applications, is exposed to a variety of errors which could affect any of the reported PMU measurement quantities- voltage (or current) phasor magnitude and angle, frequency and rate of change of frequency (ROCOF). Electrical noise, due to harmonic distortions, wiring of input signals, leakage effect caused by phasor estimation windowing function, was discussed in [20, 33-35]. Time mis-synchronization issues [36-42], caused by clock delays, intermittent reception of global positioning system (GPS) signals, loose cable wiring, spoof attacks, and which lead to phasor angle errors have also been mentioned in the literature. These error types are often attributed to the internal working mechanism of the device;

are manifested in the data, and thus result in low quality data reported by the device. The authors in [43-45] also show that data quality issues result from of low latency, low bandwidth, data drop-offs, wrong data alignment and limited capacities of the communication network. These errors are external, and reflect the limitations of the existing PMU network infrastructure. Finally, as observed from an application level, [46] reported on how an increased deployment of endpoint phasor applications can also reduce PMU data quality.

Several categories of PMU data error sources have been identified in the literature. According to [22], PMU data were classified into groups using three levels of attributes: attributes of single data points, dataset and data stream availability. Attributes of single data point are concerned with the accuracy of the individual, time-stamped measurements, while data set attributes relate to the accuracy and logical consistency of a group of data points or an entire set of PMU data. Data set attributes are related to the condition of the underlying communication network through which PMU data are transmitted. Based on these attribute levels, [47] divided PMU error sources into three categories, and shown in Table 2.1.

Table 2.1. Categorization of PMU error sources

Categories	**Error Sources**
Data point	Accuracy, noise, phase-error, harmonic distortion, estimation algorithms, asynchronous local behaviors (e.g. time-skew), instrument error
Dataset	Status code error, improperly configured PMUs, abnormal or loss of phasor data concentrator (PDC) configuration, frequency calculation discrepancies, mislabeling due to erroneous timestamps, CRC error, invalid timestamp
Data stream	Network limitations - Data loss or drop-outs, network latency; increase in endpoint applications

2.2 Synthetic PMU for Power Grid Studies

A critical challenge in the use of PMU data for research purposes is the confidentiality issues associated with obtaining actual field data. Even when they are made available, they are often devoid of the desired dynamic patterns which need to be studied. These challenges prompt the development of synthetic networks, which are fictitious, but realistic, models of power grids [48-54]. They are statistically similar to real power grids since they are designed with respect to publically available data e.g. size and locations of generators, population density, etc. Thus, they do not contain any critical energy infrastructure information (CEII) and can be freely shared, used in project publications and freely provided to other researchers [55]. A 60-Hz operating, synthetic 2,000-bus network spread out over the geographic region of Texas is shown in Fig. 2.2.

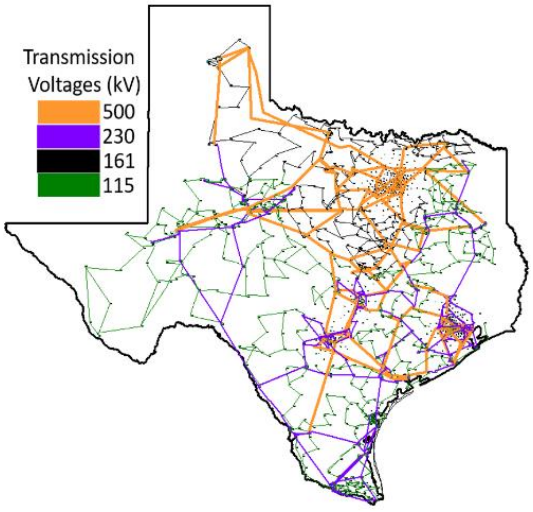


Figure 2.2 A 2,000-bus synthetic network

Containing 1,250 substations, 432 generators, 3,209 transmission lines and different component dynamics set up in the system, the network is designed to simulate the operation of an actual power

grid. More details on this, and a 10,000-bus system used in this work are presented in Appendix A.

The unique feature set of synchrophasor measurements can be attributed to the complex operation of the grid, influence of ancillary components working alongside the phasor measurement device and a host of disturbances occurring on the system [56, 57]. While synthetic networks can be used to generate artificial data for research purposes, these measurements are often devoid of actual PMU data attributes, such as inputs from random load variations and noise. Mostly comprising of only the simulated system dynamics, these measurements are not true representations of real synchrophasor datasets, and could cause researchers to make inaccurate conclusions based on idealistic, experimental results.

The production of synthetic datasets, with similar characteristics as those obtained from a PMU, can be used to circumvent some of the mentioned challenges. This will help address some of the confidentiality issues associated with accessing real data. An ability to generate artificial measurements will also aid research studies, such as the study of grid disturbances, such as GMDs/EMP, which often rely on grid data.

2.3 Wide-Area Detection of PMU Data Errors

Large-scale systems are characterized by interconnections with varying strengths of coupling among sub-networks of nodes (buses and substations). System response to actual grid events is thus non-uniform, and gives rise to varying levels of signal correlations even for the same event.

The authors of [58] explored the low dimensionality of PMU measurements to identify the source of data errors from among a data set. A two-stage state estimation technique was employed

by [59] to detect bad data in PMU measurements, while in [60, 61], neural network techniques were used to train historical measurements and predict an expected maximum deviation beyond which measurements were classified as bad data or outliers. However, the need to constantly update state estimators with the most recent grid model to ensure reliable results, extensive iterative computations and long training times in neural networks oftentimes, hinder the performance and accuracy of these techniques. As systems interconnect to form larger-scale grids, and system topologies become more complex, data-driven techniques which are independent of system model information and without the burden of long computation times are required to provide more robust means of detecting bad data measurements.

2.4 Oscillation Monitoring

As large amounts of synchrophasor data become more widely available from PMU devices, an important task for Engineers is to check for disturbances in the system by carrying out online oscillation analysis on these high resolution phasor measurements. A critical activity in preserving the safe operation of the power grid, the objective of oscillation analysis, monitoring and control is to search for sources of low-frequency oscillation disturbances that may threaten the stability of the system in order to eliminate them [25, 62].

Given an observed time-varying measurement, $y(t)$, the goal of modal analysis is to obtain a re-constructed signal $\hat{y}(t)$ that is a sum of (un)damped sinusoids, and as shown in (1).

$$\hat{y}(t) = \sum_{j=1}^{q} A_j e^{\sigma_j t} \cos(\omega_j t + \phi_j) \tag{1}$$

The j^{th} mode is characterized by the modal parameters: damping factor (σ_j), frequency (ω_j) and mode shape consisting of amplitue (A_j) and phase (ϕ_j). The number of modes is given by q. The error between the original and reconstructed signal, $e(t)$ is computed using (2).

$$e(t) = \sum_{j=1}^{q} \left\| y(t_j) - \hat{y}(t_j) \right\|_2^2 \tag{2}$$

Different modal analysis techniques have been proposed for use in power systems to reveal the underlying low frequency signals intrinsic to power system measurements. The traditional Prony analysis computes the roots of a polynomial to determine the modal frequencies of a signal. These characteristic polynomials are associated with a discrete linear prediction model (LPM) which are used to fit the observed measurements [63, 64]. In the matrix pencil technique, a singular value decomposition is performed on a Hankel matrix, after which the eigenvalues and other modal parameters are obtained [65-67]. One of the advantages of this method is its tolerance to the presence of noise in the observed measurements. A nonlinear least squares optimization method, which encapsulates the linear variables into nonlinear variables, is used by the variable projection method (VPM) to simultaneously estimate all the modal parameters [68]. However, [69] showed that the initial modes provided by the matrix pencil method are usually sufficient. Also, a fast method of dynamic mode decomposition was proposed in [70] for off-line and on-line simultaneous processing of multiple time-series signals.

The above-mentioned modal techniques can estimate modal contents contained in power system disturbance data. However, in large-scale systems, an accurate determination of all system-wide oscillation modes, while minimizing the error function in (2) within reasonable computation times can be a challenging task.

In presenting modal information, and power systems data in general, some authors have made use of visualization tools which include contour maps, dynamic objects, animations and movies to present information about system voltage, frequency, transmission line power flows and other dynamic grid information [14, 15, 28, 71, 72]. Authors in [29] make use of phasor diagrams and

animation to identify coherent group formations and mode shape of a specified inter-area mode at spatially distributed system nodes respectively. Using a combination of 2-D and 3-D graphs, phasor diagram and data table, [30] reports grid modal information. Spatial and temporal variation of mode amplitudes are presented in [70].

Large-scale, interconnected systems generate huge amounts of modal data, and current techniques become inadequate in presenting wide-area system information. Here, there are tendencies for the modal decomposition process to generate several component frequency and damping values, and associated with these components are mode shape and reconstructed signal $\hat{y}(t)$ for each of the observed measurements. In addition, large amounts of processed data are also obtained from the computation of individual transmission line power flows used for the detection of oscillation sources. As mentioned in the problem statement, there is a need for an improved method to present large-scale modal information such that observers can gain better understanding of the behaior of the system.

2.5 Contributions

2.5.1 Generation of realistic synthetic PMU measurements

Here, a proposed framework is comprised of two stages. Firstly, input data made up of annual, seasonal generation, and white-noise, load variations are fed into a power flow solver. Inclusive of other actions, such as automatic controls and disturbances, a power flow solver is used to obtain monitored states of the system. Secondly, measurements obtained from the solver are further modified. Errors and other fictitious measurements, fit enough to retain system dynamics and introduce data randomness, are included to add more realism to the synthetic measurements. This work is addressed in chapter 4.

2.5.2 Event detection and a distributed error analysis of PMU data measurements

An application of windowing-techniques of principal component analysis, a variation trend and modal analysis is used for system event detection in real and synthetic measurements. Fast computation from the use of small time-windows, and ability to discriminate from measurement errors makes these methods attractive. This part of the work is discussed in the last section of chapter 4.

Furthermore, a data-driven technique for analyzing PMU measurements, and applied using a distributed wide-area approach on a large-scale system is proposed. The method is supported by the density-based clustering technique, and is used to assess the level of deviation of the data segments of each phasor voltage magnitude and angle measurement relative to the other phasor measurements. This work is addressed in the second section of chapter 5.

2.5.3 Similarity-based PMU time error detection

Given the unique pattern in which time-based errors manifest in PMU measurements, a post-processing technique for the source identification of PMU time-related errors which is based on the sole use of reported phasor measurements is proposed. It leverages on defined PMU error mechanisms to generate prototype data error patterns, which are then used as training sets in the error analysis and source error identification in synthetic and actual PMU data. This work is addressed in the third section of chapter 5.

2.5.4 Presentation of error information

Given that large amounts of data are generated and transmitted to control centers, a method to present the error information obtained after carrying out data error analysis is shown. Currently, there does not exist a host of research works focused on the subject of PMU data error visualization,

however, the method used here leverages on the use of a multidimensional scaling to view different aspects of data errors. This work is addressed in the chapter 6.

2.5.5 Wide-area visualization of large-scale electric grid oscillation modes

The focus of this work is on the visualization of power system mode oscillations, and is addressed in chapter 7. It does not dwell on the exact methods used in identifying low frequency signal disturbances, however, the results will always be applicable regardless of the chosen method.

3 PMU DATA ERROR MECHANISMS*

To develop the contributions discussed in chapter 2, we need to generate test data which are used in subsequent stages of this work. In this chapter, we discuss the PMU error mechanisms and time error propagation models which are used to generate the synthetic data used in this work. In the next sections, prototype and actual time errors in data measurements are presented.

3.1 PMU Error Mechanisms

The unique time-synchronization aspect of PMU operation requires a prior knowledge of the operation mechanisms associated with data errors before prototype synthetic errors can be generated. Based on the developed models, the appropriate modifications are effected on bus phasor values (magnitude or/and angle). This is in addition to other no-time based errors (e.g., noise, repeated values and dropped data frames).

3.1.1 Time Errors and Error Propagation Models

A loss of synchronism between a reference coordinated universal time (UTC) signal, obtained via the use of a global positioning system (GPS) receiver, and a PMU device internal sampling clock causes time-skew errors [36-42], and have been observed to manifest as phasor angle biases in reported measurements. However, they are observed not to affect the phasor magnitude [37]. Assuming an off-nominal, system frequency of f_i Hz, the phase angle deviation $\Delta\delta_\varepsilon$ due to a time error Δt_ε, is computed as,

$$\Delta\delta_\varepsilon = 360\Delta t_\varepsilon f_i \tag{1}$$

* Part of this section is reprinted with permission from " PMU Time Error Detection Using Second-Order Phase Angle Derivative Measurements" by I. Idehen and T.J. Overbye , Feb. 2019 IEEE Texas Power and Energy Conference (TPEC), ©2019 IEEE, with permission from IEEE

where $f_i = f_o + \Delta\varepsilon$, f_o is the nominal frequency and $\Delta\varepsilon$ is the deviation from f_o. The component of $\Delta\delta_\varepsilon$ due to $\Delta\varepsilon$ is $360\Delta t_\varepsilon \Delta\varepsilon$. Δt_ε is in the order of microseconds, and in normal operating conditions, $\Delta\varepsilon \in (0,0.05)$. Ignoring $\Delta\delta_{\Delta\varepsilon}$, the updated equation becomes

$$\Delta\delta_\varepsilon = 360\Delta t_\varepsilon f_o \tag{2}$$

A corresponding phase angle error due to an observed time difference at each reported sample, however is dependent on the source of timing error. Thus, the instantaneous phase angle error introduced in any reported sample at time, t from the moment of error initiation is given by a generalized error propagation model,

$$\Delta\delta_\varepsilon(t) = 360\Delta t_\varepsilon(t) f_o \tag{3}$$

$\Delta t_\varepsilon(t)$ is the instantaneous, accumulated time drift (or time-skew) at time t.

PMUs report equal time-interval samples of data measurements in cycles, such that the number of data samples reported at any cycle is known as the report rate. The accumulated time drift at any sample point is dependent on the source of error.

1. Clock drift

Here, the internal clock of a PMU is observed to gradually drift away in time due to a delay, which then causes an uneven, accumulating time-interval between samples within a report cycle. A periodic, re-synchronization attempt with a GPS pulse per second (PPS) signal only resets the synchronization status of the first data sample in the next report cycle before the clock drift begins all over again. The error propagation model for this time error behavior is given as,

$$\Delta t_{\varepsilon,i}(t) = (i-1)\Delta t_\varepsilon, i = 1,2 \dots n \tag{4}$$

where n is the reporting rate of the PMU

2. Intermittent GPS Signal

Due to issues, such as loose wiring or incorrect placement of PMU GPS receiver, the device loses connection to the GPS reference signal. A time error, due to a delay, is observed to appear uniformly on subsequent data samples. The time error is observed to appear randomly on consecutive sets of data samples when GPS connectivity is intermittent (e.g. due to loose wiring), and an accumulating time error observed on all samples during a total GPS signal loss (e.g. due to improper placement or malfunctioning of device). The models for the intermittent and total GPS siganal loss (GSL) time error behaviors are states respectively as,

$$\Delta t_{\varepsilon}(t) = \Delta t_{\varepsilon} \tag{5}$$

$$\Delta t_{\varepsilon}(t) = t\Delta t_{\varepsilon} \tag{6}$$

3. Spoofing of GPS receiver signal

Here, an attacker initially acts as an authentic source of correct external reference signal to the PMU, and then attacks the device by gradually leading its signal away from the authentic GPS signal mode. The attack model is given as,

$$\Delta t_{\varepsilon}(t) = \Delta t_{\varepsilon,capture} + tdt \tag{7}$$

$\Delta t_{\varepsilon,capture}$ is a time error at the instance when an attacker completely captures the device receiver, and dt is the rate of time signal divergence induced by the attacker.

3.1.2 Non-Time Related Errors

Similar to data measurements obtained from other grid-installed sensors, PMU data are prone to the effects of unwanted noisy signals, data drops due to communication issues which affect network data streaming ability and repeated measurement values.

Noise in data measurements is modeled as an additive, Gaussian distributed signal, which is parameterized by a zero-mean and finite variance (σ^2). The standard deviation (σ), associated with each of the measurement time points, is obtained from a Signal-Noise Ratio, (SNR, which is in decibels, dB),

$$\sigma = 10^{-SNR/20} \tag{8}$$

Data drop is measured by a drop-out rate attribute, and defines the rate at which packets are lost in a data stream [22]. No data is reported at time points during which packets are lost or delayed, and [21] suggests the use of NaN (not a number) or 0x8000 (-32768)- corresponding to zero values - as filler data, which are not used in actual computation.

3.1.3 Updating Derived Measurements (Frequency and ROCOF)

Depending on the type of synthetic data error that is prototyped, a re-computation of the frequency and ROCOF signals is required. Currently, no specific estimation technique for these quantities has been defined by the IEEE reference documentation [21]. However, since the power system simulation software used for this work mimics high voltage, transmission grid operations, an approach based on [20] is used. It assumes a balanced set of three-phase input signal, and devoid of the iterative computations associated with nonlinear frequency estimation associated with unbalanced inputs.

Let $\omega(t)$, ω_o, $\Delta\omega$ and ω' denote the instantaneous, nominal, deviation values and rate of change of angular frequencies respectively; and ϕ_o, $\phi(t)$ denote the values of the initial and instantaneous phase angles respectively. It follows that:

$$\omega(t) = \omega_o + \Delta\omega + t\omega' \tag{9}$$

$$\phi(t) = \int \omega(t) = \phi_o + t\omega_o + t\Delta\omega + \frac{1}{2}t^2\omega' \tag{10}$$

Neglecting the nominal angular velocity which is uniform for all phase angles,

$$\phi(t) = \phi_o + t\Delta\omega + \frac{1}{2}t^2\omega' \tag{11}$$

(3) is a quadratic expression in *t*, and expressed as:

$$\phi(t) = a + bt + ct^2 \tag{12}$$

where *a*, *b* and *c* correspond to ω_o, $\Delta\omega$ and $^1/_2\,\omega'$ respectively. Solving for *a*, *b* and *c*, using a least-squares methods, the frequency deviation and ROCOF are evaluated as follows:

$$\Delta f = \frac{b}{2\pi}(Hz), \;\; f' = \frac{c}{\pi}(Hz/s) \tag{13}$$

The actual frequency, f_{act} is then computed as:

$$f_{act} = f_o + \Delta f \tag{14}$$

Alternatively, based on the implicit definition of frequency as the rate of change of phasor angle, the derived measurements can be computed in the frequency domain as,

$$f = \frac{s}{1 + sT}\theta \tag{15}$$

$$ROCOF = \frac{s}{1 + sT}f \tag{16}$$

where $T \sim 0.2$ second is a time-delay used to capture a window of data samples.

3.2 PMU Data Prototypes for Time Error

Figs. 3.1 and 3.2 illustrate voltage angle (VA) profiles based on the time propagation models in (4) – (7). Four different PMU time error prototypes were generated using original data from a test bus in the 2,000-bus network after a 30 second simulation. Error injection is initiated at the 5th second, and exists for 20 seconds. The report rate of the PMU is 30 samples per second.

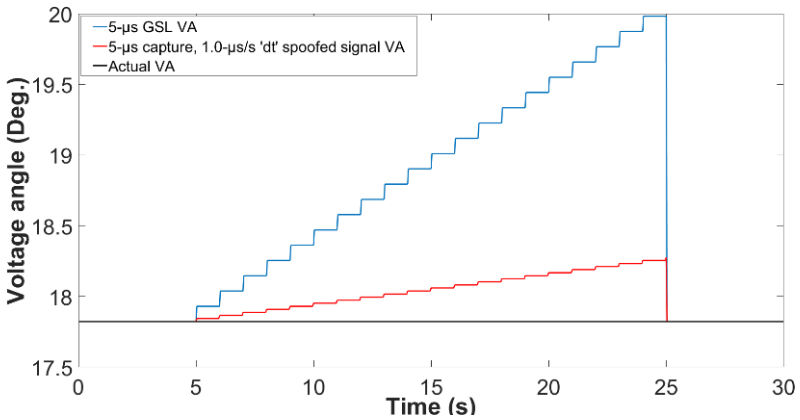


Figure 3.1 Voltage angles for GSL and signal spoof. Reprinted with permission from [73]

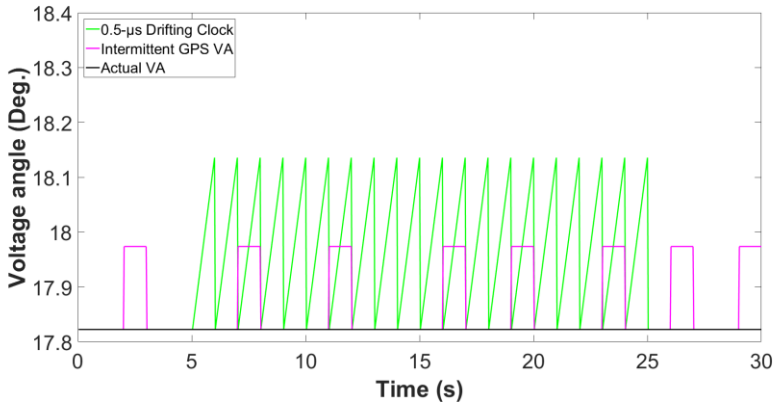


Figure 3.2 Voltage angles for clock drift and intermittent GPS. Reprinted with permission from [73]

In Fig. 3.1, the VA waveforms of the GPS signal loss (GSL) event and a spoofed-GPS time signal are shown. The black horizontal line is the original, steady state VA. The blue-colored GSL event has a pulse per second (PPS) time error (Δt_ε) of 5 μs, and the red-colored spoofed signal event has a time error divergence rate (dt) of 1 μs/s. For each of the events, the phase angle error $\Delta\delta_\varepsilon(t)$ is applied uniformly on all 30 samples in a one-second reporting period prior to the next set of reported samples.

Fig 3.2. illustrates VA profiles due to two different causes of PMU clock time offsets – a constant 0.5 μs time error due to a drifting internal clock, and a 1.0 μs error due to intermittent GPS clock signals received by the PMU device. The green-colored ramp for the cloc drift error is indicative of the accumulatng time error at each sample. In contrast, a uniform time error is observed for all samples in the case of intermittent GPS signal.

PMUs report ROCOF data which can also be used to monitor phasor angle changes. The derived ROCOF data for the voltage angle errors are shown in Fig. 3.3 and 3.4.

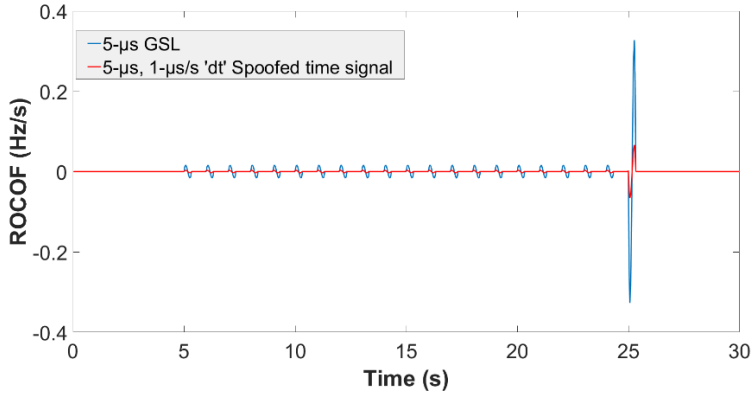


Figure 3.3 ROCOF data reprinted with permission from[73] for prototyped PMU voltage angles in Figure 4.1

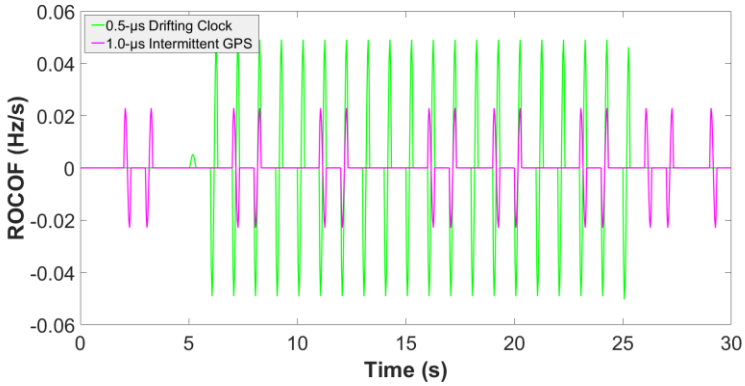


Figure 3.4 ROCOF data reprinted with permission from [73] for prototyped PMU voltage angles in Figure 4.2

Periodic ripples observed in Fig. 3.3 for both GSL and spoofed GPS signal are attributed to the small jumps in voltage angles due to the incremental time errors. However, at the point of error removal, the accumulated voltage angle deviation results in a sudden spike in the ROCOF. We observe that the GSL event generates a significant spike (0.33 Hz/sec) which is due to the large angle deviation as compared to the case of the spoofed time signal.

In Fig. 3.4, the observed uniform ROCOF measurements for a PMU internal clock offset is consistent with the periodic VA ramp-and-reset observed in Fig. 3.2. With an intermittent GPS signal, we observe a pairwise formation of positive and negative edges of ROCOF measurements.

Real Data with Time Error

A time skew error in real current measurements obtained from a 50-Hz operating system, with a report rate of 25 samples per second, is shown in Fig. 3.5, while an artificially generated one for a 60-Hz system whose PMU reports 30 samples per second is in Fig. 3.6.

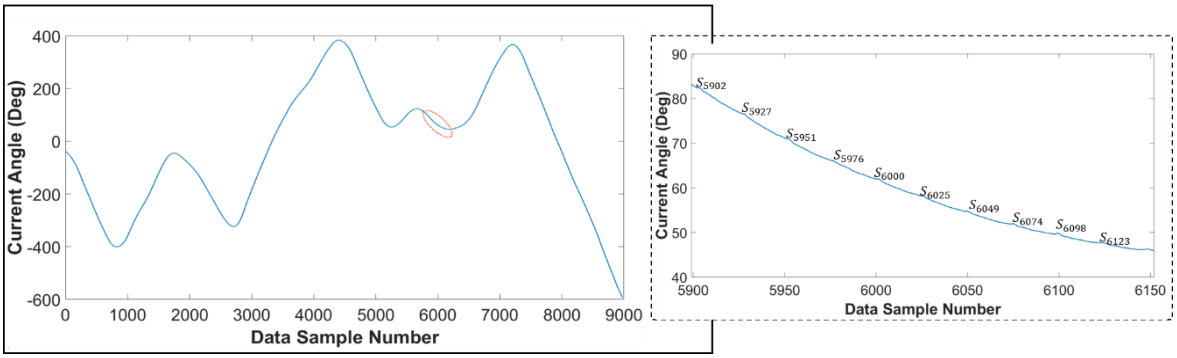


Figure 3.5 Time skew error in real current angle data

Synthetic Data with Time Error

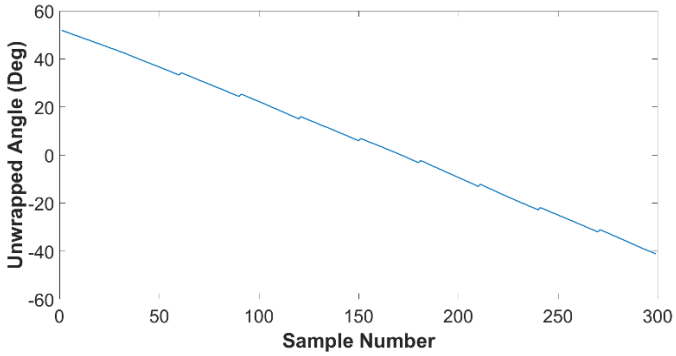


Figure 3.6 Time skew error in synthetic voltage angle data

3.3 Time & Message Quality in IEEE C37.118 PMU Data

As one of the four message types in the C37.118 framework, the data frame packet holds phasor measurements of the sample being reported. In addition, it contains time and message quality information about the generated data. Fig. 3.7 shows Table 5 of the IEEE documentation [21] describing component fields in a data message.

Table 5—Data frame organization

No.	Field	Size (bytes)	Comment
1	SYNC	2	Sync byte followed by frame type and version number.
2	FRAMESIZE	2	Number of bytes in frame, defined in 6.2.
3	IDCODE	2	Stream source ID number, 16-bit integer, defined in 6.2.
4	SOC	4	SOC time stamp, defined in 6.2, for all measurements in frame.
5	FRACSEC	4	Fraction of Second and Time Quality, defined in 6.2, for all measurements in frame.
6	STAT	2	Bit-mapped flags.
7	PHASORS	4 × PHNMR or 8 × PHNMR	Phasor estimates. May be single phase or 3-phase positive, negative, or zero sequence. Four or 8 bytes each depending on the fixed 16-bit or floating-point format used, as indicated by the FORMAT field in the configuration frame. The number of values is determined by the PHNMR field in configuration 1, 2, and 3 frames.
8	FREQ	2 / 4	Frequency (fixed or floating point).
9	DFREQ	2 / 4	ROCOF (fixed or floating point).
10	ANALOG	2 × ANNMR or 4 × ANNMR	Analog data, 2 or 4 bytes per value depending on fixed or floating-point format used, as indicated by the FORMAT field in configuration 1, 2, and 3 frames. The number of values is determined by the ANNMR field in configuration 1, 2, and 3 frames.
11	DIGITAL	2 × DGNMR	Digital data, usually representing 16 digital status points (channels). The number of values is determined by the DGNMR field in configuration 1, 2, and 3 frames.
	Repeat 6–11		Fields 6–11 are repeated for as many PMUs as in NUM_PMU field in configuration frame.
12+	CHK	2	CRC-CCITT

Figure 3.7 IEEE documentation on data frame structure

The SYNC field provides information about time synchronization and frame identification followed by a 2-byte field of the current frame size. Each data stream is identified by an IDCODE field specifying the message source or destination. Time stamp (32-bit unsigned number), fraction of second and time quality information are provided by the Second of Century (SOC) and FRACSEC fields. A 2-byte STAT field makes use of flagged bits to provide quality status - time and measurement quality - of the reported measurement quantity. Depending on the error type being prototyped, bit modifications are carried out in accordance with the IEEE documentation. Fig. 3.8 shows the information conveyed by each of the constituent bit segments in the STAT field.

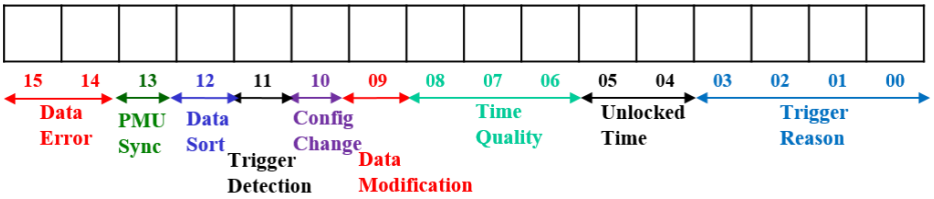


Figure 3.8 Bit segment information in a data frame STAT field

Detailed information about the content and bit status for all 16 bit segments are provided in pages 16 and 17 of [21]. For the purpose of this work, the focus is on bit 13, which provides synchronization status information for every reported data frame. As part of a modification step after generating synthetic PMU data, this bit is altered to reflect the data error being prototyped.

3.4 Summary

In this chapter, some of the common time propagation error models associated with PMU time-based, data errors were presented. Using these models, and showing figures of phasor angle errors, we were able to observe the unique patterns when these errors appear in measurements. In subsequent chapters, we will use these models during the creation of synthetic PMU data.

4 GENERATION OF POWER SYSTEMS SYNTHETIC PMU DATA

In this chapter, a process for the creation of synthetic data for research purposes is presented. Based on publically-available and pre-processed data pertaining to grid generation and load patterns, these artificial data are generated from simulations carried out on a power systems simulation software. The preferred choice of a transient stability (TS) power flow solver over a steady-state power flow analysis was to capture transient effects during selected system contingencies.

4.1 Background

The unique feature set of synchrophasor (or PMU) measurements can be attributed to the complex operation of the grid, influence of ancillary components working alongside the phasor measurement device, and effects of extraneous activities on the system [56, 57, 74]. A consequence of constantly-changing consumer loads (residential, commercial and industrial), control device actions (transformer tap changing, breaker operation, shunt capacitor switching), and several range of disturbances is the consistent variation observed in high-resolution, time-series synchrophasor measurements. In addition, low-accuracy levels of instrument transformers, improperly-connected wires and phasor estimation errors cause deviations in measurements, significant affect noise levels and introduce outlier measurements often observed in actual synchrophasor data [20, 22, 33, 34, 36, 37, 75, 76] . While synthetic networks can be used to generate artificial data for research purposes, these measurements are often devoid of actual PMU data attributes, such as inputs from random load variations and noise. Mostly comprising of only simulated system dynamics, these measurements may not be true representations of real synchrophasor datasets.

4.1.1 Real vs Simulated Data

Fig. 4.1 show two sets of per unit synchrophasor voltage measurements spanning a 20-second duration, and obtained after a generator outage. The red plots in Fig. 4.1 (a) are obtained from real PMU dataset (see Appendix B), and the blue plots in Fig. 4.1 (b) are obtained from a study carried out on a power systems simulator.

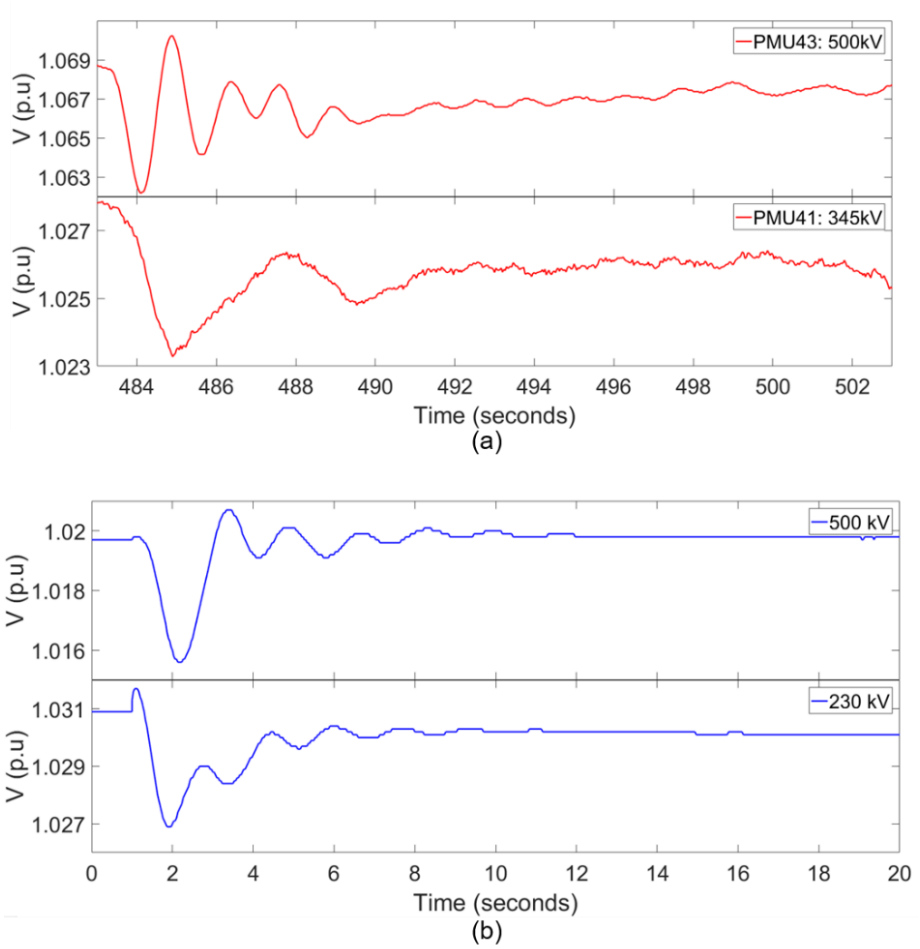


Figure 4.1. 10-sec per unit voltage for 2 PMUs from real and simulation data

The operating frequency for both systems is 60 Hz, and have PMU report rates of 30 samples per second. To ensure a fair comparison of both systems, the resolution on all voltage axis have been set to three decimal places. Regardless of the voltage ratings of the nodes being monitored in the real system, a trending voltage profile is observed for both PMU measurements, and is indicative of the continuous state of operation and dynamic interactions of the grid. In contrast, a predominantly, flat voltage profile is observed in the simulation data, even after an occurrence of a generator event. While there may be few measurement variations, they are mostly attributed to a pre-defined, input time step during which the grid state is evaluated. The measurements obtained from the inactive simulation system is thus a sharp contrast to those obtained from a steady, dynamics-driven, real power system. This deviation of real synchrophasor measurements from error-free simulations becomes more apparent when, in addition to system dynamics, issues such as malfunctioning PMU device components, limited network bandwidth and communication lags occur which then manifest as errors in real data.

Common features of industry-grade, PMU measurements have been identified in the literature, and they include well-damped, low-frequency system modes [25],[27, 77], disturbance events[78], measurement outliers due to system transients and noise [34, 79-81], missing data points [82], and several anomalies attributed to device errors already discussed in the prior chapter.

Fig. 4.2 shows 1-minute, real frequency data obtained from four different PMUs in the same system.

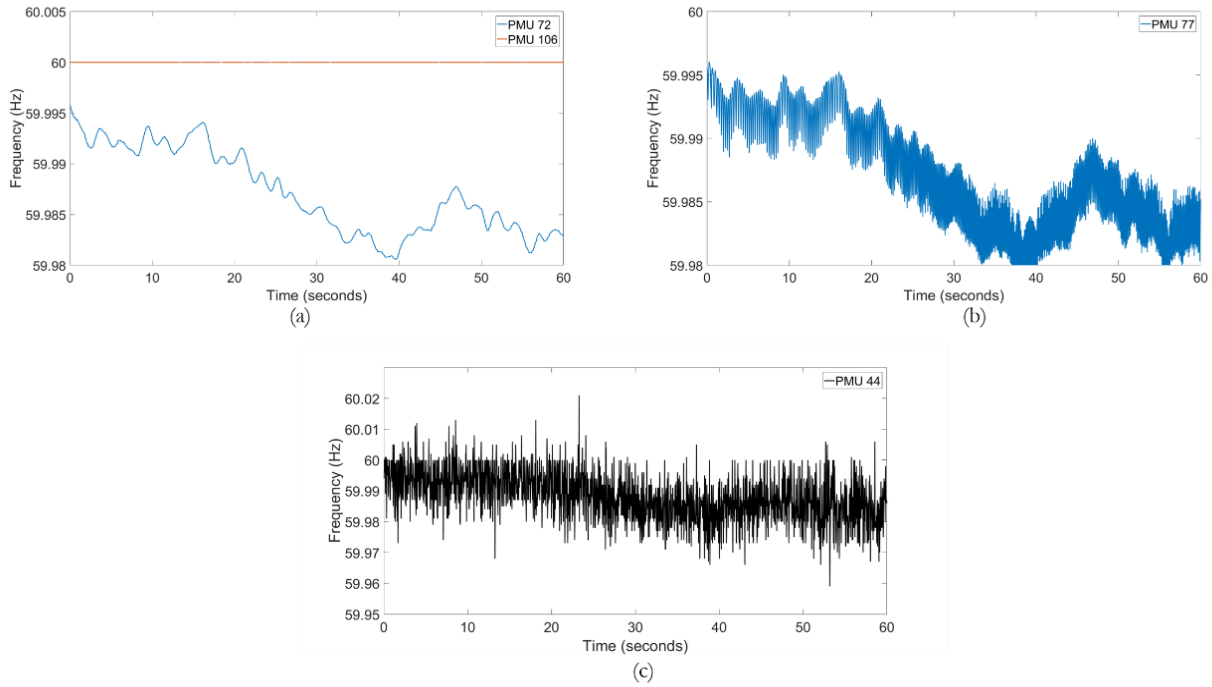


Figure 4.2. 1-min frequency measurements

The observations from the real PMUs are discussed thus: erroneous, constant 60-Hz frequency reported by PMU ID 106 even in the midst of grid disturbances; low frequency oscillations of 0.026-Hz and 0.069Hz observed in PMU 72, exceptional 0.2-Hz component in PMU 77, and the significant noise level in PMU 44. In contrast, simulated, error-free frequency measurements are deficient of these anomalies, and will often exhibit much lesser variations.

Fig. 4.3 shows 1-minute, real voltage angle measurements obtained from five PMUs.

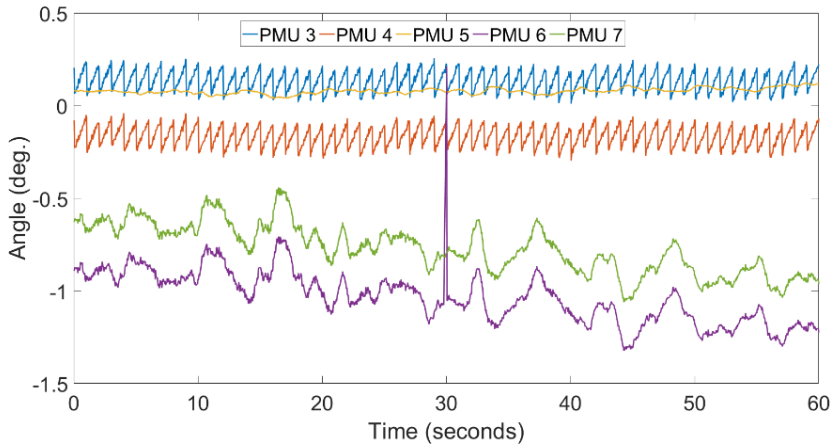


Figure 4.3. 1-min voltage angle measurements

Neglecting the slow, and true trending angle measurement samples of PMU 5, the other PMUs are a reflection of some of the several, anomalous features which can occur in industry-obtained data, and must be handled prior application usage. The observations in Fig. 4.3 are thus: PMUs 3 and 4 exhibit time-skew errors due to clock drift errors (discussed in Section 3.1); low frequency oscillations are observed in PMUs 6 and 7; and an outlier data point in PMU 6. In simulated error-free voltage angles, transitions between measurement samples are, if any, smooth, and lacking of any of the above attributes thus, causing them to differ from true industry data.

Finally, in the event of an actual contingency, the features of real measurements will often be more complex than its simulated counterpart. Fig. 4.4 is a 1-minute, voltage measurement from ten real PMUs during which there is a generator outage.

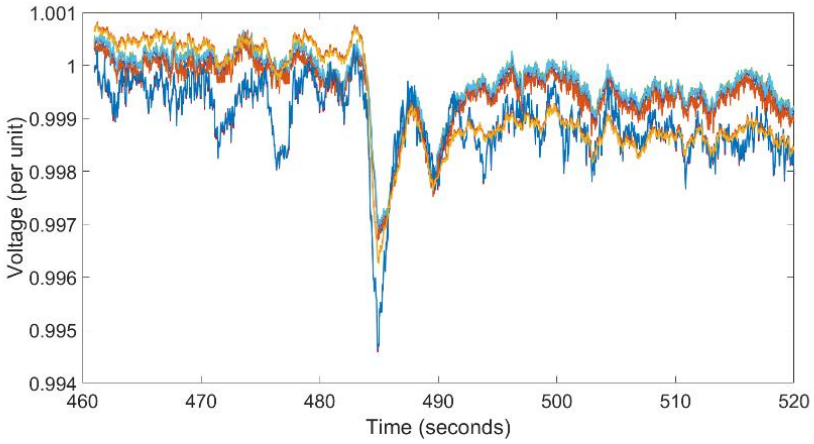


Figure 4.4. 1-min per unit voltage of 10 PMUs during generator outage

Voltage dip, as a result of the generator outage event is uniformly captured by all devices. Additional variations are also observed in the system, and attributed to other system dynamics which could include changing load-generation mix, local shunt-switching and the effects of operator control actions in the system. This complex, spatio-temporal relationship among PMU locations introduces features that may not be captured by simulated measurements.

4.1.2 Variability in PMU Data

Power system data measurements can be classified as non-stationary time series since the operating condition of the grid is known to evolve over time. According to [78], measurement variations of grid voltage is represented in (1).

$$\sigma_{\Delta V_M}{}^2 = \sigma_{\Delta V}{}^2 + \sigma_{\eta}{}^2 \quad (1)$$

$\sigma_{\Delta V}{}^2$ and $\sigma_{\Delta V_M}{}^2$ are voltage signal variances before and after an introduced noise measurement variance, $\sigma_{\eta}{}^2$. While the value of $\sigma_{\eta}{}^2$ can be obtained directly using any of the several filtering techniques in literature, $\sigma_{\Delta V}{}^2$ is often a by-product obtained from its filtered signal. Fig. 4.5 shows

a 5-second, per-unit real voltage profile extracted from a real PMU (see appendix B for full description of real PMU dataset) with much longer duration of measurements, and a report rate of 30 samples per second.

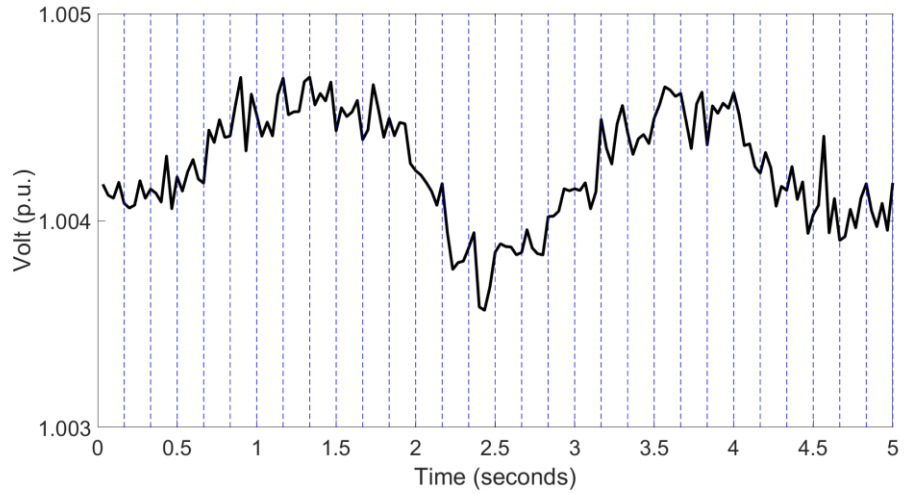


Figure 4.5. 5-sec per unit voltage magnitude

The non-stationarity feature of this time series is observed by the continuous, seemingly-erratic state of the voltage samples. Research shows that signal-noise ratios (SNR) for most power system measurements often lie within a range of 43-47 decibels [34]. However, a computed value of 70 decibels for the first minute of this measurement proved that a high proportion of the signal was relatively noiseless. This is reflected by the three decimal-place representation before signal variations could be observed. Given a high SNR, we can assume the value of σ_η to be relatively small, such that $\sigma_{\Delta V_M}$ in (1) is predominantly composed of the actual voltage variance, $\sigma_{\Delta v}$. Further analysis can be carried out on this signal by observing other trends in its down-sampled forms.

The blue plot of Fig. 4.6 shows the average voltage of every non-overlapping window of five data samples, and a blue plot shows its corresponding window variance.

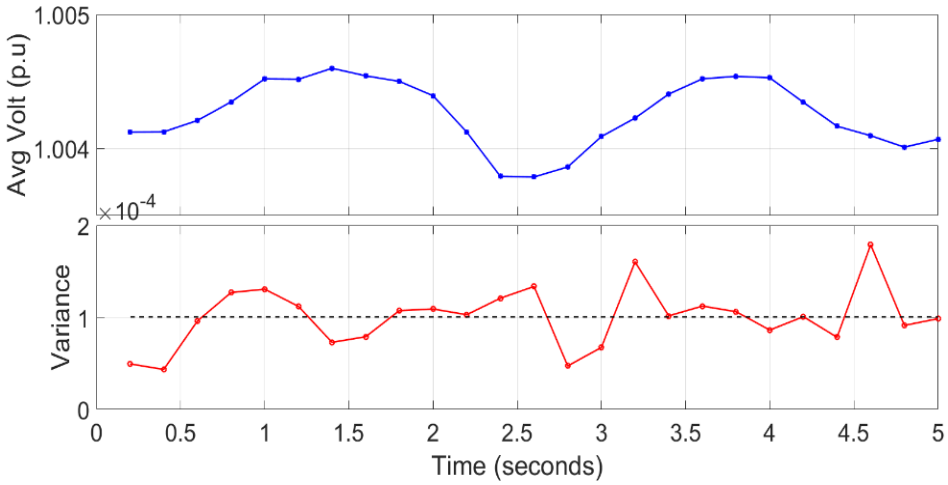


Figure 4.6. Down-sampled: 5-sample window mean and variance

With respect to the per-unit voltage, an average variance of 10^{-4} is observed for all windows when only a steady change is observed in the window average voltage, furthermore indicating the extent of total voltage variability $\sigma_{\Delta v}$ in this clean measurement segment. Given a previous assumption of negligible value of σ_{η}, we can regard only $\sigma_{\Delta v}$ as the only component $\sigma_{\Delta V_M}$, thus providing a standard for true system voltage variability. However, in segments of the measurements where more random variations are associated with lower computed SNR values, the assumption of small σ_{η} no longer holds true.

Based on the above argument, the average variability and SNR value for the first 5-second duration voltage measurement of all the real system PMUs are computed, and illustrated in Fig. 4.7 (a) and (b) respectively.

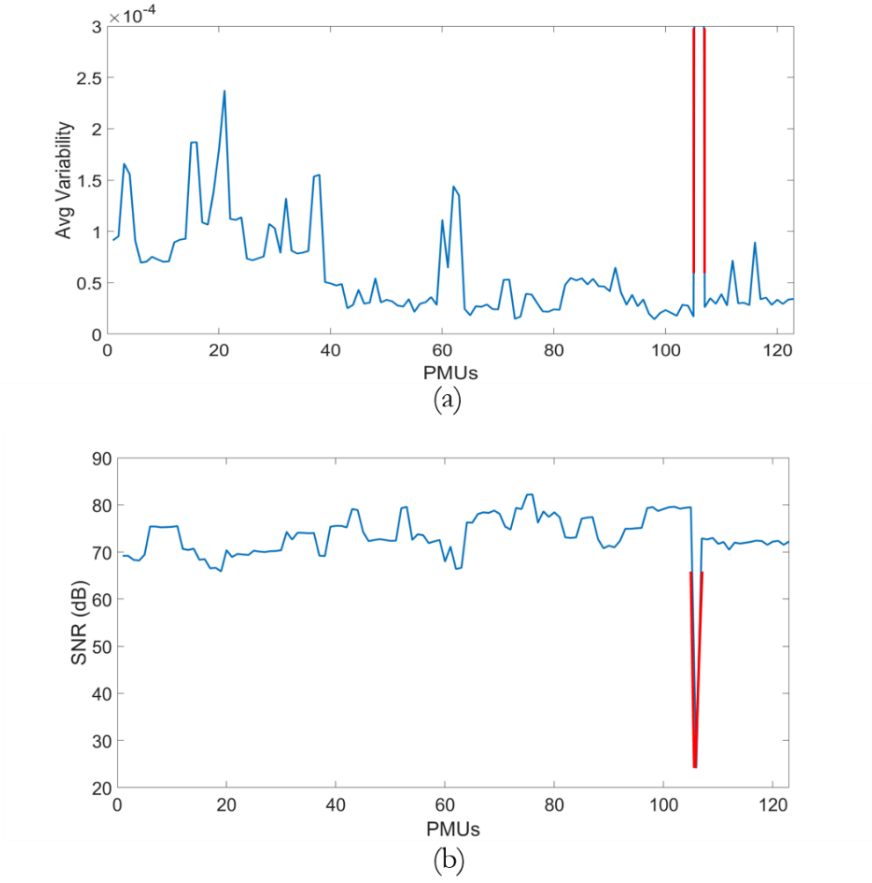


Figure 4.7. Average variability and SNR of all 123 real voltage measurements

The red spikes in both figures are due to the abnormal, average variability level noticed at one of the sources (with ID: 106) which was observed to have a value of 2.23×10^{-3}. Neglecting this erroneous PMU, an average measurement variability of about 0.3×10^{-4} $(i.e., 3 \times 10^{-5})$ is observed across the system, which is also confirmed by the almost uniform SNR values of the PMUs.

Measurement Noise

The unwanted disturbance of measurement noise injects a measure of variability into synchrophasor measurements. Noise in power systems is assumed to be Gaussian, and thus modeled as a normal distribution with a zero-mean and a standard deviation. Fig. 4.8 shows the

first 1-minute frequency obtained from a real PMU (with ID: 44), and from which noise has been extracted.

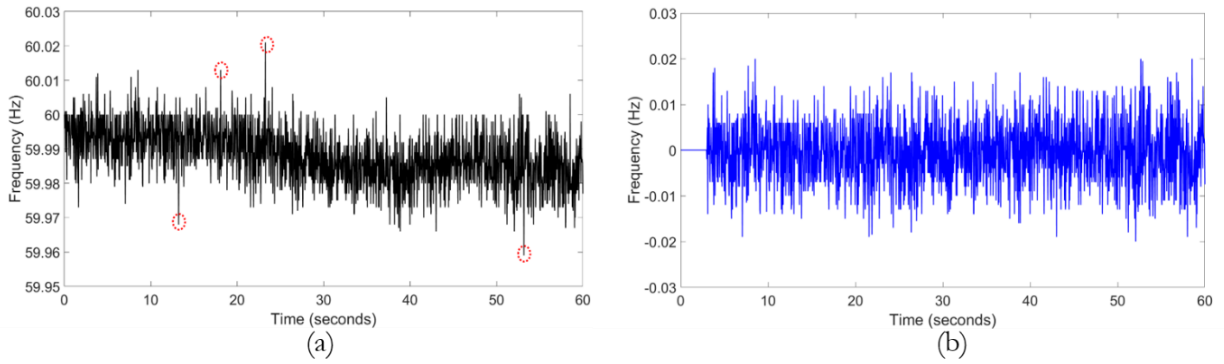


Figure 4.8. Noise in 1-min frequency measurement

Prior to noise extraction, a moving-window, median filter [81, 83] of different orders of 90, 150 and 300 samples were tested to eliminate outlier data samples which exceeded 3-standard deviations of the median value of a moving window. The red circles in Fig. 4.8 (a) are the eliminated outliers for a filter order of 90, after which the noise signal shown in Fig. 4.8 (b) was extracted. An average mean of approximately zero, the computed SNR of 43.1 decibels is adjudged to be typical of power system measurements. Other attributes of the signal are shown in Table 4.1.

Table 4.1. Noise signal attributes

	Frequency (Hz)	**Per unit (p.u)**
Mean	4.3×10^{-4}	7.2×10^{-6}
Standard Deviation	6.9×10^{-3}	6.9×10^{-3}

A further assessment of the signal to confirm the 'independent and identically distributed (*i.i.d*)' property of the noise signal samples can be carried out by computing an autocorrelation coefficient, ρ as a function of different values of lag k [84, 85].

$$\rho_k = \frac{E[(z_t - \mu)(z_{t+k} - \mu)]}{\sqrt{(E[(z_t - \mu)^2]E[(z_{t+k} - \mu)^2])}} \tag{2}$$

Z is a measurement sample; μ is a constant mean over the entire range of measurements; $E[(z_t - \mu)(z_{t+k} - \mu)]$ is an auto-covariance function which measures the co-variance between any sample that are k distance apart; and $E[(z_t - \mu)^2]$ is a self-correlation of a sample (i.e., correlation with respect to itself). Fig. 4.9 is the generated autocorrelation function (ACF) for k values up to 20.

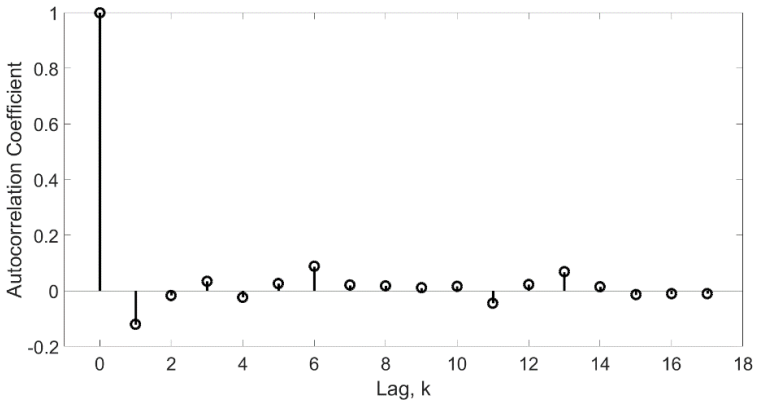


Figure 4.9. Autocorrelation function for noise signal

In an ideal scenario, noise samples are independent of each other, such that at any lag, $k \neq 0$, the ACF should equal zero, and one if otherwise. Fig. 4.9 approximates this behavior with a small ACF value of less than 0.2 at $k = 1$ before rolling off to zero at the next lag value. Neglecting its

self-correlation (i.e., $k = 0$), an average ACF value of 0.06 over all the different lag values is indicative of the strong *i.i.d* feature in this noise signal.

Local Outlier Measurements

Intermittent data samples, known as local outliers, which show significant deviations among neighboring time points are a common feature of rich datasets [86]. This is not uncommon with real PMU measurements where the transient nature of line or capacitor switching, noise, extra-terrestrial effects of sharp climate change (e.g. lightning) introduce data spikes in measurements. PMU 6 in Fig. 4.3 is a typical example of a local outlier sample where a data sample is observed to show a significant deviation from its neighbor samples. Fig. 4.10 shows the results obtained after an analysis of outlier samples in voltage magnitude (VM) and angle (VA) measurements of all real PMUs in the system over a total duration of 30 minutes (i.e., 54,000 samples for each PMU).

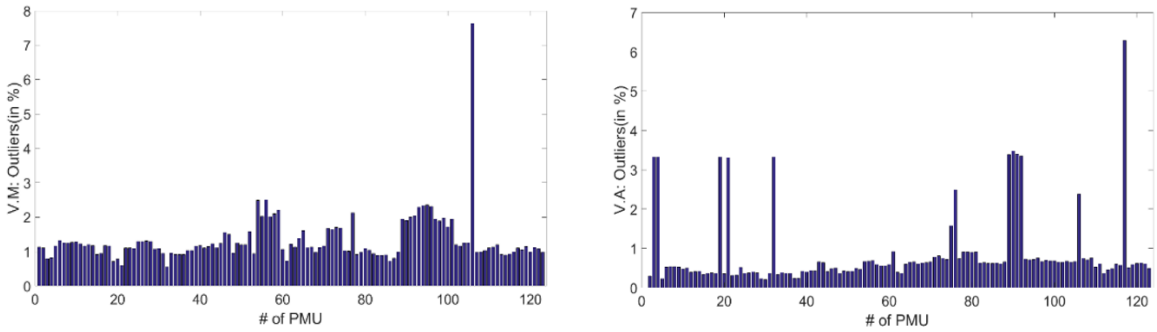


Figure 4.10. Percentage outliers in voltage magnitude and angle

A similar, moving-window, median filter of order 90 (or 3 seconds) has been used to identify outliers beyond 3-standard deviations of the window median. Both figures indicate the consistent

presence of outliers in power system measurements. Neglecting the PMU with over 7% VM outliers, an average of 1%, corresponding to 540 samples, are noted to be outside their local vicinities. Significant outlier angle measurements in Fig. 4.10 (b) is a unique feature of real PMU measurements where sporadic measurements with large deviations as much as 100 degrees need to be eliminated prior to data processing. While working directly with angle measurements can be challenging as a result of its significant non-stationarity, a first-order angle difference is used to obtain a stationary time-series for voltage angle outlier analysis. Here, the difference between consecutive voltage angles for any PMU is used to obtain a new time-series. That is,

$$\boldsymbol{VA}_{d,i} = \boldsymbol{VA}_{i+1} - \boldsymbol{VA}_i; i = 1,2, \ldots 53999 \tag{3}$$

$\boldsymbol{VA}_{d,i}$ is the new, stationary, angle-difference time-series for the PMU.

Fig. 4.11 (a) is the first-order, angle difference for all PMUs, and shows the wide variations that are possible with PMU-obtained voltage angle outliers, while Fig. 4.11 (b) shows a closer view of angle outliers in two selected PMUs with significant levels of outliers as observed in Fig. 4.10 (b).

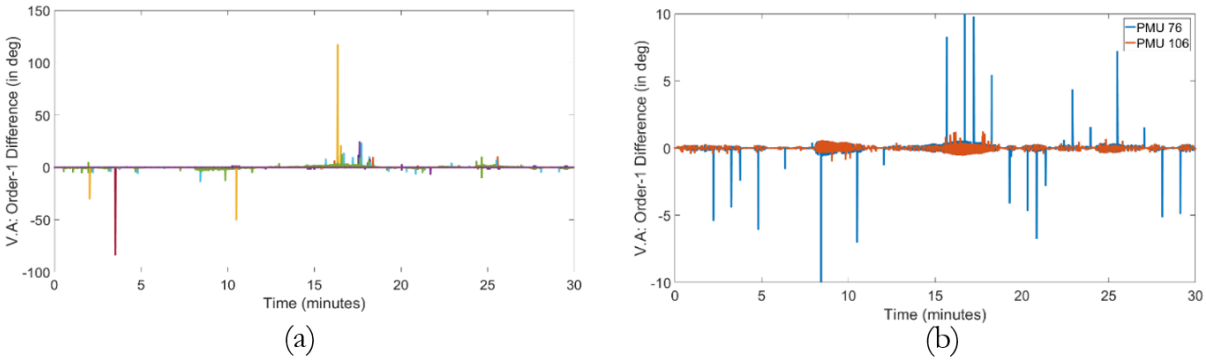


Figure 4.11. First-order, stationary, 30-min voltage angles

PMU Data Errors

The ubiquitous presence of measurement errors and anomalies in PMU data also increases the variabilities in real PMU measurements. Reliability issues associated with PMUs, and some of the prevalent PMU errors were previously discussed in chapter 3.

As an example, real time-skew errors due to drifting PMU clocks was observed for PMUs 3 and 4 in Fig. 4.3. Given the true existence of these anomalous data, a literature search for the statistics on the occurrence of different time errors was carried out. Authors in [39] reported the majority of GPS signal loss cases to be short time loss durations often occurring for less than a minute, however affecting a large number of PMUs. An average, daily loss rate of five times a day, and an average loss duration of 6.7 seconds were observed. Other notable measurement errors in existence include repeated (or stale) measurements and error due to instrument channel bias.

Missing Measurements or Packet Drops Due to Network/Communication Issues

An IEEE standard [21] stipulates that beyond a maximum wait or delay time, PDCs and other PMU reporting infrastructure replace time points with data fillers, if no measurement is available thus leading to missing measurement samples. In addition to NaN, the use of other arbitrary values such as 9999 or -9999 to represent missing samples is a common feature. The unavailability of true, actual measurement samples can be attributed to poor performance of PMU infrastructure (inclusive of data aggregators known as phasor data concentrators), or data packet drop outs due to the underlying communication network.

A data completeness problem, this PMU data attribute is defined in terms of a *drop-out rate* (ρ), which quantifies the number of packets dropped in a time period, and a maximum *drop-out* (*or gap*) *size* (χ), which defines the largest contiguous set of time points when no data sample is available [22].

$$\rho = \frac{\text{Number of dropped samples}}{\text{Total number of expected samples within time period}} \tag{4}$$

Fig. 4.12 is an illustration of missing values in 30 samples of PMU data with a drop-out rate of 60%, and maximum drop size of 6.

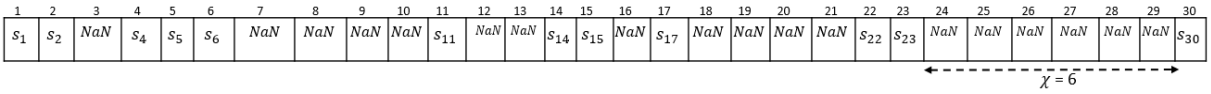


Figure 4.12. Missing data samples in PMU measurements

Using the 30-minute data of the real PMU measurements, the statistics of these phenomena can be determined. Each PMU has a report rate of 30 samples per second, thus giving a total of 54,000 (=30 × 60 × 30) data points. The percentage drop-out rates, ρ for voltage magnitude (VM), and voltage angle (VA) for all PMUs are shown in Fig. 4.13 (a) and (b), respectively.

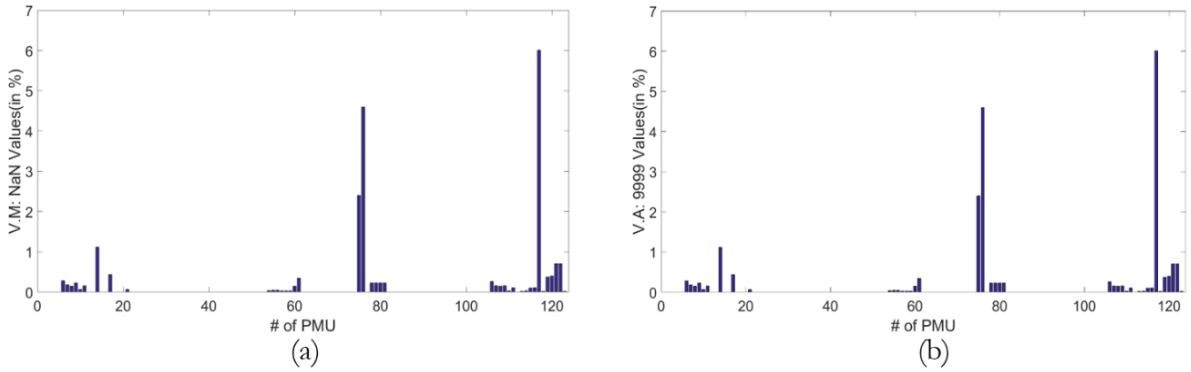


Figure 4.13. Drop-out rates in voltage magnitude and angle measurements

The correlated statistics observed for both voltage quantities indicates that actual packet loss results in complete loss of information for that time point. Three PMUs (with IDs 75, 76 and 117) are observed to show significant levels of missing data (i.e., beyond 2%). A further analysis of the severity of missing data, assesses the maximum drop size in each PMU shown in Fig. 4.14.

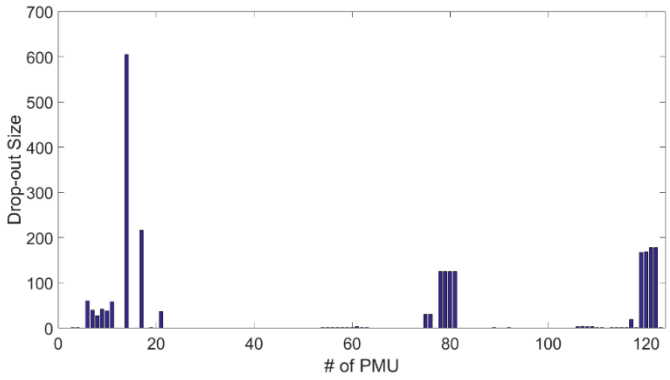


Figure 4.14. Maximum drop-size in 123 PMUs

From the figure, it is observed that 47 PMUs, accounting for 38% of all PMUs, had at least one instance of missing value, while unavailable measurements were observed from ten PMUs for a significant length of time (i.e., > 3seconds). The statistics presented in Fig. 4.13 and 4.14 is to demonstrate the prevalent cases of missing data samples primarily due to data packet drops, and which ultimately defines some of the unique features observed in industry-grade PMU measurements.

4.2 Synthetic PMU Data Creation

The production of synthetic data for research purposes have been addressed in several fields related to software testing, machine learning, and social networks [86-91]. Majority of these approaches utilize intelligent techniques, such as genetic algorithms, ensemble-based methods, R-programming, and rely on pre-defined models, patterns or random number generators to create artificial data. However, due to the multiple component and human operator interactions with the grid, power system measurements are unique as they embed underlying system dynamics which

reflect the state of the grid. As a result, they are not random nor do they strictly follow any pre-defined pattern. In order to circumvent the reliance on power system component models, [91] proposed the use of an intelligent generative adversarial network (GAN) machine learning technique to synthesize a realistic PMU time-series measurement. A limitation with this method is its significant level of dependence on real data, and for which the confidentiality issues associated with accessing real data was the original motivation for synthetic data production. Furthermore, the ability to modify or make certain inclusions to features in the synthetic dataset, using the proposed method, may be limited since the artificial data is based on an original measurement. In situations where large-scale, multivariate datasets are required from multiple, geographically-dispersed sources, such that they embed all underlying system dynamics in addition to local behaviors, and simultaneously capture the intricate spatio-temporal relationships known to exist in electric grids [79], an inability of the current methods to train multiple real data while satisfying the above requirements would significantly restrict their implementation [87].

To address the above-mentioned issues, a proposed framework for the generation of realistic, synthetic PMU measurements has been developed in Fig. 4.15.

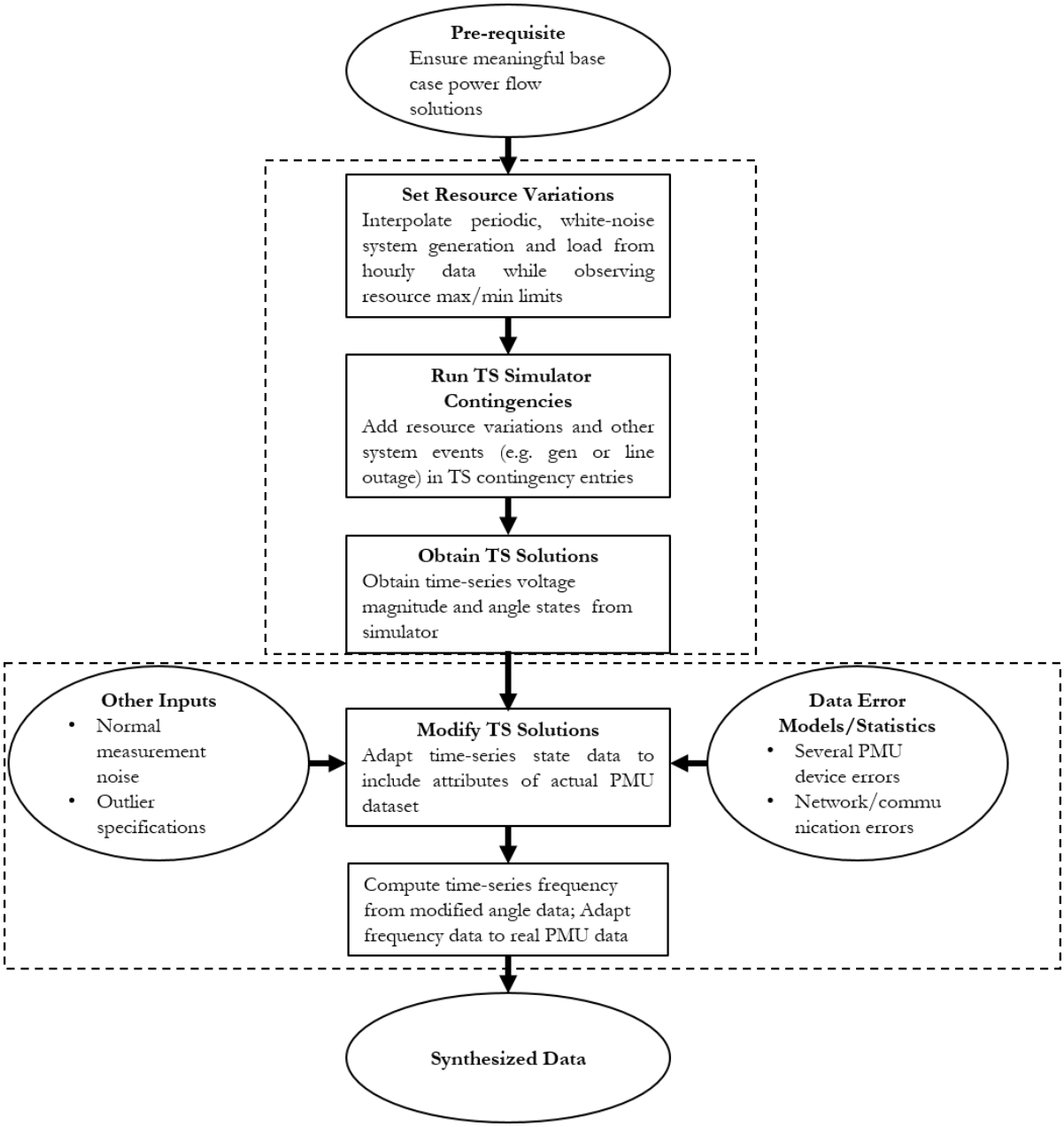


Figure 4.15. Framework for creating synthetic PMU data

The framework comprises of two main steps. Firstly, input data made up of annual, seasonal generation, and white-noise, load variations are fed into a simulator power flow solver. Inclusive of other actions, such as automatic generation controls (AGCs) and temporal or permanent line

outage disturbances based on historical data, the solver -Transient Stability (TS) Simulator - is used to obtain grid state measurements in a synthetic grid within a given time resolution level. In a second step, data modification activities are performed on these simulation measurements. Given a pre-defined PMU sample report rates, these measurements are padded with fictitious data points fit enough to simultaneously satisfy system dynamics and the random variations observed in real measurements. Finally, an integration of these measurements with different PMU data errors is used to add more realism to the generated artificial dataset. The proposed method is aided with the use of a power systems simulator with the sole purpose of executing transient-level stability analysis on the grid. For validation, principal component analysis (PCA) is used as a tool to verify the retention of the underlying, true system dynamics in the synthetic dataset; and in addition to the PCA, an average variability metric is used to assess its resemblance with real PMU data.

4.2.1 Power System Operations

Several dynamics, belonging to a wide range of time scales, have been known to occur on the grid during normal system operations. Fig. 4.16 shows a typical timeframe for different events during the operation of a grid [92], and which in turn affects the feature sets of PMU measurements.

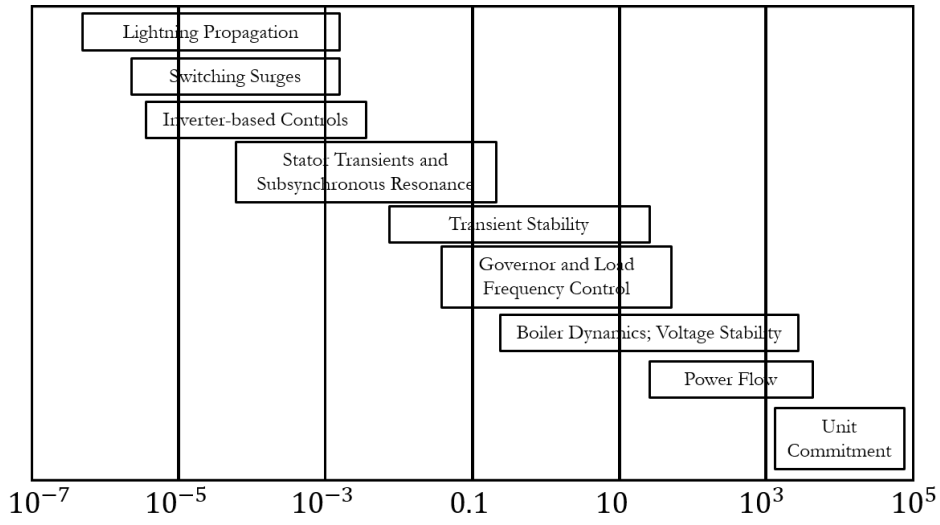


Figure 4.16. Time frames in power system operation

Generally, small time-scale events translate to small-duration transients in measurement samples, while low frequency and system variation patterns are associated with longer time-scale events. Since the availability of a comprehensive simulation which captures all these different dynamics is limited, this work focuses on only time-scales associated with, and longer than, the transient stability duration.

The electric dynamics associated with generator governor, load frequency control and boiler have been preset in generator and load models implemented in the synthetic grid used for the simulation. Shunts and generator reactive powers are used to realize voltage control, while the solver in a power systems simulator software is used to compute the states of the system at predefined time steps. The availability status (ON/OFF) of all generators and their corresponding automatic generator control (AGC) settings are left unchanged in the simulation. Thus, the variables for control are selected system generators and total system load.

System Generator and Load Variation

Power system load fluctuations have been observed to follow specific trends, which often aids system planners and operators to prepare for a concomitant level of generation required to match future load forecast [74, 93-95]. Fig. 4.17 shows identical, daily load profiles for three customers within a 24-hour period [74].

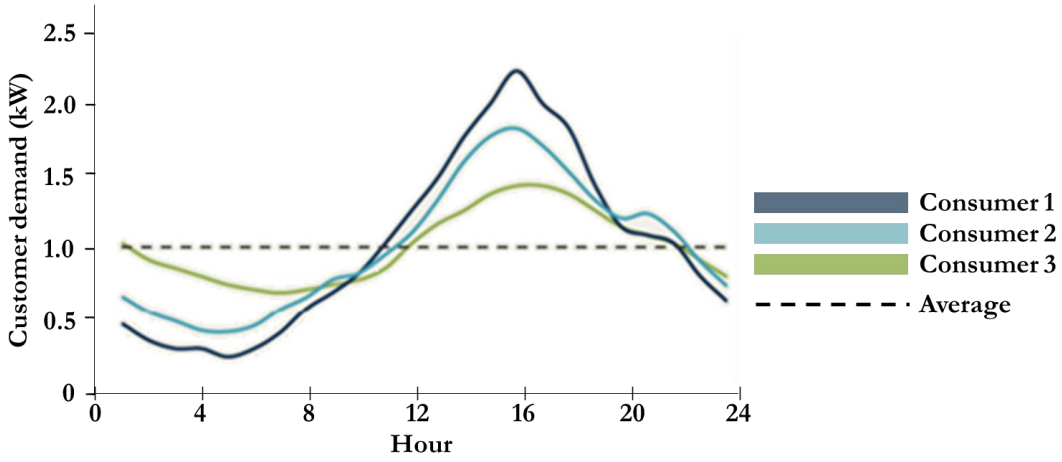


Figure 4.17. 24-hr load demand

Electricity power demand is observed to follow expected social and human behavior – peak and base electricity usage in the late afternoons and early mornings respectively. To match this behavior from a grid perspective, power generation is made to track the trend in the changing-load. A combination of these variations partly gives rise to the random patterns observed in real, high resolution PMU data.

The smooth load curves in Fig. 4.17 show hourly load variations, however they do not capture the small, time-scale load changes which occur in practice. In this work, we leverage real power system load data currently available to the research group to scale the individual loads at different

bus locations of the synthetic grid [96]. Hourly load profiles for any load bus are then decomposed into smaller, per-second resolutions to generate a time-series of load values used during the simulation.

Given the bus loads at two, consecutive hours as L_1 and L_2, a load level L_i at any i^{th} second after the first hour at L_1 is computed by interpolating between L_1 and L_2, i.e.,

$$L_i = L_1 + i \times \frac{(L_2 - L_1)}{3600} \tag{5}$$

Here, 3,600 is the number of seconds in an hour. The procedure (5) is separately implemented for both the active (MW) and reactive (Mvar) components of the load, thus maintaining a constant load power factor. Depending on the extent of expected load variations, the inclusion of a white-noise component aids the realization of load randomness observed in practice [97, 98].

$$L_{i,rdm} = (1 + \sigma_L)L_i \tag{6}$$

$L_{i,rdm}$ is an improved load obtained from a variation, σ_L in a given load signal-noise ratio.

In similar fashion, the interpolation computation of (5) is performed for a generator power output at the i^{th} second (G_i) between two consecutive, hourly generations of G_1 and G_2. That is,

$$G_i = G_1 + i \times \frac{G_2 - G_1}{3600} \tag{7}$$

Due to the short duration that has been simulated, and in accordance with real life expectation, predominantly renewable wind generation sources have been considered as nodes with changing levels of power output. The research works of [99, 100] break down wind power variation, $P_W(t)$ into three components - a slowing moving average (P_a), a zero- mean fluctuating part (P_t) and a ramp event (P_r) as shown in (8)

$$P_W(t) = P_a(t) + P_t(t) + P_r(t) \tag{8}$$

In this work, only P_a and P_t have been considered since ramp component P_r, is observed to be affected by several other events. P_a indicates the per-second trend power output, and obtained from the interpolation step in (7), while P_t is modeled as a noise component, similar to the load. Substituting all load symbols L in (6) with G, and using a σ_G-parameter associated with generator signal-noise ratio, the improved, instantaneous generation at an i^{th} second is computed likewise. That is,

$$G_{i,rdm} = (1 + \sigma_G)G_i \tag{9}$$

However, in practice, true generator power output is limited by a manufacturer capability curve which defines the extents of real and reactive powers produced. A logical assumption is to constrain the real and reactive components of $G_{i,rdm}$ for any generator unit to the pre-defined power limits that has been set for that generator in the simulator.

Line Outages and Other Disturbances

Transmission line outages due to planned and unplanned events contribute to varying levels of system disturbance. Depending on the nature of outage (sustained or momentary), its effect can often be significant on the grid. Using historical outage data [101, 102], the frequency of line outages can be obtained, and incorporated into the simulation activity. The same is applicable to other disturbance types, such as line faults and generator trip.

4.2.2 Simulator Specifications

Prior to the creation of realistic, synthetic PMU datasets, there is a need to generate base, first-level, simulation measurements with a common and meaningful grid dynamics. This is achieved using a power grid software, PowerWorld simulator [103] that is capable of executing transient

stability (TS) level simulations in high voltage power system operations, and has the ability to provide simulation data with common-mode dynamics for any size of power system.

In order to capture transients and full system dynamics during grid events, a TS study is chosen over a steady state power flow solution which is carried out only at defined steps without transient information. The time steps during which loads and generators are observed to vary are then set as contingency entries in the TS simulation options. Upon successful completion of a simulation, the desired power flow results of voltage magnitude, angle or any other quantity are extracted.

4.2.3 Simulator Results Enhancement

Error-free simulation results (e.g., voltage magnitude and angle) are obtained after a TS experiment, and are adjusted to emulate industry-type data. The data modification is implemented based on statistics obtained from real measurements. Currently, only voltage magnitude and angle data obtained from the simulation have been enhanced.

System and Noise Variability in Measurements

In this work, variability is introduced to the error-free simulator measurements in three major steps.

1. Firstly, based on an average system variability in real voltage measurements (shown in Fig. 4.7 (a)), a multiplier is chosen to scale the observed average variability in the simulation data.
2. Secondly, new sets of measurement values are obtained for each PMU by computing the average value of non-overlapping windows, followed by the inclusion of a white-noise deviation, $\sigma_{\Delta V}$ for randomness. This step ensures that new measurements mimic real data variability.

3. Finally, given a pre-defined deviation level, σ_η measurements from step 2 are enhanced with noisy signals.

Further enhancements to the synthetic dataset are carried out by the inclusion of outlier measurement samples, data errors such as repeated measurements, bias, measurement noise, different time-based errors, and missing data samples based on some of the statistics already stated.

Derived Measurements: Frequency and ROCOF

Depending on the type and extent of modifications carried out on the original transient-stability phasor angle measurement, re-computing derived measurements of frequency and ROCOF may be necessary to reflect phasor angle anomalies. For this purpose, any of the computation methods described earlier in Section 3.1.3 can be implemented.

4.3 Case Scenarios and Samples of Synthetic Voltage, Angle and Frequency Measurements

Realistic, synthetic PMU voltage and angle measurements are created using the framework described in Section 4.2. The simulation is run on a 2,000-bus network, with an operating frequency of 60Hz, using the PowerWorld TS option after which PMU measurements of a duration of 90-seconds are obtained from all the buses. Periodic variations in total system load were set to occur every 5-seconds, and those of selected generators to change every 7-seconds, and thereafter set up as contingency entries in the simulation. A time step parameter of 0.25 cycles, and set to store power flow results every 8 time steps was used in the TS solver specifications, thus mimicking a PMU report rate of 30 samples per second. Following the extraction of simulation results, modification steps to adjust simulation data to realistic, synthetic PMU data were implemented in MATLAB programming.

Some of the case scenarios that were simulated are:

1. A base case where only load and generation were set to vary over time
2. The base case and a generator trip
3. The base case, generator trip and a switched-in shunt

Besides the event detection implemented for all three cases in Section 4.4, all analysis of synthetic measurements in other sections refer to the base case.

4.3.1 Variability Assessment in TS Simulation Measurements

The extent of system variation as observed in both simulation results of voltage magnitude and angle are shown in Fig. 4.18 (a) and (b), respectively. They illustrate the average variability observed over the entire duration of the experiment in ninety-nine selected PMU locations. (The criteria for selecting these PMUs are given in Appendix B).

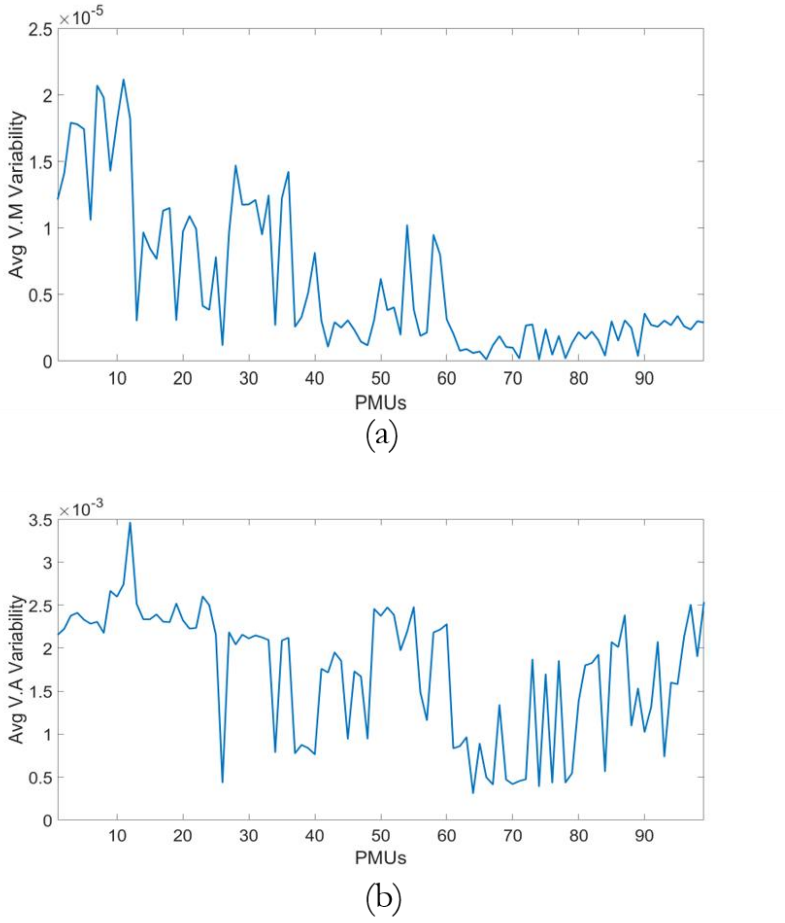


Figure 4.18. Variability in simulated voltage magnitude and angle measurements

It is observed that voltage angles exhibit wider variations (in the order of 10^{-3}) than its magnitude counterpart (in the order of 10^{-5}). This might be attributed to the fact that the analysis is performed on per unit values of voltage magnitude, while angle measurements are left unchanged in degrees. In contrast with the real system which has average magnitude and angle variabilities of 3×10^{-5} and 1.5×10^{-2} respectively, the low average variations observed in the simulation magnitude and angle results (i.e., 5×10^{-6} and 1.5×10^{-3} respectively) still indicates a lower level of system activity in the simulation. Given an assumption that average variabilities computed for the real

system were obtained from a 'clean' segment of the real data based on the high SNR values (see Fig. 4.7), we can infer these variabilities to be the reference values. We can thus scale average variability values of the simulation data to match those of the reference values. Here, multiplier values of 2.0 and 10 have been used for the average magnitude variability, and average angle variability respectively.

4.3.2 Simulation Data Re-creation

The process for re-creating new data measurements, $\mathbf{S}' = \{s_{1'}, s_{2'}, s_3, \ldots\}$, from an initial, set of simulation data samples , $\boldsymbol{S} = \{s_1, s_2, s_3, \ldots\}$ is shown in Fig. 4.19.

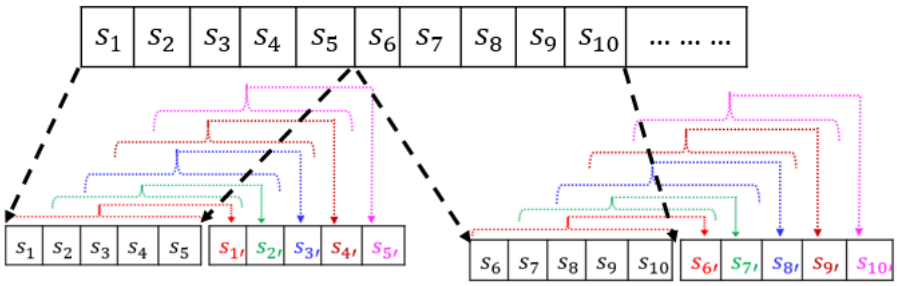


Figure 4.19. Introducing variability to simulated measurements

A data sample (s_i) in a non-overlapping data window, extracted from the original measurement, is replaced by a new sample ($s_{i'}$), which is the sum of its window mean (s_{ave}) and its variation component due to a variation factor (F_v).

$$s_{i'} = s_{ave} \times (1 + F_v) \tag{10}$$

This method of a moving averaging window through different data segments ensures the retention of the underlying dynamics, while systemic white-noise, variabilities are introduced in the measurements.

Fig. 4.20 compares a 5-second, re-created, synthetic per-unit voltage measurement with its original, simulation data.

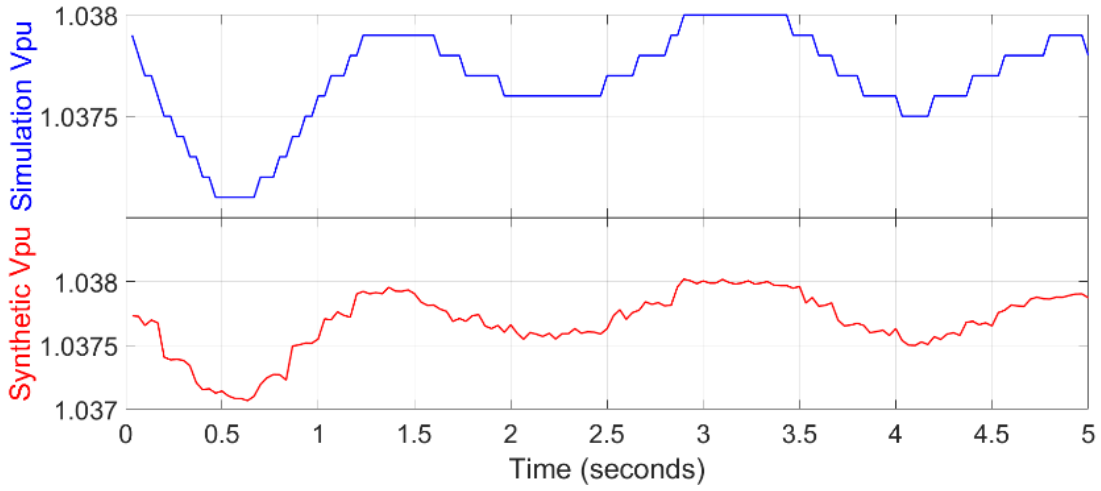


Figure 4.20. Per unit voltage measurements: simulation versus synthetic data

The discrete nature of voltage observed in the blue, simulated measurements is due to the report time step of 0.033-sec during which a value is held constant. As previously stated, this report interval corresponds to eight individual step sizes, each of 0.25-cycle at which the state of the system is evaluated. The longer periods of constant values, during which no perturbation is observed to occur, can be attributed to the largely inert state of the synthetic grid used in the simulation. Upon transformation to the red, synthetic measurement, the proposed moving-average scheme described in Fig. 4.19 is able to ensure smooth transitions between consecutive voltages, amidst the introduction of white-noise disturbances typical of real system. The new system average variability is distributed across different samples similar to the real PMU data in Fig. 4.6.

To increase signal variability, further enhancements can be carried out by injecting additional levels of noise to the synthetic measurement.

4.3.3 Validation

The initial validations which have been used to evaluate the accuracy of the proposed framework in this study were categorized into two - the ability of the synthetic dataset for the system to retain the underlying, electrical behavior or dynamics inherent in the original TS-simulation dataset, and a comparable average variability level with the real data. For this purpose, principal component analysis (PCA) has been utilized.

Electrical Dynamics Behavior

Principal Component Analysis (PCA)

Given a power systems dataset comprising of several bus measurements, PCA technique [104, 105] is used to extract the major underlying system dynamics by re-representing the dataset in a lower dimension, while retaining all primary attributes of the data. The technique re-expresses a multidimensional data set, $\boldsymbol{X}$ consisting of n measurement locations, into its most meaningful set of basis. An orthonormal matrix, $\boldsymbol{P}$ (the basis) is known to diagonalize a covariance matrix, $\boldsymbol{S_Y}$ (= $\frac{1}{n-1}\boldsymbol{YY^T}$), such that $\boldsymbol{Y} = \boldsymbol{P^T X}$. An eigen-decomposition of $\boldsymbol{S_Y}$ yields an ordered set of eigenvalues and eigenvectors,

$$\lambda = \{\lambda_1, \lambda_2, \dots \lambda_m\};\ \lambda_1 \geq \lambda_2 \dots \geq \lambda_m \qquad (11)$$
$$\boldsymbol{P} = \{\boldsymbol{PC_1}, \boldsymbol{PC_2}, \dots \boldsymbol{PC_m}\};\ \boldsymbol{PC_i} \in \boldsymbol{R^n}$$

m is the number of retained principal component vectors, $\boldsymbol{PC_i}$ in $\boldsymbol{P}$, and the order of importance of these vectors is given as $\boldsymbol{PC_1} > \boldsymbol{PC_2} > \cdots \boldsymbol{PC_m}$ based on the decreasing magnitudes of eigenvalues, λ_i.

Understanding the large number of time points at which the system is to be evaluated ($i.e., \sim 90 \times 30 = 2{,}700$), the dataset is divided into equal window segments after which PCA is performed. Given a threshold, percentage variance, $Variance_{th}$, we can compute the m number of principal components (PCs) whose percentage cumulative variance, computed by the aggregation of each eigenvalue and beginning with the largest value, λ_1 , just equals $Variance_{th}$ in each window. That is,

$$Variance_{cumm} = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \times 100 = \ Variance_{th} \tag{12}$$

Fig. 4.21 shows the number of PCs per window segment in the simulation and synthetic voltage datasets. Here, $Variance_{th}$ has been set to 98%, and each segment is 4-second (or 120 data samples) duration.

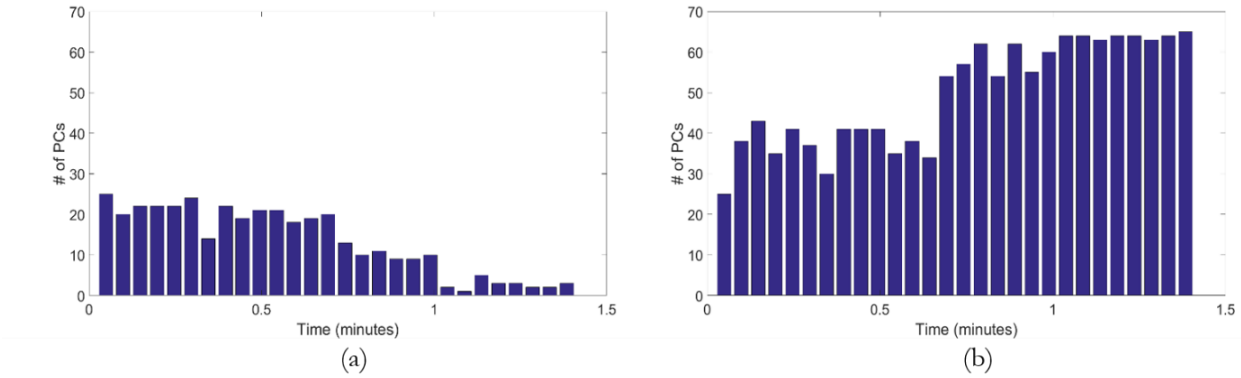


Figure 4.21. Number of principal components per window: simulation versus synthetic data

The following observations can be made from the figures:

1. Given local effects of voltage, the number of PCs during the period of initial, significant system load and generation changes (in Fig. 4.21 (a)), is observed to be relatively high, as

a result of the multiple, unique grid dynamics occurring at different segments of the system. The reduction (and non-existence) of grid dynamics results in much smaller PCs when the behavior of the system is constant across all buses.

2. In comparison with synthetic data in Fig. 4.21(b), boosting overall system dynamics, by a variation factor causes a surge in local grid, and hence an increased number of PCs in the initial window segments. However, a similar trend in the segment-by-segment PC counts in both simulation and synthetic datasets is observed
3. In contrast with Fig. 4.21(a), which has low counts of PCs attributable to grid uniformity towards the simulation completion, larger PC counts are recorded in Fig. 4.21(b). This can be attributed to the fictitious variations introduced separately for the different measurements. Several behaviors, independent of each other, are thus observed on the grid resulting in more PC counts.
4. In conclusion, the underlying electrical behavior of the simulation system is retained in the synthetic measurement amidst the newly-introduced, increased activity levels.

Real System Behavior

Neglecting the effects of measurement errors associated with real PMU data, and considering only the clean segment of real data as stated in previous sections, a comparison between the average variabilities in both real and synthetic voltage measurements is shown in Figs. 4.22.

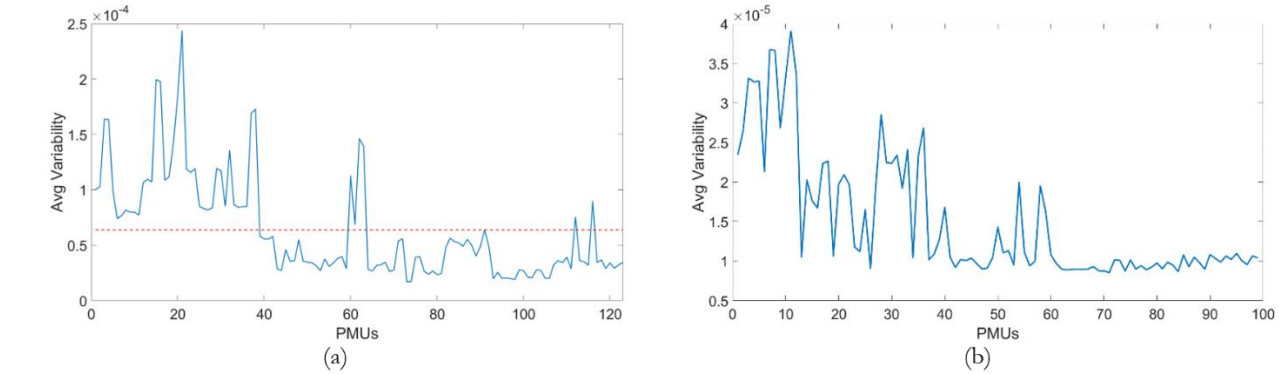


Figure 4.22. Average variability: real versus synthetic data

Without taking into consideration the erroneous PMU observed with abnormally high variability (see Fig. 4.7 (a)), a system average variability of 0.6×10^{-4} or 6×10^{-5} shown by the red, dotted line in Fig. 4.22 (a) was used as a reference value for the synthetic data generation process. Fig. 4.22 (b) is the average variability across all PMUs in the synthetic data after scaling the variability in the original simulation by 2.0, and introducing fictitious system variations. The average PMU variability in all PMUs of the synthetic dataset are observed to be the same order as the system average value in the real system. In terms of SNR, the synthetic system, as shown in Fig. 4.22 (b), is observed to show higher degrees of non-uniformity across all PMUs. Given the figure, a recommendation for a higher variation factor, and applied more homogeneously across PMUs is suggested to ensure an SNR graph similar to the real system in Fig. 4.22 (a).

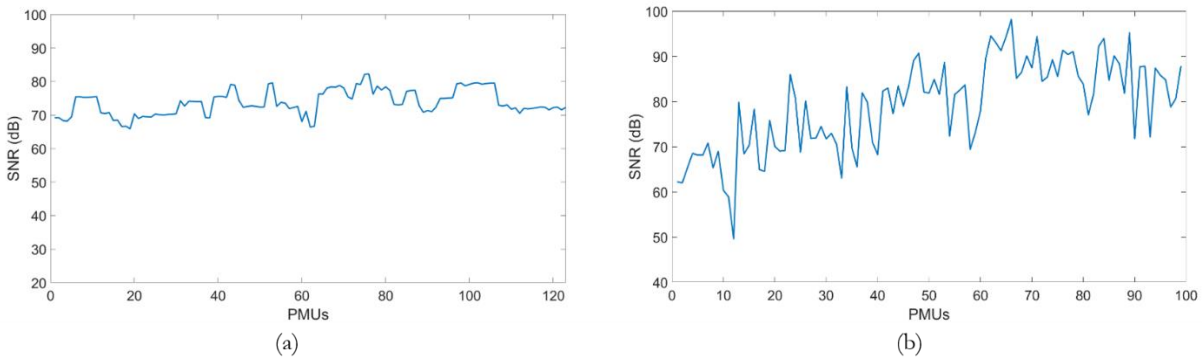


Figure 4.23. SNR: Real versus synthetic data

Further comparisons between the synthetic and real system by assessing their respective PC counts is given by Fig. 4.24.

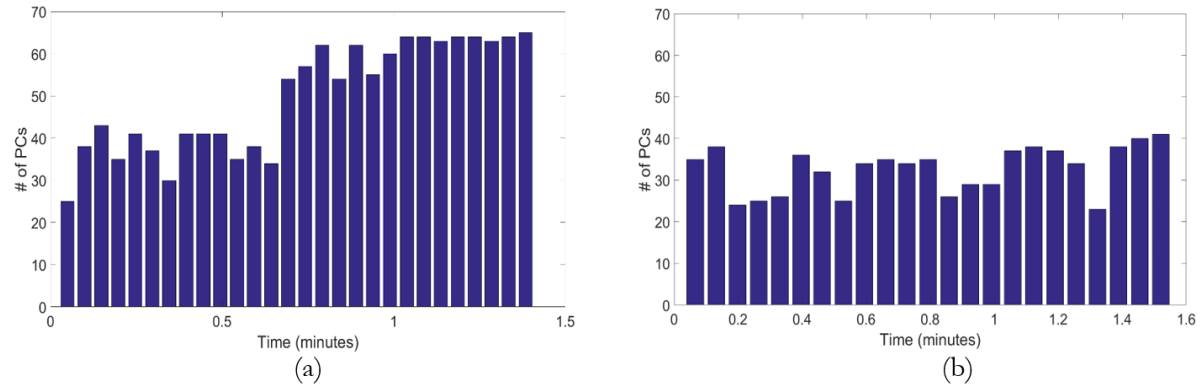


Figure 4.24. Window number of principal components: synthetic versus real data

Prior the time of non-significant, dynamics interactions in the synthetic system (i.e., 0.7 minutes, or approximately 42 seconds) as shown in Fig. 4.24(a), a similar average PC count is observed with that obtained from the real system in Fig. 4.24 (b). In contrast to the global behavior of frequency, grid voltage, as a local phenomenon, tends to vary slightly across the grid. Hence, the

relative large number of principal components required to capture the defined variance level in each window. However, this is only valid when no significant, uniform voltage disturbance is observed across the system

4.4 Event Identification Analysis

The real dataset and synthetic PMU datasets, obtained from the scenario cases in Section 4.3, are further explored for error and event identification.

4.4.1 Data Pre-processing

An effective analysis of PMU data requires a prior removal, and possible replacement, of data samples or segments containing outliers, data sources reporting all anomalous measurements, and smoothening out data sections containing high frequency transient disturbances.

Visual Inspection

1. Elimination of out-of-range and inconsistent measurements. For example, in the real data set PMUs observed to report steady-state frequency values of 27.2 Hz and 60Hz were removed. The former related to an out-of-range measurement, and the latter was a measurement inconsistent with the actual trend of varying system frequency.
2. Replacement of missing PMU samples, represented as ‘NaN’ and ‘-9999’ values in voltage magnitude and angle measurements respectively. Here, an average of a moving window containing past samples was used to provide measurement samples at time points when no data was available.
3. Removal of suspicious measurements exhibiting significant resemblance to analog-type, measurements. For example, the sinusoidal-like voltage measurements from a PMU.

4. Optionally, the removal of signals with ‘excessively high’ noise levels relative to other measurements even though noise level is within acceptable values. For example, a PMU reporting frequency measurements with noise level of 43 decibels signal-noise ratio (SNR) was eliminated when an average of 55 decibels was observed for all other sources.

In addition, an unwrapping of all PMU voltage angles, followed by a re-calculation of all measurements based on a chosen reference is needed for the usability of voltage angle measurements.

Data Filtering & Outlier Removal

Due to its ability to retain actual disturbance events while still robust to the presence of measurement noise, the median filter is preferred [76, 81, 83]. Given the median of a moving window of past samples and a pre-defined m threshold number of standard deviations, a signal sample is rejected if it falls outside the tolerable limits. A choice of $m = 3$ standard deviations, and a filter window order of 90 is used for this work. However, depending on the time-constant of desired transients to be eliminated, these parameters can always be selected to fit the purpose.

4.4.2 Event Detection Using Moving Window Methods

Given that large amounts of data streams are obtained from multiple grid PMU locations, moving window approaches are proposed to carry out comparisons between time windows in the search for common-mode system events. Depending on the measurement quantity being monitored, the occurrence of a global, system disturbance will be captured by several PMUs at their different grid locations, while local events will be observed in a relatively lower number of PMUs within the vicinity of the event. Apparently, disturbances which reflect in only single PMU

locations will be classified as measurement errors. In this section, a deviation trend analysis and principal component analysis are utilized as analytic tools to identify event time points in datasets.

A Variance Trend Analysis (VTA)

The red, down-sampled plot in Fig. 4.6 was previously described as showing the variances in consecutive segments within a synchrophasor measurement. Given a careful selection of time window, an erroneous PMU reporting bad measurements can be detected when sudden, abnormally high deviation values occur. In the analysis of multiple synchrophasor measurements, inconsistent measurements can be identified when large deviations are observed in sole PMUs. On the contrary, actual system disturbances, with large deviations, will reflect in more than one PMU simultaneously. Similar to the detrended fluctuation analysis (DFA) which was used in [106, 107], however, this proposed approach improves on runtime by avoiding an unnecessary computation of a high-dimension, integration signal vector obtained from the large number of measurement samples of each PMU.

Fig. 4.25 is the *VTA* analysis performed on 1-minute of per unit voltage data.

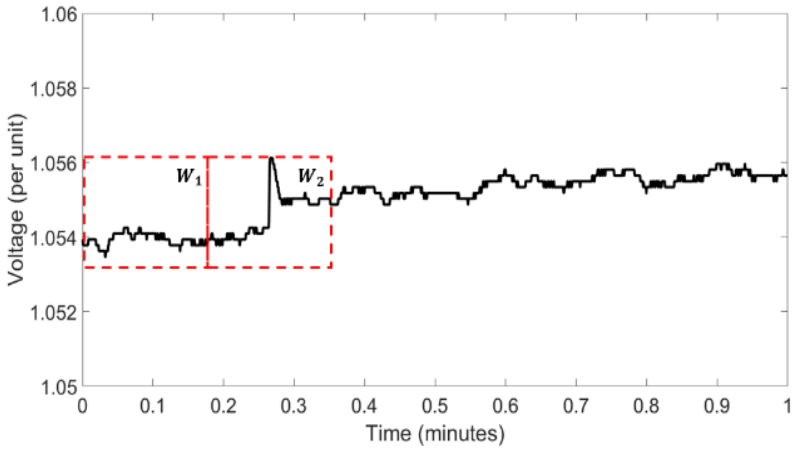


Figure 4.25. Window variance in a VTA analysis

Only two windows, W_1 and W_2, are shown in the figure. As observed the variance of W_2 is significantly larger than that of W_1. A similar analysis performed simultaneously on several PMUs will reveal the true identity of these observed glitch in the data.

Principal Component Analysis (PCA)

In a previous section (Section 4.3.3), the PCA technique was discussed. The windowing method which has been used to search for events is illustrated in Fig. 4.26

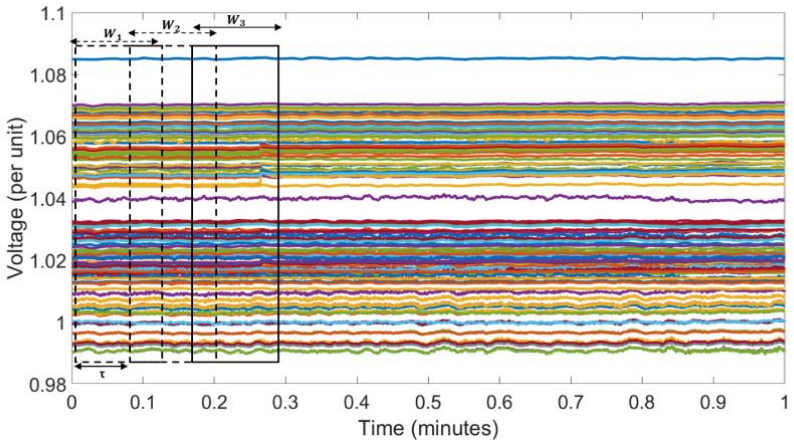


Figure 4.26. PCA window-window comparison on 1-min per unit voltage magnitude data

Using a pre-defined window-width size (l) and window-step size (τ), a similarity assessment of consecutive multi-variate, time-series windows $\boldsymbol{w_1} \in R^{l \times n}$ and $\boldsymbol{w_2} \in R^{l \times n}$ is performed to check for major system changes. In order to mitigate the interfering effects of individual PMU measurement errors and at the same time, magnify only true system dynamics, each window is re-represented with its top *k*-principal components which capture a threshold variance. That is,

$$\text{PCA}(\boldsymbol{w_1}) = \boldsymbol{P}^{n \times k} \tag{13}$$

$$\text{PCA}(\boldsymbol{w_2}) = \boldsymbol{Q}^{n\times k}$$

A disparity metric between consecutive windows is carried out by comparing the strength of their respective principal component vectors. This can be computed as a weighted, average sum of squares of the cosine angles between each of the principal component in $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ [108]. That is,

$$dsim(\boldsymbol{w_1}, \boldsymbol{w_2}) = \frac{\sum_i^k \sum_j^k (\lambda_{w_{1i}} \lambda_{w_{2j}} Cos^2\theta_{ij})}{\sum_i^k \sum_j^k (\lambda_{w_{1i}} \lambda_{w_{2j}})} \quad (14)$$

The disparity metric, *dsim* is observed to increase when a sudden change or large discrepancy is observed between time windows.

VTA and PCA Analysis on Real Dataset

The VTA and PCA techniques were separately used to analyze real PMU voltage measurements. Fig. 4.27 shows 30-miutes, per unit voltage measurements after the data pre-processing stage.

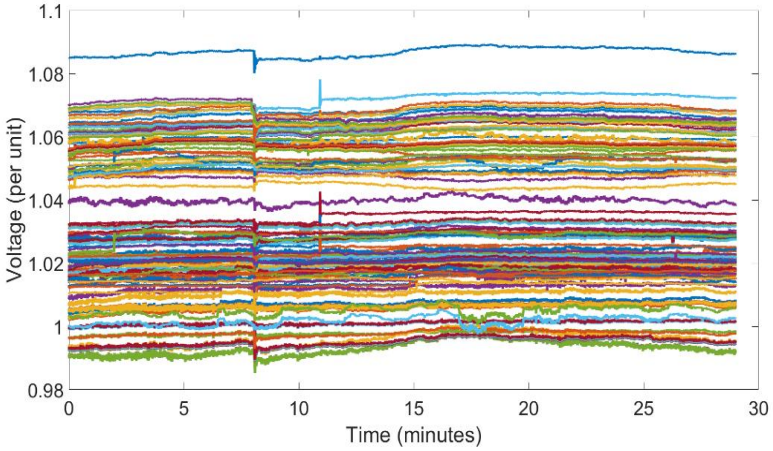


Figure 4.27. 30-min per unit voltage for all PMUs

The results obtained after VTA was performed on a moving, non-overlapping time-window of 1-second (or 30 samples) are shown in Fig. 4.28.

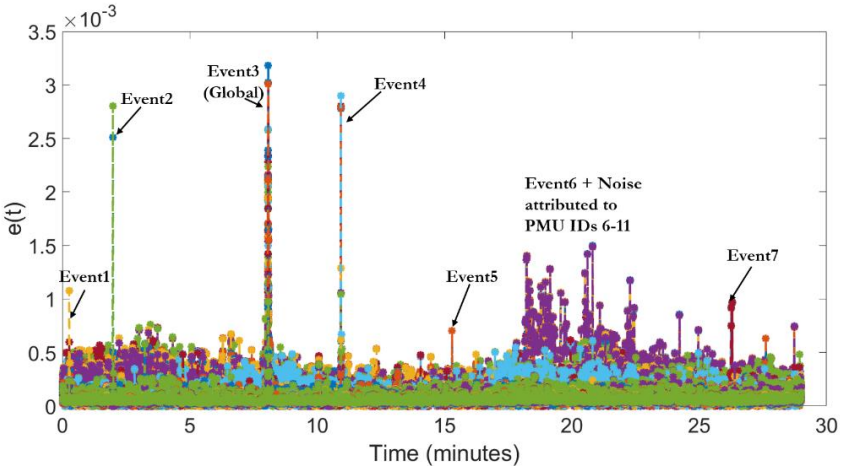


Figure 4.28. 1-sec window, e-values for 30-min duration of voltage measurements in 123 PMUs

A significant level of activity occurring on the real system is captured by the constantly, fluctuating values of the variance, $e(t)$ across the different PMUs. However, these values pale in comparison with actual grid events whose high variation levels are simultaneously observed in more than one PMU. The largest disturbance of *Event3*, and a maximum e-value of 3.18×10^{-3} at PMU 43, is the 8th minute generator trip during which a significant voltage dip, and large e-values, are observed at all PMUs. Here, *Event3* is classified as a global system disturbance. *Event2* and *Event4* are observed at 2nd and 11th minute respectively, and indicate local events concentrated on certain parts of the grids. *Event2*, with a maximum e-value of 2.8×10^{-3}, is concentrated at PMU locations 106 and 100, while *Event4*, with a maximum e-value of 2.9×10^{-3}, is concentrated at PMU s 77, 78, 79 and 83.

Table 4.2 shows maximum e-value of other events, extent at which they are observable based on the number of PMUs with significant variances, and a classification of either global or local event.

Table 4.2. Attributes of detected events

Event ID	Time (minute)	Max e-value($\times 10^{-3}$)	# of PMUs	Global/Local
1	0.25	1.08	> 4	Local (weak)
2	2.00	2.80	2	Local (strong)
3	8.00	3.18	> 65	Global
4	11.00	2.90	> 15	Local (strong)
5	15.00	0.70	2	Local (weak)
6	21.30	1.50	>10	Local (weak)
7	26.30	0.97	4	Local (weak)

The detection of PMUs 6-11 with abnormally high e-values, as a result of significant noise levels, paved for an identification of the local event behavior inherent in *Event6*. Fig. 4.29 is the updated e-value plot, after removal of these noisy PMUs.

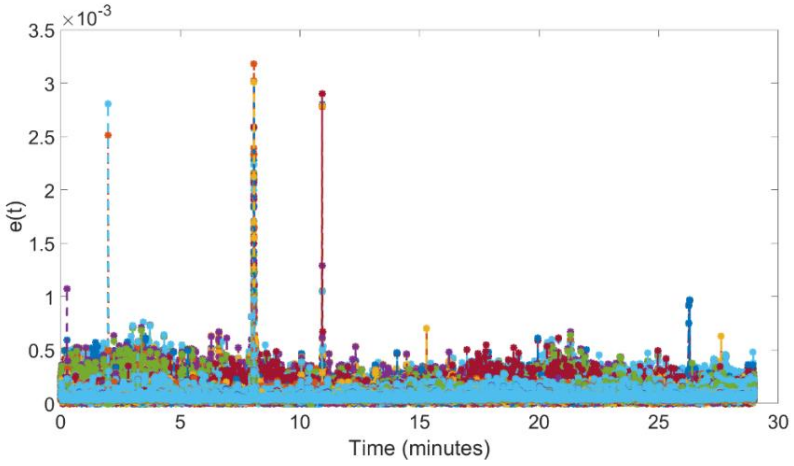


Figure 4.29. 1-sec window, e-values for 30-min duration of voltage measurements in 123 PMUs after removal of noisy PMUs

A further event detection analysis using a moving time-window of 3 seconds (90 samples), and applying the PCA-based, consecutive window, disparity check to search for events is carried out. Fig. 4.30 shows the disparity of each window relative to a previous, non-eventful time window.

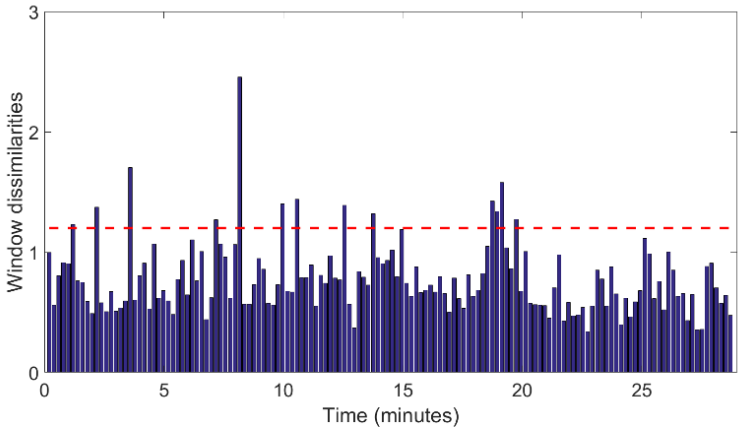


Figure 4.30. System time-window relative disparity levels for 30-min voltage measurements

The dissimilarity values in the figure have been normalized with respect to the first time window, and the observations discussed thus:

1. The data window containing the generator trip event in the 8th minute (i.e. global *Event3*) is observed to significantly differ from a preceding, non-eventful window, thus manifesting as a large dissimilarity value.
2. Local events, *Event2* and *Event4***,** previously observed to show large deviations in Fig. 4.28, are rather less pronounced with the PCA method in Fig. 4.30. However, the impact of *Event4,* is observed to be stronger than *Event2* since it is captured by more PMUs in the system. A similar observation is made for *Event1, Event5 and Event7,* whose local events have low impacts on nearby PMUs.

3. A notable revelation is the relatively, significant level of system activities detected at 3.6 minutes (or 216 seconds). Small magnitudes of variation, however observed in most PMUs, exposed a common-mode system disturbance occurring at this time period.

Generally, it can be seen that the PCA window comparison method ranks the effects of more, global or common-mode events higher than activities locally concentrated on parts of the grid. The benefit of this proposed method thus stems from its ability to amplify creeping, common-mode events that may be undetected by the VTA method (as shown in Fig. 4.29). In turn, a benefit of the VTA technique is to quickly identify PMUs with strange behaviors by visually observing the large deviations from the rest of the system irrespective of the cause of deviation (that is, measurement errors, locally-based or global-impact disturbances).

VTA and PCA Analysis on Synthetic Dataset

The event detection techniques were also applied on the 90-seconds, synthetic voltage measurements for the case scenarios in Section 4.3.

1. Generator trip after 18 seconds. Fig. 4.31 (a), (b) and (c) are the voltages, VTA and PCA results respectively for all selected 99 PMUs.
2. Generator trip, switched-in shunt and noise. Here, the amount of noise signal in PMUs 1 and 10 was increased to 45 decibels for 15 seconds. The results from the analysis are shown in Fig. 4.32.

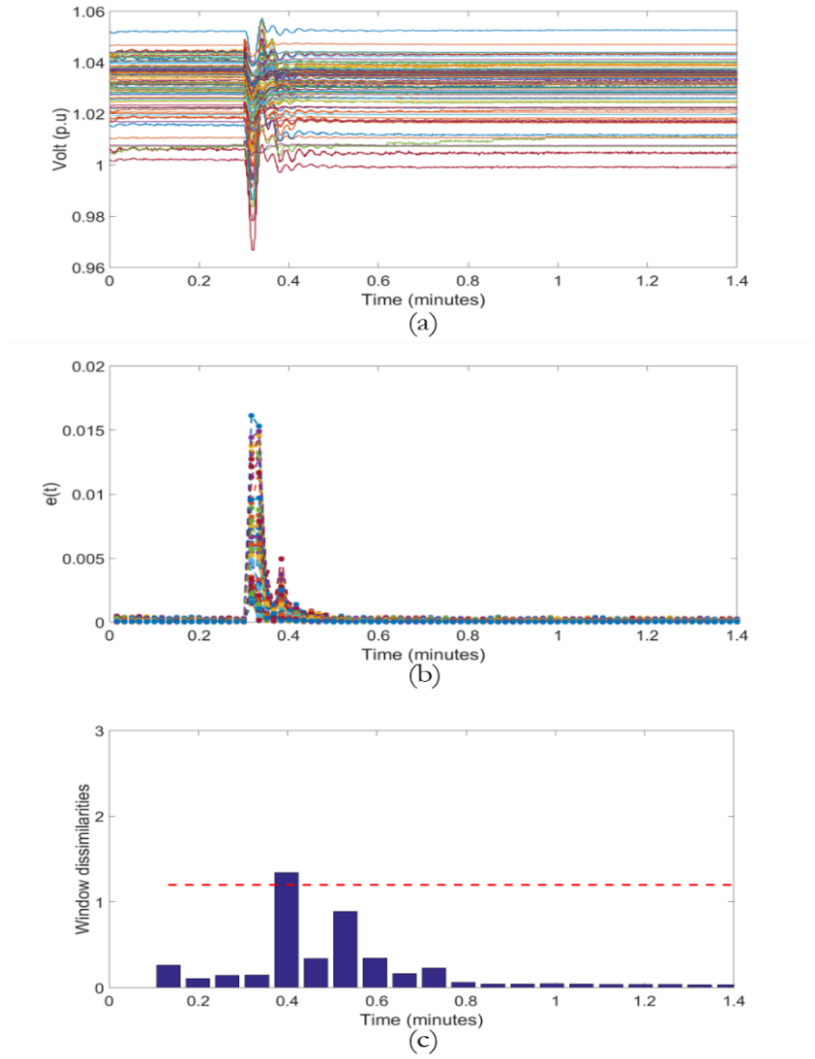


Figure 4.31. Voltage, e-values and system time-window disparities for generator trip

The generator trip event is captured by virtually all 99 PMUs when a maximum, distinct *e*-value of 0.016 in Fig. 4.31 (b), which also masks out other time windows of PMU variations in the order of 10^{-3}, is observed for the time duration. It is also observed to violate the preset threshold in Fig. 4.31 (c) when the properties of that time segment are in discordance with a previous, event-less, system window. In the event of local events, such as a switched-in shunt at 0.5-mins (or 30-sec)

in Fig. 4.32 (b), the disturbance is, however, mostly observed in fewer PMUs. The persistent data errors in PMUs 1 and 10, due to increased measurement noise levels, are detected by the extraordinary and consistent deviations in *e*-values observed in the two noisy PMUs.

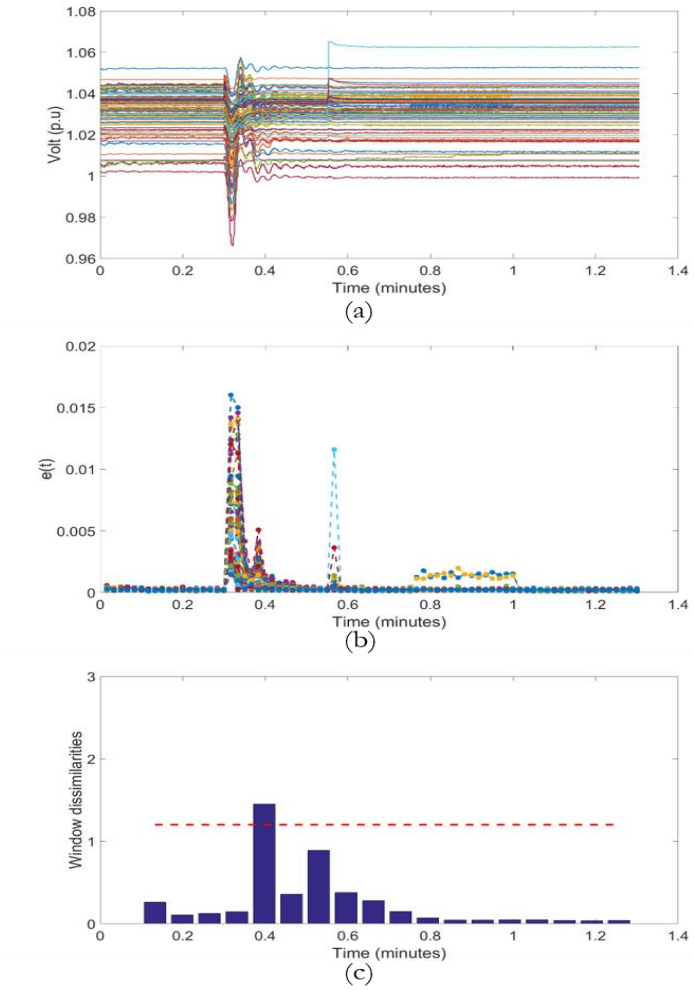


Figure 4.32. Voltage, e-values and system time-window disparities for generator trip, switched-in shunt and noise

4.4.3 Steady-state Oscillation Analysis

As a result of transients and the constantly changing nature of the grid, performing a small-signal stability analysis will often unravel ambient, low-frequency disturbance signals (also known as modes) present in the system [25, 62]. By applying any of the available small-signal stability analysis techniques in the literature [63, 65, 68] using a moving-window approach on PMU datasets, the strength of these sinister, low-frequency disturbances can be estimated. Beyond a pre-defined threshold, threatening activities or events can be detected in the system.

Fig. 4.33 shows real, 1-minute frequency measurements within the time of generator tripping.

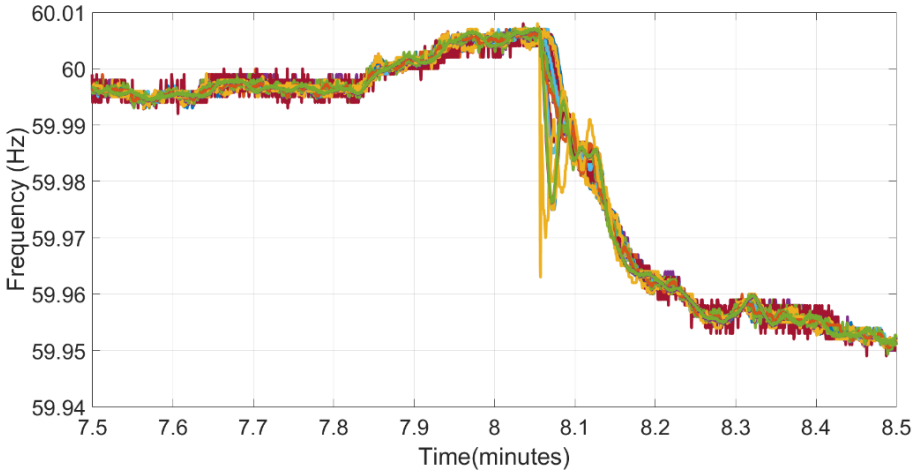


Figure 4.33. 1-min, real frequency measurements around the time of generator trip

As described further in Chapter 7 and appendix E, the method of modal analysis, and which has been used in this work, aims to determine the damping factor (σ_j), frequency (ω_j) and mode shape (consisting of a magnitude (A_j) and phase (ϕ_j)) of all j^{th} low-frequency modes in a signal.

Fig. 4.34 (a), (b) and (c) show the frequency, magnitude and damping factors of the detected modes respectively when modal analysis was performed on the frequency dataset in Fig. 4.33. The configuration of the moving-window which has been used to analyze the dataset has a time window size and step of 6-second, and 3-second respectively.

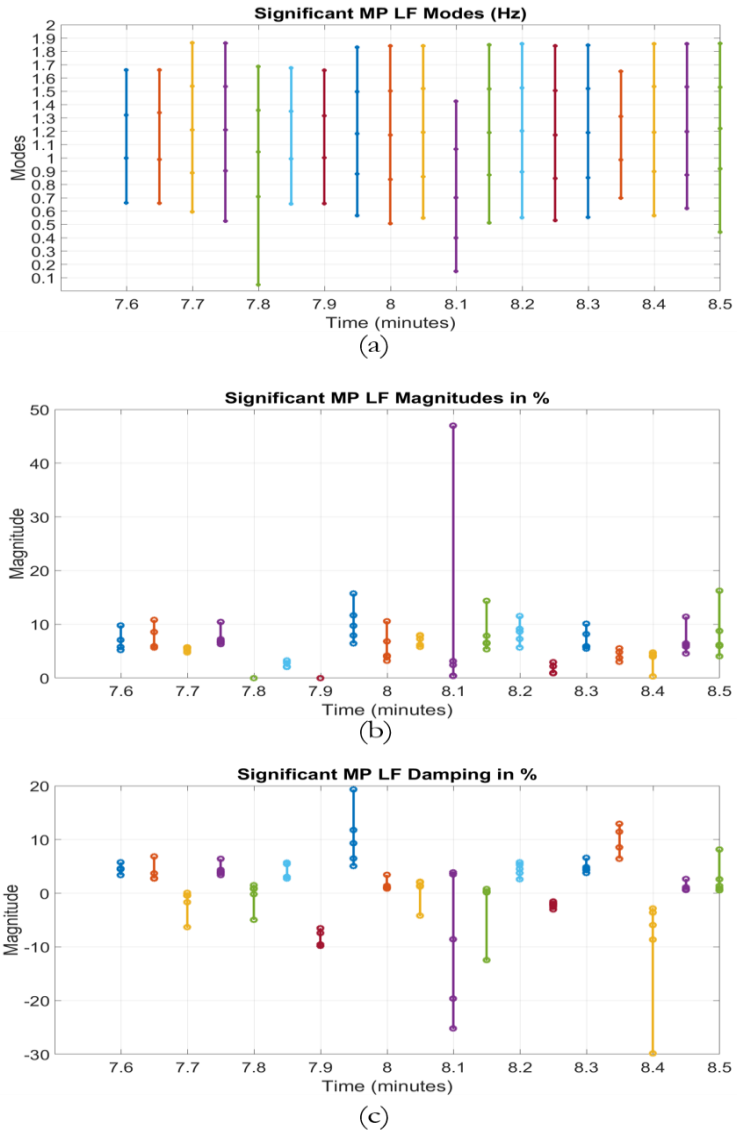


Figure 4.34. Real data: frequency, amplitude and damping factor of observed modes

Time stamps, such as 7.8-mins and 7.9-mins in Fig. 4.34(b) , during which zero magnitudes is reported for all modes are attributed to the computational issues associated with low-ranked matrix observed in the dataset for that time window. However, this does not affect the accuracy of the analysis performed at other time windows. With respect to other time windows, the generator trip, transient disturbance is immediately flagged given the mix of low frequency signal modes observed at 8.1-minutes: (a) 0.14-Hz with 47% amplitude and 3.9% damping, (b) 0.4-Hz with 3% amplitude and -25% damping, and (c) 0.7-Hz with 0.4% amplitude, and -19% damping. Apparently due to good system damping, these modes are observed to quickly decay (i.e. 0.4-Hz and 0.7-Hz) or completely die out (i.e. 0.14-Hz) in subsequent time-windows. While frequency modes above 1.0 Hz can be attributed to the effects of local generator control oscillations, excitation and DC controls [27] , electromechanical inter-area or local oscillations (between 0.15 – 1.0 Hz) are usually of more interest since they are mostly associated with system disturbances. Another observation made in Fig. 4.34 is the consistent appearance of frequency modes of 0.5-0.6 Hz and 0.8-0.9 Hz in the system. As part of an interconnected system, we conjecture that these modes are the ambient electromechanical modes present in the system from which the real PMU dataset used in this study was obtained [109]. These ambient modes will often be caused by the random input variations (e.g. constantly-changing load) in the system. It is important that the system is sufficiently-damped to prevent very large or forced excitations (e.g. during the generator trip) of these ambient modes to threatening levels that could affect the stability of the grid.

4.5 Summary

In this chapter, some of the features of industry-grade PMU data have been discussed to enable the creation of realistic synthetic measurements. The relevance of outlier measurements and other

data errors, in order to achieve true realism in synthetic datasets is elaborated as observed by the statistics obtained from the real dataset. Two moving-window techniques have been proposed to search and discriminate between data errors and actual system events, while ambient, system oscillation modes were shown to be an inherent characteristic of any power grid.

A selection of 99 measurements, out of 2,000 available bus measurements from the synthetic network, have been used for the analysis in this work. However, we acknowledge the possible existence of a better criteria for selecting a subset of signals for analysis (e.g. increasing the number of PMU measurements or selecting measurements from specific hotspot grid locations where these modes are concentrated) which may have provided a better observation of the full strength of the system ambient modes. Nevertheless, all the synthetic dataset-related analysis have been based on the fact that only a limited set of measurements, on which data analytics is performed, is truly available to operators or control centers of real, large-scale grids.

5 EVALUATING PMU TIME-SERIES DATA FOR ERRORS*

A distributed, error analysis technique is used to evaluate data segments of time-series measurements for errors using bus data obtained from a large-scale system. Based on observations made on the unique data patterns of time-related issues, a similarity-based method is proposed to detect timing errors in PMU data measurements.

5.1 Local Outlier Factor

Given a dataset $\boldsymbol{X}$, composed of different time series measurements obtained from n number of PMU nodes, the aim is to be able to identify inconsistent data segments within any of the n measurements.

$$\boldsymbol{X} = \{\boldsymbol{x_1^t}, \boldsymbol{x_{2,}^t} \ldots \boldsymbol{x_n^t}\};\ \boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots x_{ip}\} \tag{1}$$

$\boldsymbol{x_i}$ is the data time series obtained from the i^{th} PMU device, and consists of p data points.

The region of influence of power system events could be negligible, reflecting only in measurements obtained from PMUs located within a local geographical area. Spatio-temporal correlations of PMU measurements, and key features distinguishing bad data from good measurements (containing event time points) were discussed in [79]. A local outlier factor (LOF) method was implemented in [79, 110] to compute degrees of correlation of all measurements relative to a PMU dataset, from which computed error metric were used to identify data measurements considered to deviate from the dataset.

The LOF technique is an unsupervised outlier detection algorithm which is based on the density of the k-nearest neighborhood of each object [104, 111]. By comparing the relative densities of

* Part of this section is reprinted with permission from " PMU Time Error Detection Using Second-Order Phase Angle Derivative Measurements" by I. Idehen and T.J. Overbye , Feb. 2019 IEEE Texas Power and Energy Conference (TPEC), ©2019 IEEE, with permission from IEEE

each neighbor, it is able to detect an outlying object in the set. A summary of the procedures for computing LOF values are given in appendix C.

In this work, a windowing technique which searches for erroneous data segments across all time-series measurements obtained from a large-scale system is implemented. It is observed that in a large system where buses are separated by large geographical and electrical distances, wide-area variations in bus signal trends during system events can mask out bad data due to wrongly-computed LOF values. In addition, it is observed that a direct implementation of the LOF method in a large system is computationally intensive since the k-nearest neighborhood of a measurement can be very large.

5.2 Distributed Local Outlier Factor

The method of computing LOF assigns a metric representing the extent of deviation of a given measurement from a dataset. In large-scale power grids where the degree of signal correlations often differ among regions (e.g., due to local effect of voltage), a central computation of LOF values using all grid signals in one single dataset could mis-represent the true quality of bus measurements.

5.2.1 Illustrating Example (Using contingency case label- 2,000bus (Case 1)).

Fig. 5.1 shows eleven locations in the grid from which voltage measurements were obtained after the contingency outage of a 230-kV line in the 2,000-bus grid. The simulation was run for 3-seconds, and the report rate of all PMUs set to 30 samples per second.

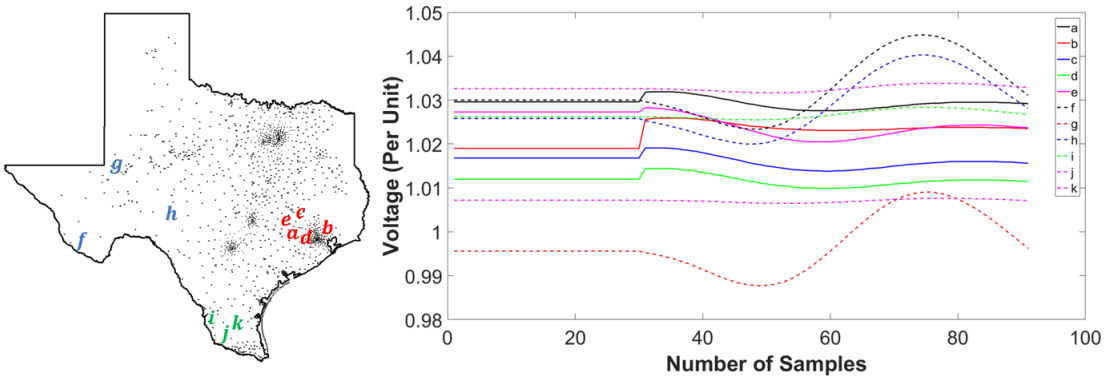


Figure 5.1 3-second voltage measurement

Using all signals in an LOF-based data error analysis, the computed error metrics for the eleven different bus signals are given in Table 5.1.

Table 5.1 LOF results for all 11 signals

Bus	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***
LOF	1.015	0.982	1.015	1.015	0.985	0.987	0.982	0.984	1.015	1.015	1.015

Given different bus trend signals with varying correlations, computed LOF values are observed to vary among different locations even for the same event. Considering 2,000 grid signals, a wider range of LOF variation obscures the true presence of measurement errors. In addition, the runtime for computing the LOF value for each of the bus measurement can be prohibitively high.

To address these limitations, a distributed computation method, which takes into consideration local signal variations in the wide-area network is proposed. Table 5.2. shows computed LOFs when the signals are aggregated into two different clustering formations: A - $\{a, b, c, d, e\}$,$\{f, g, h\}$, $\{i, j, k\}$; and B - $\{b, e\}$, $\{a, c, d\}$, $\{f, g, h\}$, $\{i, j, k\}$ respectively.

Table 5.2 LOF results for all 11 signals using two sets of clustering

Bus	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***
LOF-A	1.045	0.945	1.045	1.045	0.945	1.0	1.0	1.0	1.0	1.0	1.0
LOF-B	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

The flowchart which has been developed for the distributed LOF computation of measurements obtained from a synthetic 2000-bus system is illustrated in Fig. 5.2.

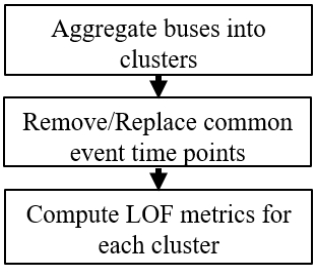


Figure 5.2 Wide-area search for data errors

5.2.2 Cluster Formation for Error Identification

The method used to partition the system into smaller and local groups is governed by the concept of voltage control areas [112] - they consist of groups of buses exhibiting strong electrical coupling among each other. In this work, augmented local bus information – bus geographical location and derived information from bus voltage measurement – are used to determine these bus clusters. Augmenting bus geography with voltage information ensures that generated bus clusters are able to track the dynamic response of the system rather than a sole use of fixed, bus geography coordinates (longitude and latitude) which are devoid of any electrical information.

Extraction of bus voltage information

Due to its simplicity of use in large data sets, PCA technique has been used to extract the major underlying dynamics in a dataset. In an earlier Section 4.3.3, a summary of PCA technique had

been provided. To validate the information contents provided by the principal component vectors, PCA is performed on the data set, however with noise signals injected at buses 1 and 2. The generator outage event is at bus 1506. Fig. 5.3 shows plots of the first eight principal vector components, and the corresponding eigenvalues, while Table 5.3 gives the variance percentage for each of the components.

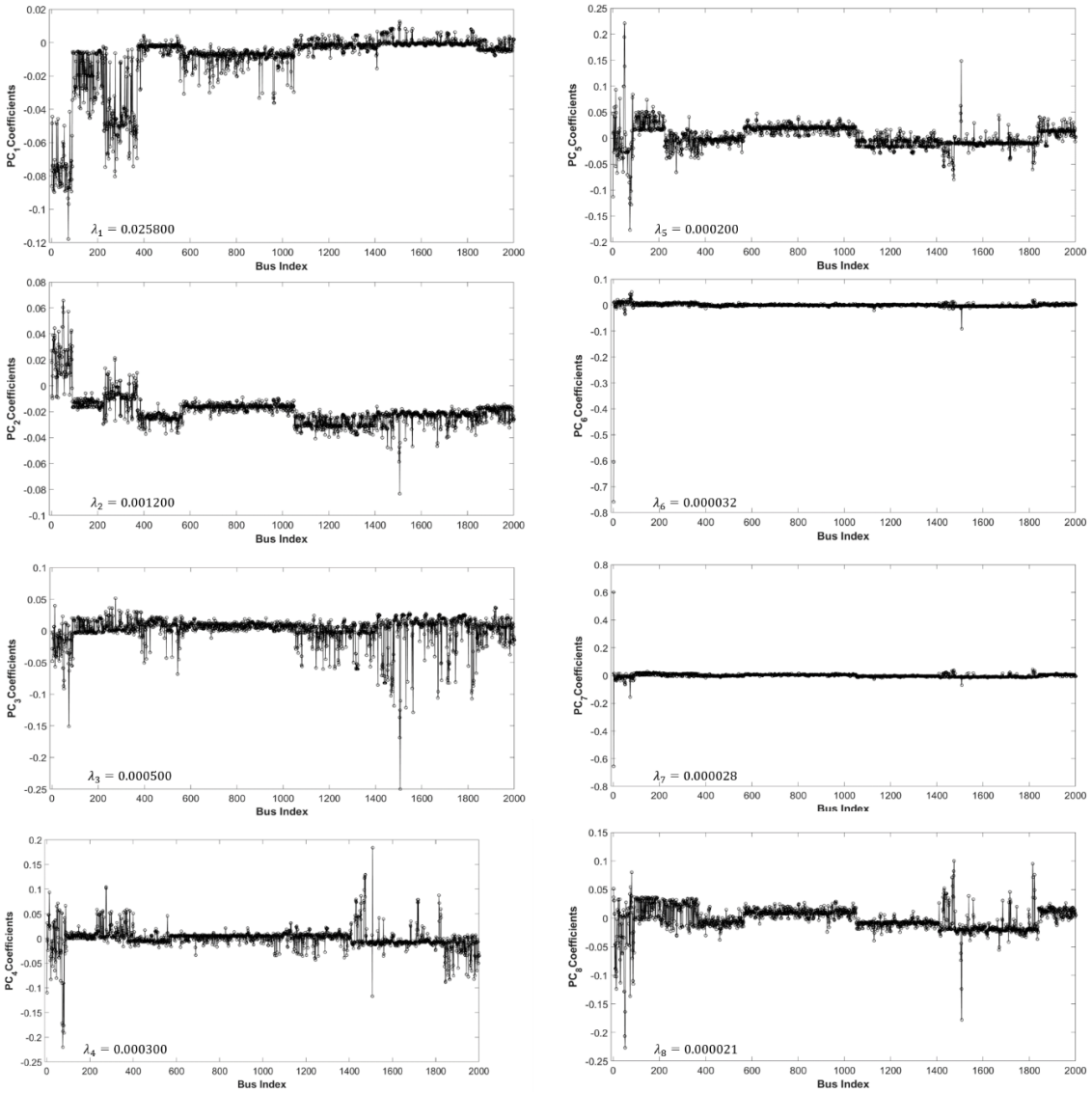


Figure 5.3 First eight principal component vectors

Table 5.3 Variance percentage of principal components - $(\boldsymbol{PC_i}/\sum \boldsymbol{PC_i}) \times \mathbf{100}\%$

Principal Component	$\boldsymbol{PC_1}$	$\boldsymbol{PC_2}$	$\boldsymbol{PC_3}$	$\boldsymbol{PC_4}$	$\boldsymbol{PC_5}$	$\boldsymbol{PC_6}$	$\boldsymbol{PC_7}$	$\boldsymbol{PC_8}$
Variance %	91.880	4.270	1.780	1.068	0.712	0.114	0.0997	0.075

As expected, the variance percentage for $\boldsymbol{PC_1}$ (i.e., 91.88%) is observed to be dominant since it bears most of the voltage information in the system. Beyond the first 3/4 dimensions, other components can be ignored since they contribute minimal or no system information. The extent of activity of any i^{th} bus is indicated by the absolute value of its coefficient in the component vector, and it is observed that as more redundant components are considered, the effect of system noise and data errors become dominant. This is indicated by the high level of activity observed at the buses of noise injections - vector coefficients 1 and 2 - in $\boldsymbol{PC_6}$ and $\boldsymbol{PC_7}$.

Integrating bus position and voltage information

The PCA results indicate that relevant system information is captured by only few, significant vector components which are identified by the magnitude of the corresponding eigenvalues. Based on this reason, the respective voltage information obtained for each bus is provided by the bus index of the significant principal components. A hybrid, data-driven approach implemented for the purpose of clustering, thus augments every bus geography information with its principal vector coefficient(s).

$$\boldsymbol{X}_i = \left[Geo_{lat,i}, Geo_{long,i}, \boldsymbol{PC}_{1,i}, \ldots \boldsymbol{PC}_{n,i}\right] \quad (2)$$

$\boldsymbol{X}_i$ is an hybrid vector for the i^{th} bus, which entries are made up of latitude and longitude coordinates, $Geo_{lat,i}$ and $Geo_{long,i}$, and bus entries in the n significant principal components $\boldsymbol{PC}_{1,i}, \ldots \boldsymbol{PC}_{n,i}$.

Clustering

The aggregation of the bus vectors in (3) is carried out using *k*-means algorithm, which is proven to be a simple and fast technique for generating clusters [113]. The *k*-means algorithm begins with an arbitrary assignment of initial cluster centers (centroids) after which object re-assignment and centroid updates are performed based on a greedy minimization of a sum of squared errors, *SSE*. A prior selection of *K* number of clusters and initializing centroids are some of the limitations of the algorithm. In this work, the initializing buses are selected such that they spread across the different regions of the grid. Different cluster sizes of five and twenty have been chosen to test the performance of the data error detection process. The pseudo algorithm for the clustering procedure is illustrated in Fig. 5.4.

Perform PCA on dataset
Step 1:
for i = 1 to $TotalBuses$
$x_{1,i}$ =Cos($Geo_{long,i}$) ×Cos($Geo_{lat,i}$)
$x_{2,i}$ =Sin($Geo_{long,i}$) ×Cos($Geo_{lat,i}$)
$x_{3,i} = \boldsymbol{PC}_{1,i}$;
$X_i = [x_{1,i}, x_{2,i}, x_{3,i}$
end
Dataset, $\Gamma = [X_1, X_2, .. X_{TotalBuses}]$
Step 2: Pre-processing: Normalize Γ to zero mean and unit variance
Step 3: Initialize area centers and cluster

Figure 5.4 Augmented voltage clustering

Identify/Replace Event Points

A pre-processing step prior to the detection of bus or PMU locations reporting data errors is to identify and isolate actual system disturbance data. The method implemented in this work further

goes to replace the disturbance or event data points with a value based on previous signal trend. The pseudo-code for this process is illustrated in Fig. 5.5.

```
Identify
for  i = 1 to k-areas
      Select all nodes in i^th area ( k^i-buses)
   for  j = 1 to length( k^i-buses)
       Event points for measurement from node_j → EvP_{j,i}
   end
      Common event points  EvP_i = ∩ {EvP_{j,i}}
end
```

Figure 5.5 Pseudo-code for event data point detection and replacement

A variance-based, event-detection technique [34] applied to a sliding moving-window was used to extract event time points from voltage magnitude measurements. A time point $V(n)$ is flagged as an event data point if a computed variance, $v(n)$ exceeded a threshold value.

$$v(n) = \frac{1}{L} \sum_{l=(n-L+1)}^{n} (V(l) - V_{\mu}(n))^2 \tag{3}$$

L is the window length over which $v(n)$ and a mean value, $V_{\mu}(n)$ are computed.

5.2.3 Check for Data Errors

A distributed LOF computation is performed cluster-wise on the grid. Using a sliding window, defined by τ- width and τ_s-sliding time, the bus signals are individually assessed for errors by computing error metric for m segments in the time-series data.

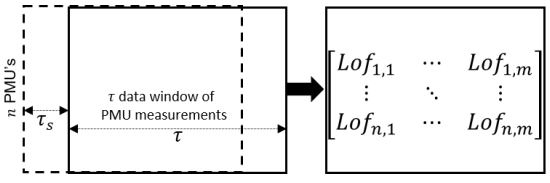


Figure 5.6 Window technique for assessing error level in data segments

Note: A phasor measurement consists of a magnitude and angle component. Because PMU errors affect either or both components, it is important to specify the measurement component being assessed for errors. If the focus is magnitude (or angle), the input data to the LOF routine is phasor magnitude (or ROCOF). Absolute values of voltage angles do not convey much information, and cannot be used to adequately monitor PMU time errors. On the other hand, ROCOF measurements provide a good indication of voltage angle dynamics with an added ability of being able to detect small and sudden changes in voltage angles. Consequently, ROCOF qualifies as a good parameter to monitor time errors in PMU devices.

Case Study: 2,000-bus (Case 2)

A 10-second simulation is carried out on the 2,000 bus network during which one of the 115-kV transmission lines is disconnected after 3 seconds. A data error analysis is carried out on the voltage measurements obtained from the system. Considering only the contingency event (a 115-kV line outage), Table 5.4 gives a summary of the computed error metrics for different system configurations.

Table 5.4 Summary of computed phasor magnitude LOFs (event only)

# Control areas /clusters	Event points removed?	Highest LOF(s)	Bus Index
1	Y	[16.69,8.355,8.358,8.355]	[6276,4174,6128,6015]
	N	[16.69,8.355,8.358,8.355]	[6276,4174,6128,6015]
5	Y	8.360	6015, 6128
	N	8.360	6015, 6128
20	Y	1.000	ALL
	N	8.360	6128
1	Y *	5.060	6276

* Event time points removed separately for individual bus measurements

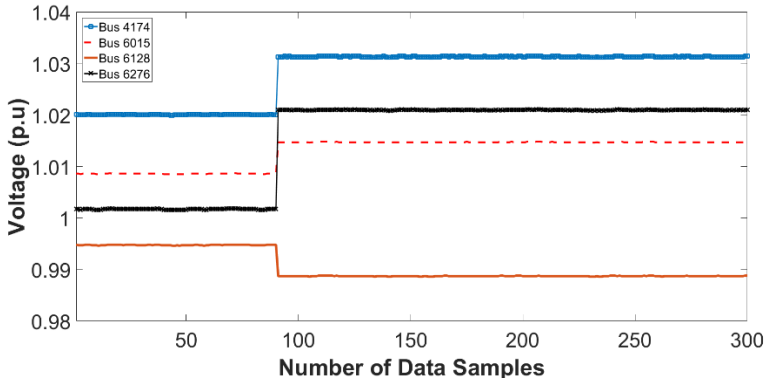


Figure 5.7 Voltage magnitude measurements at all 2,000 buses and other selected buses

According to the direct LOF computation method, false data error notifications are observed at buses (IDs: 6276, 4174, 6128, 6015) where the event is most significant, even though the disturbance has little impact on the system as observed by the measurements in Fig. 5.7. Regardless of the removal of event points, LOF analysis is not able to distinguish the event from instances of data errors, and thus mis-identifies event buses as locations with the largest system LOFs. The proposed distributed method applied to a 5-cluster system aggregates bus IDs 4174 and 6276 into a cluster where a high voltage correlation results in smaller LOF values. However, a false identification of bus IDs 6015 and 6128 with LOF values 8.36 is due to both buses belonging to separate clusters with disparities in their voltage patterns. Using a 20-cluster system, more groups containing buses with similar, dynamic voltage response are generated. Fig. 5.8 shows the re-allocation of the event buses to cluster 2, 5 and 17 such that computed LOF value is 1.0 at all buses, thus isolating the impact of the event.

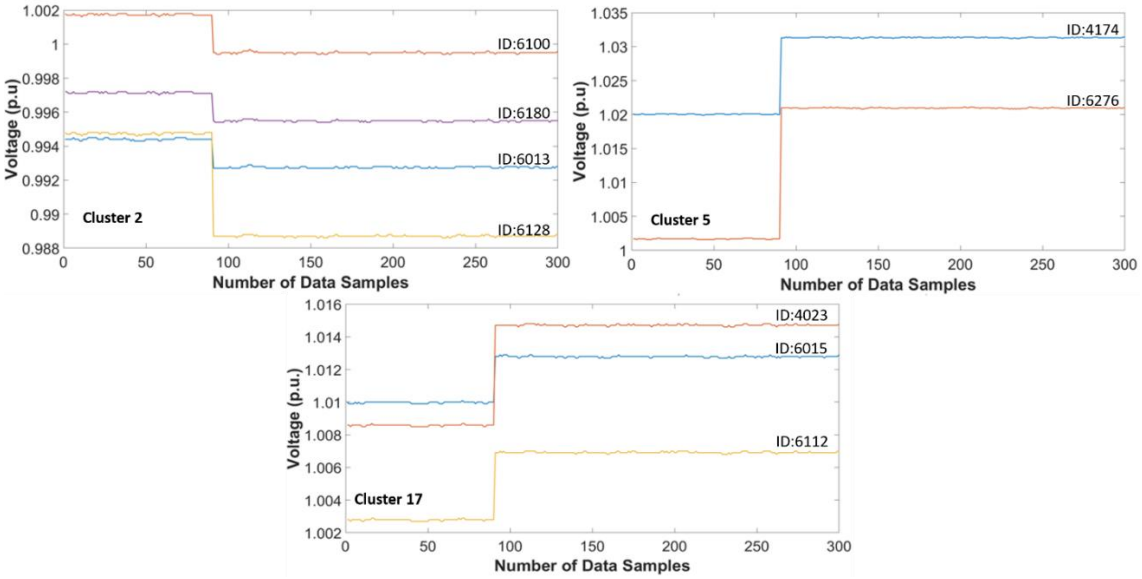


Figure 5.8 Event buses re-distributed to three clusters

Error 1: In this scenario, 45-dB of noise signal is injected in the phasor voltage measurements reported by the PMU located at substation ID 1250. Table 5.5 provides a summary of the computed LOFs.

Table 5.5 Summary of computed phasor magnitude LOFs (event and noise error)

# Control areas /clusters	Event points removed?	Highest LOF(s)	Bus Index
1	Y	[16.69,8.355,8.358,8.355]	[6276,4174,6128,6015]
	N	[16.69,8.355,8.358,8.355]	[6276,4174,6128,6015]
5	Y	[8.358,8.358,2.941,3.263,2.941]	[6015,6128,8158,8159,8160]
	N	[8.358,8.358,2.941,3.263,2.941]	[6015,6128,8158,8159,8160]
20	Y	[2.941,3.263,2.941]	[8158,8159,8160]**
	N	[8.360,2.941,3.263,2.941]	[6128,8158,8159,8160]
1	Y *	5.060	6276

* Event time points removed for individual bus measurements, ** PMU 1250 reports measurements for buses 8158-8160

Based on the large LOF values observed at the event buses, the direct LOF computation mis-identifies these locations as sources of erroneous measurements even though data errors are reported by another PMU set of phasor measurements. A distributed error analysis on the 5-cluster system identifies all three erroneous measurements, with LOF values 2.941, 3.263 and 2.941, in addition to two event bus measurements (bus IDs 6015 and 6128). Finally, with a 20-cluster system, all event buses are isolated, and thus correctly identifying only the measurements with true data errors. Table 5.6 shows bus allocations in four groups within the 20-cluster system.

Table 5.6 Cluster formation

Cluster 2	[6013,6100,6128,6180]
Cluster 5	[4174,6276]
Cluster 16	[8158,8159,8160, …]
Cluster 17	[4023, 6015, 6112]

Error 2: In this scenario, time errors were incorporated into the phasor measurements reported by the PMUs located at substation IDs 4, 538 and 764. Time error details and the voltage angles for the affected measurements are shown in Table 5.7 and Fig. 5.9 respectively.

Table 5.7 Data errors

SS/PMU ID	**# buses**	**Error Type (T)**	**Error**	**Parameters**	**Duration**
4	2	T-2	Clock drift	Skew: 1-μs	7-sec
538	1	T-2	Clock drift	Skew: 0.5-μs	10-sec
764	1	T-1	Int. GPS	Skew: 10-μs, 5 instances	1-sec/instance

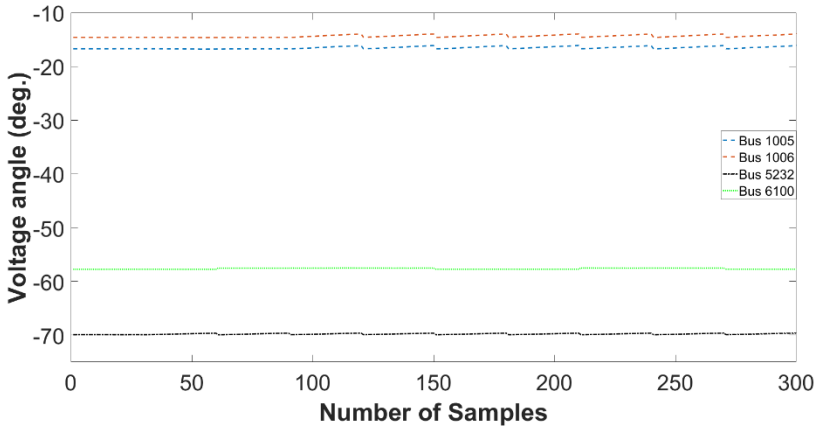


Figure 5.9 Voltage angle measurements at all 2,000 buses, and time-skews due to time errors

Similar computations were carried out in a data error analysis using the phasor angle components of bus measurements. Tables 5.8 and 5.9 provide a summary of the computed LOFs when considering only the event, and then the inclusion of the noise respectively.

Table 5.8 Summary of computed phasor angle LOFs (event only)

# Control areas /clusters	Event points removed?	Highest LOF(s)	Bus Index
1	Y	[58]	[6276]
	N	[58]	[6276]
5	Y	1.000	ALL
20	Y	1.000	ALL

Table 5.9 Summary of computed phasor angle LOFs (event and time errors)

# Control areas /clusters	Event points removed?	Highest LOF(s)	Bus Index
1	Y	[522,522,325,200,58]	[1005,1006,5232,6100,6276]
	N	[522,522,325,200,58]	[1005,1006,5232,6100,6276]
20	Y	[267,192,192,41]	[5232,1005,1006,6100]*

* PMU 4 reports measurements for bus IDs 1005 and1006; PMU 538 for 5232; and PMU 764 for 6100

Regarding the error analysis carried out on voltage angle measurements, ROCOF data has been used for the reasons stated in Section 4.2.3. As expected, direct LOF computation identifies event bus 6276 where a large change in the voltage angle occurred during the event only analysis. By dividing into bus regions and applying distributed LOFs, both 5-cluster and 20-cluster systems isolate the dynamic event and returns a uniform, LOF value of 1.0 for all buses. Inclusion of time-error events causes direct LOF to identify both buses with measurement errors and event bus as large LOF locations. The observed large error metrics (522, 522, 325, 200, and 58) are due to the pre-processing, normalization step carried out on ROCOF data prior to error analysis. Using the bus allocations in Fig. 5.6 which re-allocates event buses into groups, the 20-cluster system is able to correctly identify the error measurements, and corresponding faulty PMUs.

Error 3: Here, the data errors in both previous error scenarios (error 1 and 2) were integrated into the original case event measurements i.e. noise at PMU 1250, and time errors at PMUs 4,538 and 764.

The effects of both data errors are observed in the ROCOF data, and thus used as the input data source in the error analysis. Table 5.10 summarizes the results.

Table 5.10 Summary of computed phasor angle LOFs (event, noise and time errors)

# Control areas /clusters	Event points removed?	Highest LOF(s)	Bus Index
1	Y	[522,522,325,200,58,248,253,291]	[1005,1006,5232,6100,6276,8158,8159,8160]
	N	[522,522,325,200,58,248,253,291]	[1005,1006,5232,6100,6276,8158,8159,8160]
20	Y	[297,192,192,41,248,253,132]	[5232,1005,1006,6100,8158,8159,8160]

Similar to the other scenarios, the distributed method is able to detect all bus/PMU locations where data errors were reported from.

Note: The effectiveness of distributed error LOF analysis has so far showed that grid dynamics can be contained within local regions prior to data error search in order to reduce instances of false identifications. However, it is critical to note the importance of clustering dynamics according to the most significant, few principal component vectors of the reduced dataset dimension. Using more component vectors introduces error to the clustering process which then affects data error analysis.

Table 5.11 shows error analysis computation times for the single area, 5-cluster, and 20-cluster system configuration which was carried out on 10-second data on a 3.6 GHz processor, windows-based system.

Table 5.11 LOF execution time for different system configurations

# Control areas /clusters	1	5	20
Time (sec)	284	70	21

A notable processing time for the single area can be attributed largely to the several executions of some of the LOF analysis steps carried out within the large *k*-neighborhood of each of the 2,000 measurements. This is not the case with the clustered configurations, and it is assumed that with parallel computation, the running time for clustred systems can be much reduced.

Windowing Scheme

A distributed LOF computation for the windowing scheme in Fig. 5.6 is applied on the 10-sec data obtained for error cases 1 and 3, using time woindow and time step of 1-sec and 0.5-sec respectively. The derived error values for each segment in all 2,000 measurements for both data error cases is shown in Fig. 5.10 and 5.11.

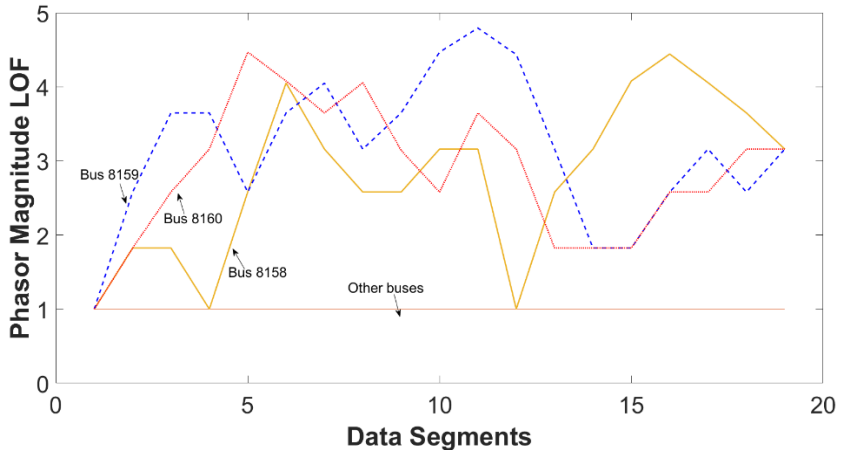


Figure 5.10 Data segment errors in all 2,000 voltage magnitude measurements for error case #1

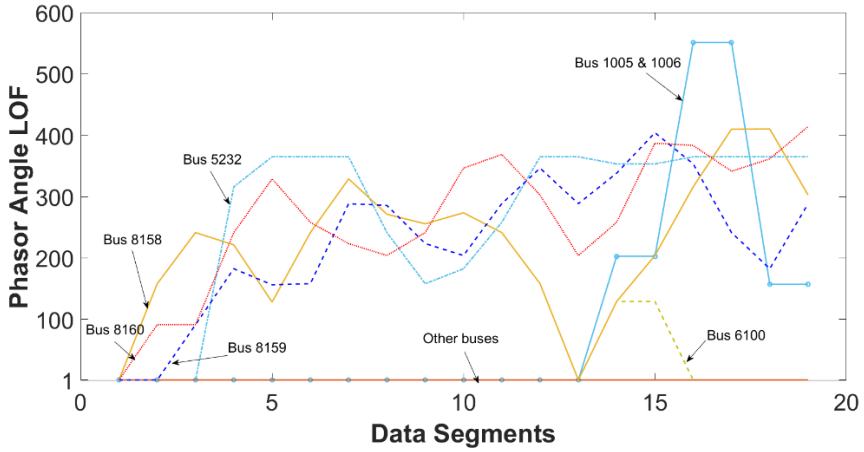


Figure 5.11 Data segment errors in all 2,000 voltage angle measurements for error case #3

5.3 Similarity-Based PMU Time Error Detection

As part of a data post-processing stage, and given the unique data patterns created on voltage angle measurements due to time errors, a method to detect synchronization issues in PMU data is proposed [73].

5.3.1 Similarity matching – Illustrating Example

Given two data sequences:

$$\boldsymbol{X} = \{1,1,1,3,3,3,3,3,3,3,1,1,1,1,1\}$$
$$\boldsymbol{Y} = \{1,1,1,1,1,3,3,3,3,3,3,3,1,1,1\}$$

A dis-similarity value (ρ) between the sequences can be computed using the L_n norm [114].

$$\rho = \sqrt[n]{\sum_{i}^{m}(x_i - y_i)^n} \tag{4}$$
$$x_i \in \boldsymbol{X} \in R^m \text{ and } y_i \in \boldsymbol{Y} \in R^m$$

When n = 2, the computed dissimilarity value (Euclidean distance) is 5.66, which indicates a disparity between both sequences. However, careful observation reveals $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same element values though shifted in time. The Euclidean measure fails to detect this time alignment

issue, and results in a similarity value that is non-intuitive. Hence, the need for a metric which captures time shift information among data sequences. Furthermore, given a pattern sequence, $\boldsymbol{A} \in R^{\alpha}$, and a longer data sequence $\boldsymbol{B} \in R^{\beta} = \{n_1, \dots, n_\beta\}$ such that $\beta > \alpha$, the goal is to determine the similarity level between $\boldsymbol{A}$ and $\boldsymbol{B}$ of mis-matched lengths.

5.3.2 Dynamic Time Warping (DTW)

It is a technique used to temporally wrap a data sequence along its time axis in order to detect similar sequence, and can be used to find the optimal alignment between both sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$, while simultaneously computing the similarity level - a *DTW* distance [115-118]. Considering two sequences, $\boldsymbol{C} = \{c_1, c_2 \dots c_M\}$ and $\mathbf{Q} = \{q_1, q_2 \dots q_N\}$, important terms in the DTW literature are briefly summarized:

1. A cost matrix, $\boldsymbol{P} \in R^{N \times M}$, which is a 2-dimensional matrix, and contains pairwise local cost between data points in $\boldsymbol{C}$ and $\boldsymbol{Q}$**.** The goal is to find an alignment between $\boldsymbol{C}$ and $\boldsymbol{Q}$ having a minimal overall cost. For this work, a pairwise Euclidean distance between data points i and j is chosen as the local cost i.e. $P(i,j) = d(c_i, q_j)$ such that $1 \leq i \leq N$ and $1 \leq j \leq M$.
2. A warping path is a sequence $p = (p_1, \dots p_L)$ with $p_l = (n_l, m_l) \in [1{:}N] \times [1{:}M]$ for $l \in [1{:}L]$. p_l is a node in *p*. It defines an alignment between $\boldsymbol{C}$ and $\boldsymbol{Q}$ subject to some conditions:

   Boundary condition: $p_1 = (1,1)$ and $p_L = (N, M)$. Similar to a stretch on the time axis, it ensures that alignment of the terminals of both sequences.

   Monotonicity condition: $n_1 \leq n_2 \leq \cdots \leq n_L$ and $m_1 \leq m_2 \leq \cdots \leq m_L$. This condition forces the different nodes to be monotonically spaced in time.

   Step size condition: $p_{l+1} - p_l \in \{(1,0), (0,1), (1,1)$ for $l \in [1{:}L-1]$. It constrains the usage of every element in $\boldsymbol{C}$ and $\boldsymbol{Q}$ during the formation of all possible p's in $\boldsymbol{P}$.

3. A total cost $d_p(\boldsymbol{C}, \boldsymbol{Q})$ associated with a warping path, p and is the sum of all local cost measures, d in p. i.e. $d_p(C, Q) = \sum_{l=1}^{L}(dp_l)$ where $dp_l = d(c_{n_l}, q_{m_l})$.

4. An optimal warping path p^* is that path which minimizes the *DTW* distance among all the different possible paths. p^* is computed as $argmin(d_p(C, Q))$, and with a *DTW* distance of $d_{p*}(C, Q)$.

Aligning two data sequences and determining the optimal *DTW* distance requires the construction of an accumulated cost matrix, $\boldsymbol{AP} \in R^{N \times M}$ in a forward direction, followed by the determination of p^* in a backward direction. Both computation steps involve recursive calculations resulting in the use of dynamic programming (DP) algorithms.

The DP algorithm used in obtaining optimal alignment between sequences is given as:

1. Forward Direction

$$\boldsymbol{AP}(i,j) = \boldsymbol{P}(i,j) + min\begin{cases} \boldsymbol{AP}(i,j-1) \\ \boldsymbol{AP}(i-1,j) \\ \boldsymbol{AP}(i-1,j-1) \end{cases}$$

$$\boldsymbol{AP}(0,0) = 0, \boldsymbol{AP}(i,0) = \boldsymbol{AP}(0,j) = \infty; \quad (1 \le i \le M, 1 \le j \le N)$$

Hence, $(C, Q) = \boldsymbol{AP}(M, N)$.

Initialization of the entries in the first row and column is carried out by an additional 0^{th} row and column with values set to arbitrarily large numbers.

2. Backward Direction (OptimalWarpingPath)

Beginning at $p_l = (M, N)$

Determine: $p_{l-1} \coloneqq \begin{cases} (1, m-1), & if\ n = 1 \\ (n-1, 1), & if\ m = 1 \\ argmin\{\boldsymbol{AP}(n-1, m-1), \boldsymbol{AP}(n-1, m), \boldsymbol{AP}(n, m-1)\}, & otherwise \end{cases}$

$m \in [1:M], n \in [1:N]$

Computation of p^* involves starting at the $(M, N)^{th}$ node and recursively tracing back the matrix through the path of least cost until p_1 is reached. An optimal path lying along the diagonal position of the matrix indicates a strong point-to-point correlation of the data points in both sequences.

Using the algorithm of the forward direction, the developed accumulated cost matrix for both sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ are given in Fig. 5.12, where $N = M = 15$.

	Y		1	1	1	1	1	3	3	3	3	3	3	3	1	1	1
	Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	0	0	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
1	1	999	0	0	0	0	0	4	8	12	16	20	24	28	28	28	28
1	2	999	0	0	0	0	0	4	8	12	16	20	24	28	28	28	28
1	3	999	0	0	0	0	0	4	8	12	16	20	24	28	28	28	28
3	4	999	4	4	4	4	4	0	0	0	0	0	0	0	4	8	12
3	5	999	8	8	8	8	8	0	0	0	0	0	0	0	4	8	12
3	6	999	12	12	12	12	12	0	0	0	0	0	0	0	4	8	12
3	7	999	16	16	16	16	16	0	0	0	0	0	0	0	4	8	12
3	8	999	20	20	20	20	20	0	0	0	0	0	0	0	4	8	12
3	9	999	24	24	24	24	24	0	0	0	0	0	0	0	4	8	12
3	10	999	28	28	28	28	28	0	0	0	0	0	0	0	4	8	12
1	11	999	28	28	28	28	28	4	4	4	4	4	4	4	0	0	0
1	12	999	28	28	28	28	28	8	8	8	8	8	8	8	0	0	0
1	13	999	28	28	28	28	28	12	12	12	12	12	12	12	0	0	0
1	14	999	28	28	28	28	28	16	16	16	16	16	16	16	0	0	0
1	15	999	28	28	28	28	28	20	20	20	20	20	20	20	0	0	0

Figure 5.12 Accumulated cost matrix for $\boldsymbol{X}$ and $\boldsymbol{Y}$

The matrix is read from left-to-right, and top-to-bottom for $\boldsymbol{Y}$ and $\boldsymbol{X}$ respectively. Node (M, N) is located at the right-bottom corner of the matrix, and with a $\boldsymbol{DTW}$ distance of 0. The heavy-shaded area is the region of least cost, while the lightly shaded area is the optimal warping path. Initializing conditions for the matrix have been set to 999 i.e. $\boldsymbol{AP}(i, 0) = \boldsymbol{AP}(0, j) = 999$.

DTW Modification

Multiple low cost nodes observed in different paths indicate exact correlation between values in $\boldsymbol{X}$ and $\boldsymbol{Y}$. At any node, p_i the choice of p_{i-1} in the backward direction algorithm requires that it is the lowest cost node. In this work, an algorithm bias is introduced to ensure that in the event of multiple nodes with equal and minimal costs, the diagonal node is always preferred.

$$p_{i-1} = argmin\{\boldsymbol{AP}(n-1, m-1), \boldsymbol{AP}(n-1, m), \boldsymbol{AP}(n, m-1)$$
$$if\ \{p_{i-1}\} > 1, and\ (n-1, m-1) \in\ p_{i-1}$$
$$Return: node\ (n-1, m-1)$$

Subsequence DTW

It is a special form of DTW which allows finding specific smaller data sequences in a long data stream. If $N \gg M$, one can search for a sub-sequence $Q(a^*: b^*) \coloneqq (q_{a^*}, q_{a^*+1} \dots q_{b^*})$ with $1 \leq a^* \leq b^* \leq$ N which minimizes the *DTW* distance to C over all the possible subsequences of Q. i.e.

$$(a^*, b^*) \coloneqq \underset{(a,b):1\leq a\leq b\leq N}{argmin}(DTW(C, Q(a: b)))$$

A distinguishing feature in the modified algorithm is the relaxation of the previously-specified warping path boundary conditions.

$$\sum_{i=1}^{m} \boldsymbol{P}(i, 1) \text{ for } m \in [1: M] \text{ and } \boldsymbol{AP}(1, n) \coloneqq \boldsymbol{P}(1, m)$$
$$\boldsymbol{AP}(i, 0) = \infty, i \leq i \leq M$$
$$\boldsymbol{AP}(0, j) = 0, i \leq j \leq N$$

Hence, the procedures to obtain b^*, and then $a^* \in [1: M]$ are stated as follows:

1. Search for all *b*'s which minimize $\boldsymbol{AP}(M, :)$ i.e.

$$b^* = \underset{b:[1:N]}{argmin}(\boldsymbol{AP}(M, b))$$

2. *To find* a^*, begin backward recursive search from $p_l = (M, b^*)$. Apply the *OptimalWarpingPath* algorithm such that $p_1 = (a^*, 1)$ for some $l \in [1: L]$

5.3.3 Case Study

Motivated by the similarity-matching technique, the SDTW method is applied in the search for time-based errors in PMU-sourced data.

Prior Step: Event Data Points and Noise in ROCOF Data

Considering that system perturbations and noise could mask out distinct signatures of PMU time-errors, a prior pre-processing step is to identify and either remove or replace the data points corresponding to these events. Given the event-detection technique already used in this work, an additional step was to remove all event-based ROCOF time points, and replace them with a past

window average. Furthermore, noisy ROCOF measurements can be identified by comparing data points with a threshold value. The standard in [119] specifies a steady-state threshold value of 0.01 Hz/s, which does not capture system dynamics (such as load variations). For this work, a dynamic which does not capture threshold of 0.02 Hz/s (similar with [37]) is used. Thus, below this value, ROCOF data points are classified to be noisy.

Prototypes - Clock Delay and Intermittent GPS

$\boldsymbol{P} \in R^M$ is a prototype pattern for a specific time-based error, and $\boldsymbol{T} \in R^N$ is the test ROCOF measurement. M and N are the number of data points in $\boldsymbol{P}$ and $\boldsymbol{T}$ respectively. Error prototypes, $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$, used for the demonstration represent a 5 μs PMU internal clock delay and 135 μs intermittent GPS clock respectively, and are shown in Fig. 5.13. The report rate is 30 samples per second.

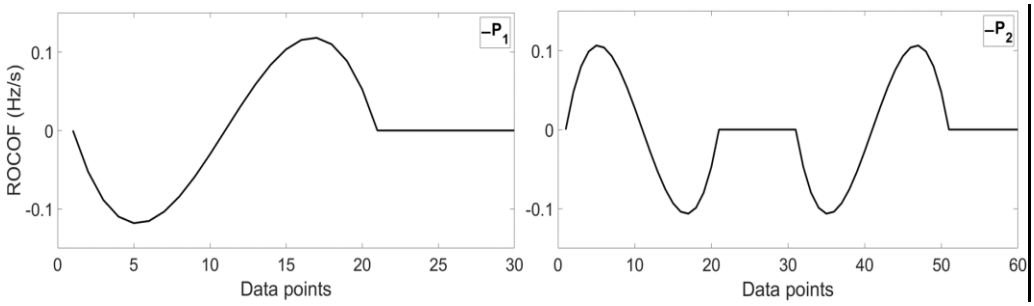


Figure 5.13 Prototype ROCOF patterns for internal clock delay and intermittent GPS signal

Test Data

The algorithm was tested on two different erroneous datasets – time errors for a 3 μs PMU internal clock delay ($\boldsymbol{T_1}$) and 75 μs intermittent GPS clock ($\boldsymbol{T_2}$). The simulation was carried out for 10 seconds, and with a PMU report rate of 30 samples per second. Thus, N was set to a length of 300 data points.

1. Steady state with small random load perturbations

Test ROCOF data $\boldsymbol{T_1}$, obtained from a test bus is shown in Fig. 5.14. The SDTW similarity–based pattern query was separately implemented using $\boldsymbol{P_1}$.

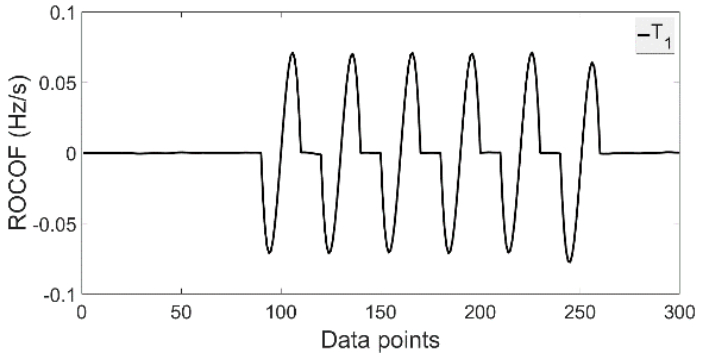
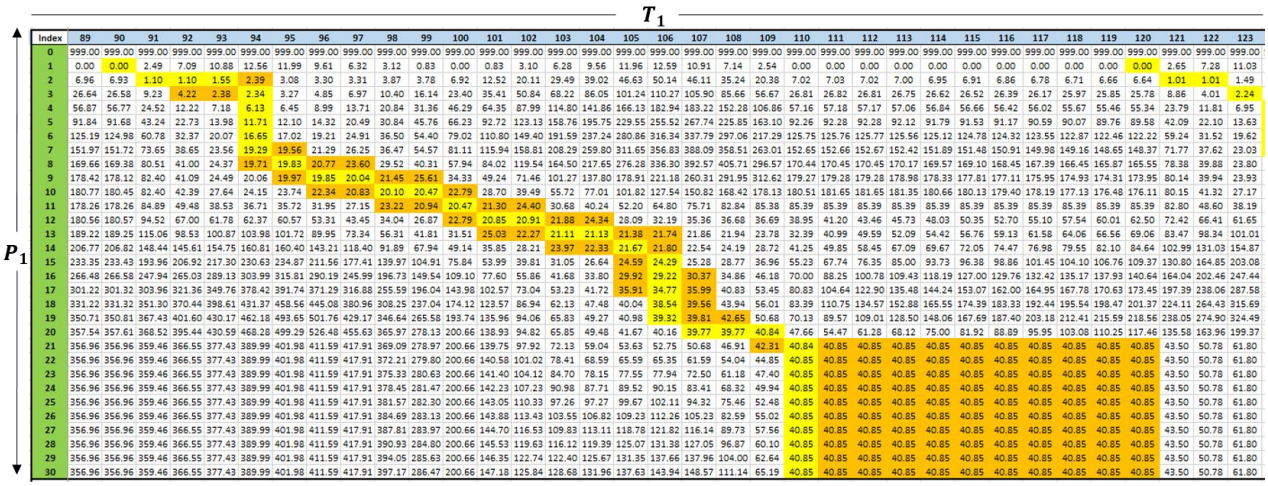


Figure 5.14 Test non-event ROCOF data $\boldsymbol{T_1}$ for case study (1)

In the absence of significant system dynamics and noise, $\boldsymbol{T_1}$ is vividly observed to have a maximum of 0.07 Hz/s when clock delay was simulated between the second and eighth second of the total simulation time.

Fig. 5.15 shows a cut-section of the accumulated cost matrix, $\boldsymbol{AP} \in R^{30\times300}$ generated when the SDTW algorithm was run on $\boldsymbol{T_1}$ in a search for $\boldsymbol{P_1}$.



Index	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123
0	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00
1	0.00	0.00	2.49	7.09	10.88	12.56	11.99	9.61	6.32	3.12	0.83	0.00	0.83	3.10	6.28	9.56	11.96	12.59	10.91	7.14	2.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.65	7.28	11.03
2	6.96	6.93	1.10	1.10	1.55	2.39	3.08	3.30	3.31	3.87	3.78	6.92	12.52	20.11	29.49	39.02	46.63	50.14	46.11	35.24	20.38	7.02	7.03	7.02	7.00	6.95	6.91	6.86	6.78	6.71	6.66	6.64	1.01	1.01	1.49
3	26.64	26.58	9.23	4.22	2.38	2.34	3.27	4.85	6.97	10.40	16.14	23.40	35.41	50.84	68.22	86.05	101.24	110.27	105.90	85.66	56.67	26.81	26.82	26.81	26.75	26.62	26.52	26.39	26.17	25.97	25.85	25.78	8.86	4.01	2.24
4	56.87	56.77	24.52	12.22	7.18	6.13	6.45	8.99	13.71	20.84	31.36	46.29	64.35	87.99	114.80	141.86	166.13	182.94	183.22	152.28	106.86	57.16	57.18	57.17	57.06	56.84	56.66	56.42	56.02	55.67	55.46	55.34	23.79	11.81	6.95
5	91.84	91.68	43.24	22.73	13.98	11.71	12.10	14.32	20.49	30.84	45.76	66.23	92.72	123.13	158.76	195.75	229.55	255.52	267.74	225.85	163.10	92.26	92.28	92.28	92.12	91.79	91.53	91.17	90.59	90.07	89.76	89.58	42.09	22.10	13.63
6	125.19	124.98	60.78	32.37	20.07	16.65	17.02	19.21	24.91	36.50	54.40	79.02	110.80	149.40	191.59	237.24	280.86	316.34	337.79	297.06	217.29	125.75	125.76	125.77	125.56	125.12	124.78	124.32	123.55	122.87	122.46	122.22	59.24	31.52	19.62
7	151.97	151.72	73.65	38.65	23.56	19.29	19.56	21.29	26.25	36.47	54.57	81.11	115.94	158.81	208.29	259.80	311.65	356.83	388.09	358.51	263.01	152.65	152.66	152.67	152.42	151.89	151.48	150.91	149.98	149.16	148.65	148.37	71.77	37.62	23.03
8	169.66	169.38	80.51	41.00	24.37	19.71	19.83	20.77	23.60	29.52	40.31	57.94	84.02	119.54	164.50	217.65	276.28	336.30	392.57	405.71	296.57	170.44	170.45	170.45	170.17	169.57	169.10	168.45	167.39	166.45	165.87	165.55	78.38	39.88	23.80
9	178.42	178.12	82.40	41.09	24.49	20.06	19.97	19.85	20.04	21.45	25.61	34.33	49.24	71.46	101.27	137.80	178.91	221.18	260.31	291.95	312.62	179.27	179.28	179.28	178.98	178.33	177.81	177.11	175.95	174.93	174.31	173.95	80.14	39.94	23.93
10	180.77	180.45	82.40	42.39	27.64	24.15	23.74	22.34	20.83	20.10	20.47	22.79	28.70	39.49	55.72	77.01	101.82	127.54	150.82	168.42	178.13	180.51	181.65	181.65	181.35	180.66	180.13	179.40	178.19	177.13	176.48	176.11	80.15	41.32	27.17
11	178.26	178.26	84.89	49.48	38.53	36.71	35.72	31.95	27.15	23.22	20.94	20.47	21.30	24.40	30.68	40.24	52.20	64.80	75.71	82.84	85.38	85.39	85.39	85.39	85.39	85.39	85.39	85.39	85.39	85.39	85.39	85.39	82.80	48.60	38.19
12	180.56	180.57	94.52	67.00	61.78	62.37	60.57	53.31	43.45	34.04	26.87	22.79	20.85	20.91	21.88	24.34	28.09	32.19	35.36	36.68	36.69	38.95	41.20	43.46	45.73	48.03	50.35	52.70	55.10	57.54	60.01	62.50	72.42	66.41	61.65
13	189.22	189.25	115.06	98.53	100.87	103.98	101.72	89.95	73.34	56.31	41.81	31.51	25.03	22.27	21.11	21.13	21.38	21.74	21.86	21.94	23.78	32.39	40.99	49.59	52.09	54.42	56.76	59.13	61.58	64.06	66.56	69.06	83.47	98.34	101.01
14	206.77	206.82	148.44	145.61	154.75	160.81	160.40	143.21	118.40	91.89	67.94	49.14	35.85	28.21	23.97	22.33	21.67	21.80	22.54	24.19	28.72	41.25	49.85	58.45	67.09	69.67	72.05	74.47	76.98	79.55	82.10	84.64	102.99	131.03	154.87
15	233.35	233.43	193.96	206.92	217.30	230.63	234.87	211.56	177.41	139.97	104.91	75.84	53.99	39.81	31.05	26.64	24.59	24.29	25.28	28.77	36.96	55.23	67.74	76.35	85.00	93.73	96.38	98.86	101.45	104.10	106.76	109.37	130.80	164.85	203.08
16	266.48	266.58	247.94	265.03	289.13	303.99	315.81	290.19	245.99	196.73	149.54	109.10	77.60	55.86	41.68	33.80	29.92	29.22	30.37	34.86	46.18	70.00	88.25	100.78	109.43	118.19	127.00	129.76	132.42	135.17	137.93	140.64	164.04	202.46	247.44
17	301.22	301.32	303.96	321.36	349.76	378.42	391.74	371.29	316.88	255.59	196.04	143.98	102.57	73.04	53.23	41.72	35.91	34.77	35.99	40.83	53.45	80.83	104.64	122.90	135.48	144.24	153.07	162.00	164.95	167.78	170.63	173.45	197.39	238.06	287.58
18	331.22	331.32	351.30	370.44	398.61	431.37	458.56	445.08	380.96	308.25	237.04	174.12	123.57	86.94	62.13	47.48	40.04	38.54	39.56	43.94	56.01	83.39	110.75	134.57	152.88	165.55	174.39	183.33	192.44	195.54	198.47	201.37	224.11	264.43	315.69
19	350.71	350.81	367.43	401.60	430.17	462.18	493.65	501.76	429.17	346.64	265.58	193.74	135.96	94.06	65.83	49.27	40.98	39.32	39.81	42.65	50.68	70.13	89.57	109.01	128.50	148.06	167.69	187.40	203.18	212.41	215.59	218.56	238.05	274.90	324.49
20	357.54	357.61	368.52	395.44	430.59	468.28	499.29	526.48	455.63	365.97	278.13	200.66	138.93	94.82	65.85	49.48	41.67	40.16	39.77	39.77	40.84	47.66	54.47	61.28	68.12	75.00	81.92	88.89	95.95	103.08	110.25	117.46	135.58	163.96	199.37
21	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	369.09	278.97	200.66	139.75	97.92	72.13	59.04	53.63	52.75	50.68	46.91	42.31	40.84	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
22	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	372.21	279.80	200.66	140.58	101.02	78.41	68.59	65.59	65.35	61.59	54.04	44.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
23	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	375.33	280.63	200.66	141.40	104.12	84.70	78.15	77.55	77.94	72.50	61.18	47.40	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
24	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	378.45	281.47	200.66	142.23	107.23	90.98	87.71	89.52	90.15	83.41	68.32	49.94	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
25	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	381.57	282.30	200.66	143.05	110.33	97.26	97.27	99.67	102.11	94.32	75.46	52.48	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
26	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	384.69	283.13	200.66	143.88	113.43	103.55	106.82	109.23	112.26	105.23	82.59	55.02	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
27	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	387.81	283.97	200.66	144.70	116.53	109.83	113.11	118.78	121.82	116.14	89.73	57.56	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
28	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	390.93	284.80	200.66	145.53	119.63	116.12	119.39	125.07	131.38	127.05	96.87	60.10	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
29	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	394.05	285.63	200.66	146.35	122.74	122.40	125.67	131.35	137.66	137.96	104.00	62.64	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80
30	356.96	356.96	359.46	366.55	377.43	389.99	401.98	411.59	417.91	397.17	286.47	200.66	147.18	125.84	128.68	131.96	137.63	143.94	148.57	111.14	65.19	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	40.85	43.50	50.78	61.80

Figure 5.15 Cut-section of accumulated cost matrix. Reprinted with permission from [73]

Unlike the example matrix in Fig. 5.12, the directions of increasing data point index have been set read from right-to-left and bottom-to-top for $\boldsymbol{T_1}$ and $\boldsymbol{P_1}$ respectively. Thus, warping paths are read looking leftward through the matrix. The figure shows the first observed warping path traced as the yellow-colored sets of nodes, and observed between data points 90 and 110 in $\boldsymbol{T_1}$. The surrounding orange-colored cells indicate alternative, low cost neighbor nodes. Since $\boldsymbol{AP}$ is non-square, no exact diagonal exists for this path in order to capture the identical relationship between $\boldsymbol{P_1}$ and the first feature in $\boldsymbol{T_1}$. The different stacks of vertically and horizontally-aligned nodes within any warping path captures a one-to-multiple correlations between data points in $\boldsymbol{P_1}$ and $\boldsymbol{T_1}$. This can be attributed to the time-stretch and data point fitting along the different time points for both measurements resulting in the identification of a time-error pattern $\boldsymbol{P_1}$ in $\boldsymbol{T_1}$.

The incremental cost of the alternative low-cost nodes are minimal or zero at worst. Combinations of these nodes with those in the warping path give rise to redundant paths which are

elongated or incomplete time-error patterns. These are similar to overlapping, neighbor subsequences in a time series.

Fig. 5.16 shows a plot of the *DTW* distance computed for every data point in $\boldsymbol{T_1}$.

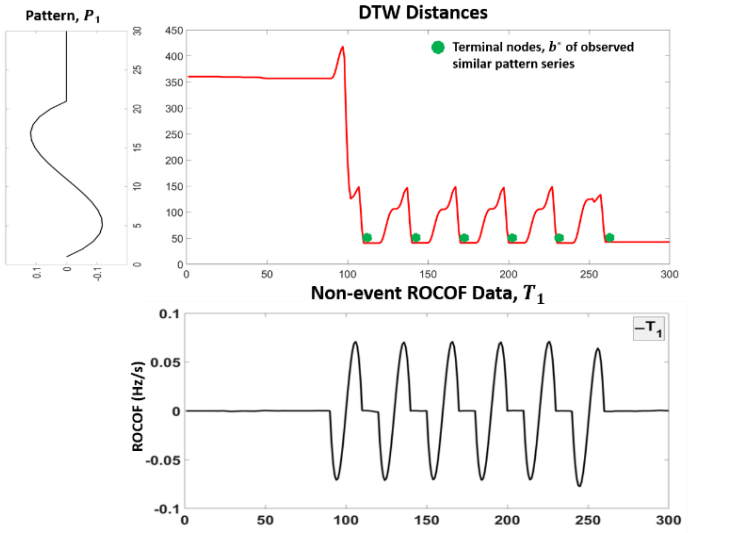


Figure 5.16 Noise-free ($\boldsymbol{P_1}$, $\boldsymbol{T_1}$) *DTW* distances in $\boldsymbol{T_1}$. Reprinted with permission from [73]

By visual inspection, an initially high *DTW* distance between $\boldsymbol{P_1}$ and the early data segments of $\boldsymbol{T_1}$ is indicative of the large disparity between both signals. This value is observed to suddenly drop off (~ 41) once the first unique feature is completely identified in the $\boldsymbol{T_1}$ subsequence ($\boldsymbol{SS_1 = T_{1,90}: T_{1,110}}$) by a full cycle of $\boldsymbol{P_1}$. This is represented by the first, green ball in the low-cost valley region of the graph. A successive rise-and-fall in the *DTW* distance is observed when another subsequence feature is detected. The non-overlapping, unique subsequences in consecutive valley regions with similar *DTW* distances all account for the remaining instances when $\boldsymbol{P_1}$ is identified in $\boldsymbol{T_1}$.

5.4 Summary

In this chapter, a distributed approach for a wide-area data error analysis has been proposed for use on a large-scale system. The preliminary results indicate that the localization of system dynamics can help discriminate true PMU data error sources from system events. Furthermore, a similarity-based technique to track time-based errors in PMU has been implemented. Here, it is important to state that for the proposed pattern search to be effective, the signal must have a low noise content for the time-error patterns to be detectable in the data.

6 PRESENTING RESULTS FROM PMU DATA ERROR ANALYSIS

In this chapter, error metrics computed for the different measurements in the 2000-bus system and PMU device information about the synchronization status of each data sample are presented in visual form. An overview of the multidimensional scaling technique, and how it was applied in this work to the error cases in the previous chapter are presented.

6.1 Data Error Visualization Using Multidimensional Scaling

Depending on the number of data segments assessed for errors, the computed time-series, LOF values for each PMU measurement can generate vectors of high dimensions. The method which has been adopted for visualizing errors in this work relies on observing correlations among different bus measurements for any given data error type, and visually displaying them as dis-similarities in a two dimension chart. The different charts generated are based on the use of the multidimensional scaling (MDS) which can be a useful method for representing power system information [120-123].

MDS is used to represent measurements of similarity or dissimilarity among pairs of objects as distances between points of a low-dimensional, multidimensional space [124-126]. Given the time-series error values ($\boldsymbol{LOF_{orig}}$) computed at m different data segments for n PMU measurements, and a pairwise, proximity matrix ($\boldsymbol{\delta}$) between them in (1) and (2) respectively.

$$\boldsymbol{LOF_{orig}} = \begin{bmatrix} Lof_{1,1} & \dots & Lof_{1,m} \\ \vdots & \ddots & \vdots \\ Lof_{n,1} & \dots & Lof_{n,m} \end{bmatrix} \tag{1}$$

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_{11} & \dots & \delta_{1n} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \dots & \delta_{nn} \end{bmatrix}, \delta_{ij} = \begin{cases} dissim\left(Lof_i, Lof_j\right); & i \neq j \\ 0; & i = j \\ & i, j = 1,2, \dots n \end{cases} \tag{2}$$

The output of the MDS algorithm is a set of n coordinates, $\boldsymbol{LOF_{coord}}$ in two dimensions, which re-re-represents $\boldsymbol{LOF_{orig}}$, and preserves or approximates the pairwise proximities in $\boldsymbol{\delta}$. That is,

$$\boldsymbol{LOF_{coord}} = \begin{bmatrix} l_{1,1} & l_{1,2} \\ \vdots & \vdots \\ l_{n,1} & l_{n,2} \end{bmatrix} \tag{3}$$

$$\boldsymbol{d} = \begin{bmatrix} d_{11} & \dots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nn} \end{bmatrix}, d_{ij} = \begin{cases} dissim(i,j); & i \neq j \\ 0; & i = j \end{cases} \tag{4}$$

The MDS optimization problem is then to identify the optimal set of coordinates in $\boldsymbol{LOF_{coord}}$ which minimizes a stress function which corresponds to a sum of squared errors.

$$\sigma = \underset{(l_{1,1},l_{1,2}),\dots,(l_{n,1},l_{n,2})}{arg\ min}, \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij} - \delta_{ij})^2 \tag{5}$$

Computation of $\boldsymbol{LOF_{coord}}$

The choice of classical MDS for this work is due to it being a non-iterative technique, and generating analytical solutions within a fast computation time. Classical MDS assumes the proximity matrix, $\boldsymbol{\delta}$ as a distance matrix, and finds the coordinate matrix, $\boldsymbol{LOF_{coord}}$ comprising of the two leading eigenvectors obtained from the eigen-decomposition of the normalized proximity matrix. The detailed steps are given in appendix D.

6.2 Generating Data Error, Hybrid Correlation Charts

In this work, the MDS is used to facilitate the display of the system structure by observing all PMU similarities using smaller LOF dimensions. It is used to transform the dissimilarities observed within a given dataset into a 2-D graphical representation. The benefit lies in visually displaying the dynamic electrical parameters (e.g. voltage magnitude and angle, frequency and ROCOF), and providing a better means of conveying the measurement errors.

Proximity Matrices

The multidimensional matrix, $\boldsymbol{LOF_{orig}}$ is used to compute the entries in the proximity matrix, $\boldsymbol{\delta}$ in (5.2), which are defined as Euclidean distances. That is,

$$dissim(Lof_i, Lof_j) = \sqrt[2]{\sum_{p=1}^{m} (Lof_{i,p} - Lof_{j,p})^2} \tag{6}$$

When the desired option is the ability to visualize the similarities among the synchronization status of all PMUs, based on their sync bits (bit 13 in the STAT field – see Fig. 3.8), pairwise, binary-based distances, in the proximity matrix, $\boldsymbol{\delta_{sync}}$ can be computed using the Rogers & Tanimoto binary similarity measure [126-128], given as

$$dissim(Pmu_i, Pmu_j) = \frac{a+d}{a+d+2(b+c)} \tag{7}$$

Each of Pmu_i, Pmu_j is a binary string formed by cascading all data frame sync bits in a given PMU measurement; a, b, c and d are obtained from Table 6.1.

Table 6.1 Expression of binary instances

		Object B	
		1	0
Object A	1	a	b
	0	c	d

a is the number of times elements in Object A and B are simultaneous bit-1;
d is the number of times elements in Object A and B are simultaneous bit-0;
b is the number of times elements in Object A are bit-1, and elements in B are bit-0; and
c is the number of times elements in Object A are bit-0, and elements in B are bit-1

The choice of selection of this binary similarity measure is based on the need to emphasize bit differences and similarities between PMUs. In addition, computed dis-similarity values always lies between 0 and 1.

Finally, when the choice is a visualization of PMU data drop-outs, periodic drop-out rates (d_{ri}) are computed for every sliding i^{th} window within the overall measurement string transmitted by each PMU.

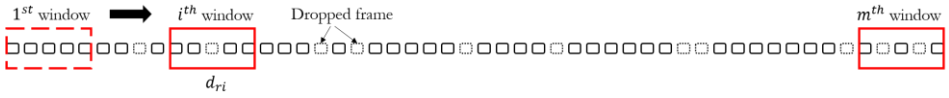


Figure 6.1 Periodic drop-out rates

The periodic rates of data drop-outs are different from the total drop-out for a PMU measurement, as large window segment drop-outs can impact the performance of an application making use of these input measurements [22]. A drop-out matrix ($\boldsymbol{d_r}$), similar to the error values, $\boldsymbol{LOF_{orig}}$, and comprising of m window drop-out rate for all n measurements is given by,

$$\boldsymbol{d_r} = \begin{bmatrix} d_{r1,1} & \dots & d_{rm,1} \\ \vdots & \ddots & \vdots \\ d_{r1,n} & \dots & d_{rm,n} \end{bmatrix} \tag{8}$$

Instead of $\boldsymbol{LOF_{orig}}$, $\boldsymbol{d_r}$ is then used as input data to compute a proximity matrix, $\boldsymbol{\delta_{dr}}$.

Selecting MDS axis for plotting

Time-series phasor measurements, obtained from PMUs, comprise of two aspects – magnitude and angle – which are both affected differently depending on the type of error. Therefore, we propose to visualize the wide-area similarity in all measurements using the relevant aspect when an error type is specified.

1. Noise Errors: Noisy signals is observed in both aspects of phasor measurements. As a result, significant noise levels are identified by visualizing the coordinate matrices of both magnitude-based and angle-based $\boldsymbol{LOF}$ matrices.

Input: Proximity matrix ($\boldsymbol{\delta_{M,LOF}}$) corresponding to magnitude-based $\boldsymbol{LOF}$; and proximity matrix ($\boldsymbol{\delta_{R,LOF}}$) corresponding to angle-based $\boldsymbol{LOF}$.

Output: First dimension of coordinate matrix, $\boldsymbol{X_M}$; and first dimension of coordinate matrix, $\boldsymbol{X_R}$

2. Timing Errors (GPS vs Clock drift): Timing errors reflect in the phasor angles. In addition, the PMU Sync bit (bit 13 in the STAT field) is flagged to '1' once time-issues are observed to have occurred. Thus, we visualize the coordinate matrices of both angle-based $\boldsymbol{LOF}$ and sync status.

Input: Proximity matrix ($\boldsymbol{\delta_{R,LOF}}$) corresponding to angle-based $\boldsymbol{LOF}$; and proximity matrix ($\boldsymbol{\delta_{sync}}$)

Output: First dimension of coordinate matrix, $\boldsymbol{X_R}$; and first dimension of coordinate matrix, $\boldsymbol{X_{sync}}$

3. Data frame drop: This type of error is exogenous to the PMU device as both phasor magnitude and angle are lost during transmission. Here, we sought to visualize only the coordinate matrix of the drop-out matrix, $\boldsymbol{d_r}$.

Input: Proximity matrix ($\boldsymbol{\delta_{dr,}}$) corresponding to $\boldsymbol{d_r}$.

Output: First and second dimensions of coordinate matrix, $\boldsymbol{X_{dr}}$.

6.3 Study: 2,000-bus (Case 2)

This case involves the outage of one of the 115-kV lines during a 10-second simulation, and the information being used are obtained from the scenario of 'Error 3' which was previously described

in chapter 4 i.e. an integration of the original event measurements with noisy signals at PMU 1250 and time errors signals at PMUs 4,538 and 764.

Based on these different types of errors (noise and time errors) which are present in the data, two different MDS visualization options are used for this purpose.

Bit Adjustments for Time Synchronization Issues

Given the occurrence of clock drift and intermittent GPS time errors as described in Table 5.7, the flag bit values indicating the status of each sample synchronization (PMU Sync bit 13) is demonstrated in Fig. 6.2 and 6.3. A bit value of zero (or OFF) indicates a synchronized sample, and a bit value of one (or ON) is a sample that is out of sync.

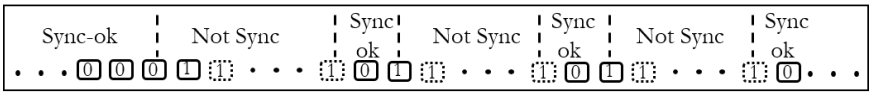


Figure 6.2 Bit flag updates for clock drift error

The clock drift is an alternating sequence of bit changes. Prior to time issues, bit 13 is 0 (Sync ok) until the moment the drifting begins with the first data sample, when bit 13 changes to 1 (Not Sync). For a reporting period, it remains out of sync, and all sample bit status is preserved as value 1. An attempt at re-synchronizing first sample in the next report cycle sets bit 13 to 0 momentarily, after which drifting continues with the second sample. Bit 13 is set back to 1 for the reporting period. The cycle is repeated for as long as the PMU clock issue exists.

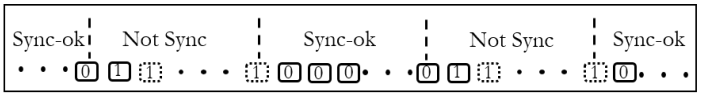


Figure 6.3 Bit flag updates for intermittent GPS error

Sequencing of bit changes in the case of intermittent GPS is to a lesser degree than the clock drift, and depends on the frequency and duration of signal loss. Hence, the non-periodic sequencing of bit changes in Fig. 7.3. A period of bit-OFF (Sync-ok, and bit value is 0), followed by another equal- or non-equal duration of bit-ON (Not Sync, and bit value is 1).

Generation of MDS Correlation Graphs

Based on Figures 6.2 and 6.3, bit adjustments due to clock drifts and intermittent GPS signal reception are performed on the bus measurements of PMUs 4, 538, and PMU 764 respectively. Execution of MDS procedures on the proximity matrices – $\boldsymbol{\delta_{M,LOF}}$, $\boldsymbol{\delta_{R,LOF}}$ and $\boldsymbol{\delta_{sync}}$ – across all 2,000 buses generates the reduced two-dimension coordinate matrices, $\boldsymbol{X_M}$, $\boldsymbol{X_R}$ and $\boldsymbol{X_{sync}}$ respectively.

Table 6.2 shows the normalized coordinates for each bus in the $\boldsymbol{X_M}$, $\boldsymbol{X_R}$ and $\boldsymbol{X_{sync}}$ matrices

Table 6.2 MDS coordinates for PMU bit-13 status flag and phasor angle error

PMU ID	**Bus IDs**	$X_{M,1}$	$X_{M,2}$	$X_{R,1}$	$X_{R,2}$	$X_{sync,1}$	$X_{sync,2}$
4	1005	0.001	0.000	-0.251	0.657	-0.512	-0.323
	1006	0.001	0.000	-0.251	0.657	-0.512	-0.323
538	5232	0.001	0.000	-0.517	-0.240	-0.591	0.031
764	6100	0.001	0.000	-0.028	0.012	-0.352	0.889
1250	8158	-0.526	0.864	-0.414	-0.070	0.001	0.000
	8159	-0.639	-0.495	-0.428	-0.190	0.001	0.000
	8160	-0.560	-0.225	-0.499	-0.193	0.001	0.000
Others	Others	0.001	0.000	0.001	0.000	0.001	0.000

In the absence of errors, the coordinates of all buses in either of the three coordinate matrices should approximately lie close to an origin i.e. (0, 0). However, isolated bus coordinates exhibiting significant deviations from this point is an indication of the presence of anomaly in the reported measurement. The hybrid-MDS graph in Fig. 6.4 is obtained by plotting $\boldsymbol{X_{R,1}}$ and $\boldsymbol{X_{sync,1}}$, and provides a method to visualize the correlations among PMU measurements when time errors are present.

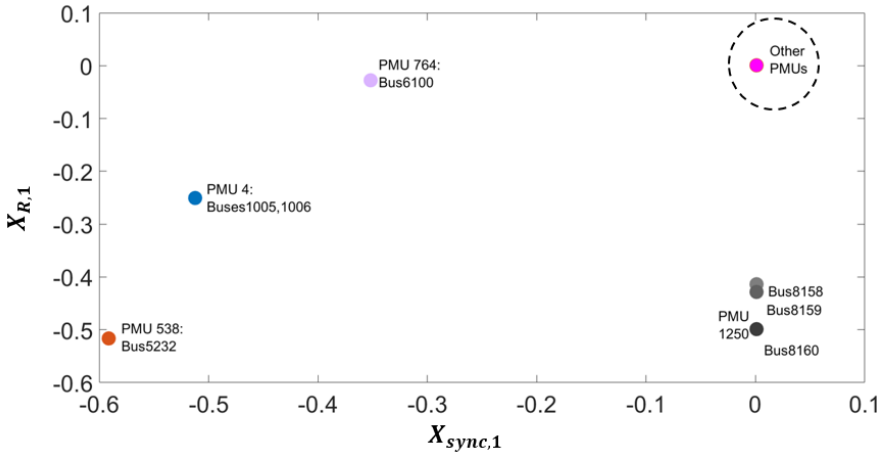


Figure 6.4 Hybrid-MDS spatial representation of noise and time errors in PMU data

The diagonal distances spanned across both sync and phasor angle correlation axis from majority of other PMU measurements is indicative of actual time errors present in some of the reported measurements. The most severe cases are observed to occur in measurements obtained from PMUs 4 and 538 where the clock drift error had occurred for the most part of the simulation (i.e. 7 and 9 seconds respectively) followed by the five instances of intermittent, external time synchronization in PMU device 764. The vertical distance between PMU ID 1250 and 'other PMUs', however does not necessarily indicates a time error as the device is correlated in the sync axis with 'other PMUs'. Further investigations, by updating the graph to a plot of $\boldsymbol{X_{R,1}}$ and $\boldsymbol{X_{M,1}}$ in Fig. 7.5, reveal the large deviation of PMU 1250 with respect to the other PMUs along the phasor magnitude correlation axis. The data error can be attributed to other causes which simultaneously impact on both components of phasor measurements, and in this case, the effect of noisy signals.

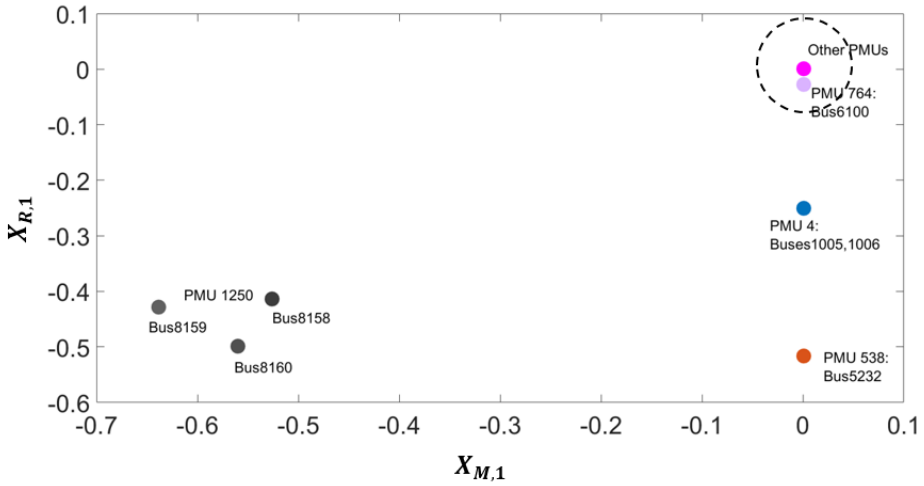


Figure 6.5 Hybrid-MDS spatial representation of errors in PMU magnitude and angle data

6.4 Summary

PMUs and other synchrophasor devices are known to report phasor quantities (comprising of magnitude and angle), it implies that measurement errors can either appear in the magnitude, angle or both magnitude or angle. In this chapter, a hybrid-MDS scheme is proposed to visualize the different aspects of PMU measurement errors.

7 VISUALIZATION OF LARGE-SCALE ELECTRIC GRID OSCILLATION RESULTS*

In this chapter, wide-area visualization methods are used to present results of low-frequency disturbance modes obtained from the analysis of a large-scale power system. In the first section, an improved modal analysis technique used in this work is discussed. Wide-area visualization of mode information pertaining to mode estimation quality, oscillation mode activities and source of oscillations are presented in the subsequent section.

7.1 An Improved, Iterative Mode Decomposition Technique

The goal of modal analysis is to obtain a re-constructed signal , $\hat{y}(t)$ which is a sum of un(damped) sinusoids and considered to be a close approximate of an original signal $y(t)$. The observations not fully captured by the reconstructed signal constitutes the error signal, $e(t)$.

$$\hat{y}(t) = \sum_{j=1}^{q} A_j e^{\sigma_j t} \cos(\omega_j t + \phi_j) \tag{1}$$

$$e(t) = \sum_{j=1}^{q} \left\| y(t_j) - \hat{y}(t_j) \right\|_2^2 \tag{2}$$

The j^{th} mode is characterized by the modal parameters: damping factor (σ_j), frequency (ω_j) and mode shape consisting of amplitue (A_j) and phase (ϕ_j). q is the number of dominant low-frequency modes captured by the analysis.

An improved method, which iteratively minimizes this error has been proposed in[129]. It uses a subset of the signals to determine significant modes after which reconstructed signals are

*Part of this section is reprinted with permission from " Visualization of Large-Scale Electric Grid Oscillation Modes" by I. Idehen, B. Wang, K. Shetye, T. Overbye and J. Weber, Sept. 2018 North American Power Symposium (NAPS) © 2018 IEEE, with permission from IEEE

calculated for the observed measurements. If a system comprises of n number of time-series signals, the system modes can be approximated using m multiple signals obtained using heuristic methods such that $1 \leq m \leq n$. For every signal added to the subset, a signal reconstruction is carried out for all n original measurements. A good choice of m is set to balance between the computation time and better capturing dominant system modes. The improved method is a general technique which can be applied to any of the mode decomposition techniques mentioned in the literature [63-66, 68, 70] . However, this work utilizes the matrix pencil analysis (*MPA*) technique as it is tolerant to the presence of noise in any observed set of measurements.

Given a data set $\boldsymbol{Y}$, such that $\boldsymbol{Y} = \{y_1(t)^T, y_2(t)^T \dots y_n(t)^T\}$, the pseudo-code for an iterative matrix pencil analysis procedure is shown in Figure 7.1.

Input = $m, \boldsymbol{Y} \in R^{n \times \alpha}$; $count = 1$;
Step 1. Initialize: $[\boldsymbol{y}] = [\boldsymbol{y_{ini}}]$; $\boldsymbol{y_{ini}} \in \boldsymbol{Y}$
Step 2. Execute: MPA on $\boldsymbol{y}$
Output: ω_i, σ_i ; $i = 1$ to q
$A_{i,j}, \phi_{i,j}$; $j = 1$ to n
Step 3. Compute: Re-constructed signals, $\widehat{\boldsymbol{y}}_{\boldsymbol{j}}(t)$
Cost function, $\varepsilon_j = \sum_{k=1}^{\alpha} \left(\left| \boldsymbol{y_j}(k) - \widehat{\boldsymbol{y}}_{\boldsymbol{j}}(k) \right| / \alpha \right)$
Step 4. Aggregate cost functions, $\boldsymbol{\varepsilon} = [\varepsilon_1 \dots \varepsilon_n]$
Step 5. Obtain: Maximizing error signal, $\hat{\jmath} = \underset{j}{arg\ max}\{\boldsymbol{\varepsilon}\}$
Step 6. Update $[\boldsymbol{y}] = \left[\boldsymbol{y}, \boldsymbol{y}_{\hat{\jmath}}\right]$; $count = count + 1$
Repeat Step 2 - Step 6 until $count = \alpha$
Output = Re-constructed signals, $\widehat{\boldsymbol{y}}_{\boldsymbol{j}}(t)$, $\omega_i, \sigma_i, A_{i,j}, \phi_{i,j}$

Figure 7.1 The iterative MPA

α is the number of equally-spaced samples in a signal $y_1(t)$, and $\boldsymbol{y_{ini}}$ is the initializing signal used for the procedure, and which can be any of the signals in $\boldsymbol{Y}$. The cost function (CF), ε is the total absolute error over the different time points which is averaged over the number of samples.

$$\varepsilon = \sum_{i=1}^{\alpha} \frac{|y(i) - \hat{y}(i)|}{\alpha} \tag{3}$$

$\boldsymbol{y}$ is incrementally populated with the signal which maximizes the set of CFs at every iteration, after which the MPA algorithm is performed. Full details of the MPA is presented in appendix E.

The sensitivities of the computation time and maximum cost functions to the number of signals, m using frequency measurements obtained from the 2,000-bus (case 3) is presented in Fig. 7.2.

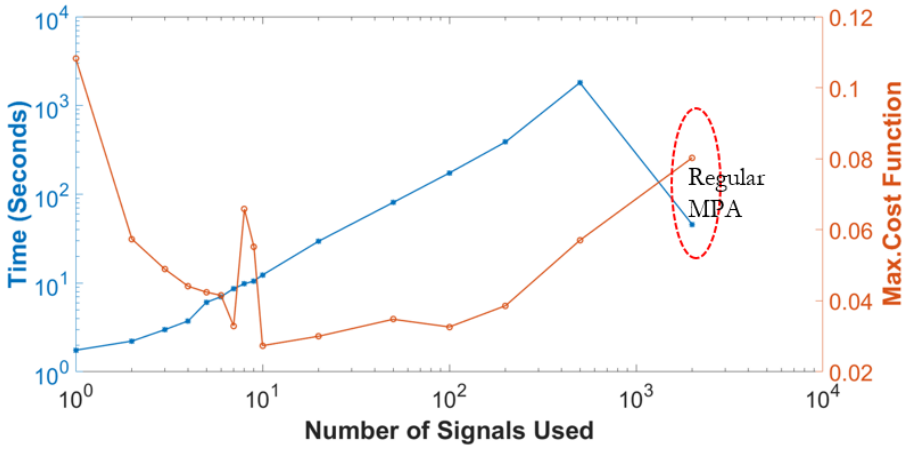


Figure 7.2 Sensitivities of computation time and maximum cost function to number of signals

The computation time is observed to increase with the number of signal inclusions in the dataset used for mode decomposition, while the maximum cost function increases after a period of initial decline. We hypothesize that the inclusion of more signals, beyond a threshold limit, in the dataset introduces more local signal variations, thus resulting in the identification of regionalized or non-system modes that are absent in other signals. Hence, the increase in the maximum cost function. It is also important to identify the point at which the performance benefits of using the proposed

iterative method are minimal or non-existent. For example, regular MPA was comparatively better in computation time than the iterative method when 100 signals (hence 100 iterations) were used.

7.2 Wide-Area Visualization of Modal Information

The wide-area display of results from the synthetic networks which we utilize are based on pre-existing one-line diagrams which are over laid on a geographical map of the United States. In addition to the use of contour plots, dynamic data are visualized using geographical data views (GDVs) providing an information layering technique to present large amounts of data [130-132]. This enables user customization of a display in order to see power system quantities by making use graphical objects. Information is encoded using the attributes of object color, size, rotation or shape [133].

7.2.1 Quality estimation of modal analysis technique

The desire is to approximate as closely as possible each original signal using the signals from (1). However, the few dominant system modes are not sufficient to fully represent the original signal and other dynamics in the system. The quality of the mode estimation process is thus measured in terms of the difference between the original and reconstructed signal.

Fig. 7.3 shows the actual (blue), reproduced (red) frequency signals and the CFs (mismatch errors, ε) at nine different buses for the 2,000-bus (Case 3). Here, the case involves a 10-second simulation during which two generators are disconnected after one second.

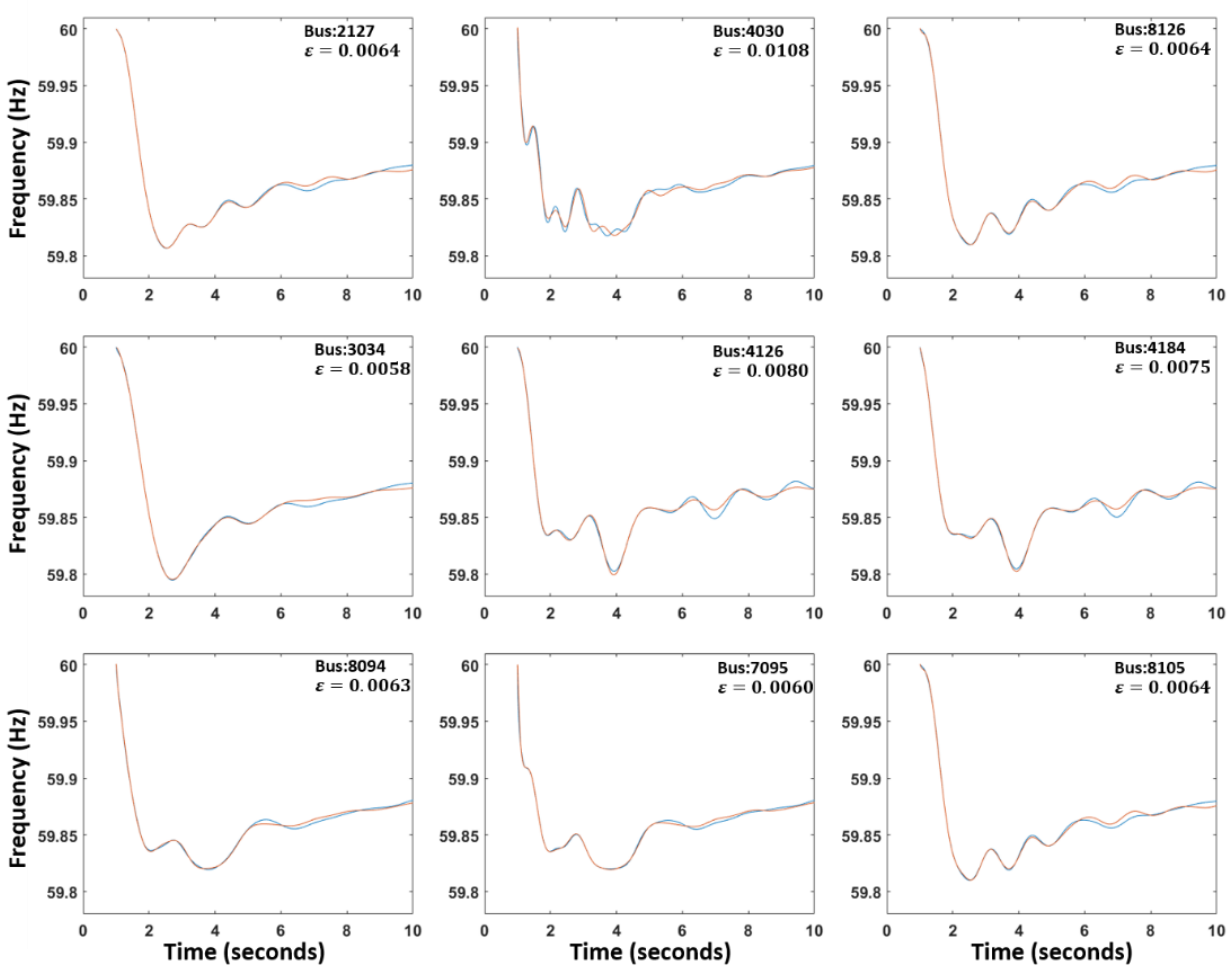


Figure 7.3 The cost functions, actual and reproduced frequency signals at 9 locations

Using the definition in (2), the computed best and worst case cost functions are 0.0058 and 0.0108 respectively, and which were observed at bus IDs, 3034 and 4030, respectively. However, relatively low values of the extreme CF quantities indicate the good matching ability of the proposed technique. The wide-area trend of the CF is shown in Fig. 7.4.

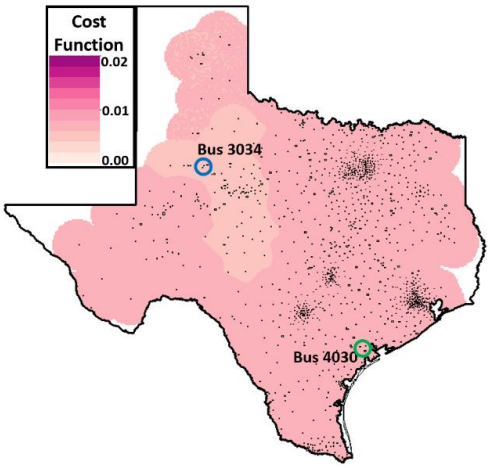


Figure 7.4 Wide-area system cost function

A color scale has been set arbitrarily [0, 0.02] to indicate the best and worst case matching errors while the signal buses 3034 and 4030 are enclosed by blue and green circles respectively. The uniform variation of the cost function is largely indicative of the global pattern of system frequency, and good quality of the modal technique used for this purpose.

Fig. 7.5 (a) shows the wide-area trend of the CF when modal decomposition was applied on the voltage measurements.

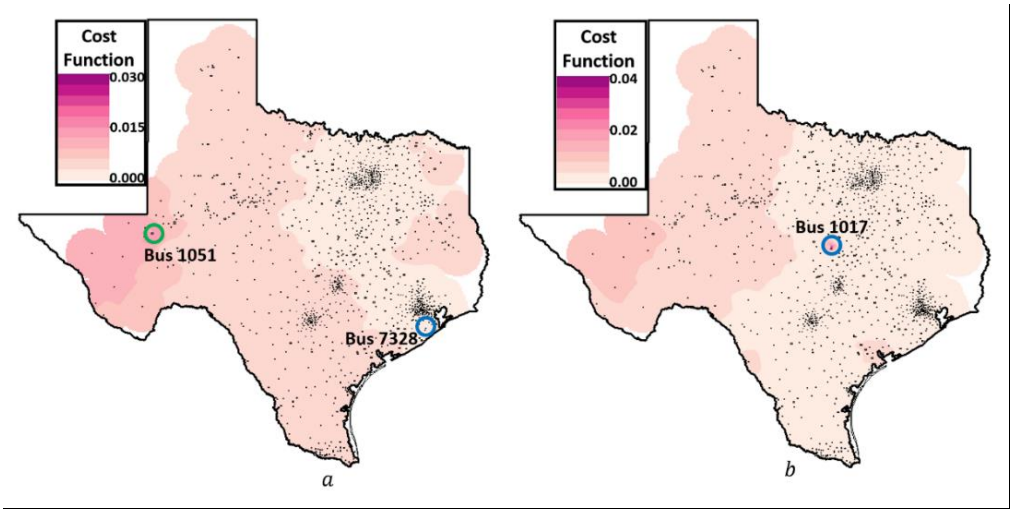


Figure 7.5 (a) Wide-area cost function using voltage measurements; (b) with noise signal at bus 1017

A wider CF range, [0.0014 0.028] at bus IDs 7328 and 1051 respectively, and more variation in all bus CFs are indicative of the local action of voltage trends observed in the system after the contingent generator outage. This can be attributed to the fact that fewer frequency modes could be prevalent at different bus locations, however they remain invisible to the system. Hence, they are not captured during signal reconstruction. Assuming a high modal decomposition quality, wide-area visualization of CFs (using voltage measurements) can become helpful to operators by providing early indicators of imminent system instability (e.g. voltage collapse).

Another unique case of CF variation that could point to an event, and thus assist operators in understanding the system is when locations report erroneous data deemed to be inconsistent with the actual system trend. Fig. 8.5(b) is the wide-area CF when noisy data is reported by a PMU device at bus location 1017. High CF at an isolated bus location indicates a prevailing, anomaly condition, and especially when nearby buses have much lower CF values.

7.2.2 Oscillation Modes

The mode shape describes the relative activity of the state within an oscillation mode. Comprising of both magnitude and angle information, this vectoral attribute can be a distraction source when visualizing individual signal mode shape information in a wide area network.

Fig. 7.6 shows the current phasor technique used to view mode shapes in different sections of the power system [25, 29, 30]. Mode shapes at twenty different bus locations (a to t) are currently being displayed.

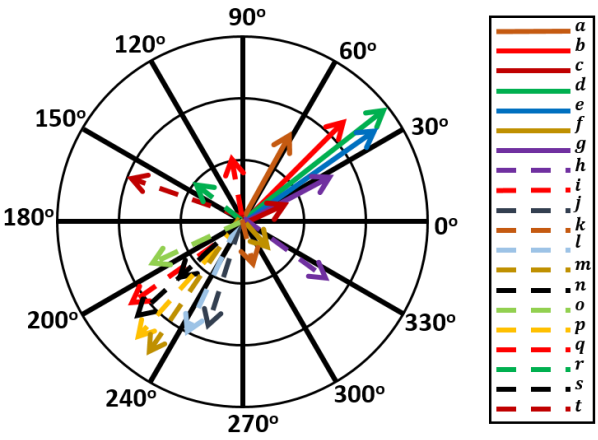


Figure 7.6 Phasor vector plot of mode shapes at 20 bus locations

Using this method, the figure is able to capture the relative magnitudes and angles at the different buses. However, as mode vectors align in the same direction, and vector magnitudes become equal, the occurrence of significant vector overlaps result in the inability to distinguish among the mode shapes at the different buses. Most importantly, extracting the underlying system dynamics information from the phasor diagram is challenging without the use of an actual geographic map.

Vector Field Visualization

The vectoral characteristics of all bus mode shape information makes them amenable to being represented as two-dimension (2D) vector fields on geographical-based, one-line diagram of the system [131, 134, 135] . Among the different forms for vector field visualization in 2D surfaces, the choice of using arrow icons on a rectangular grid (GRID) vector field visualization is predicated on its ability to convey a sense of bus swing direction at any of the grid regions. This information is more critical to an operator rather than, for example, the short time it may take an operator to identify a critical point on the vector field if line-integral convolution (LIC) were used

[135]. In addition, the GRID method has the ability to help users identify critical points within local neighborhoods on the vector field, which could indicate locations in need of attention. For example, the arrows forming the boundary of the green-colored, contour region of the grid indicates the extent of bus inclusions in the two-area swing of the system.

Based on the highlighted benefits of using 2D vector fields, a more effective, wide-area visualization is implemented to address the challenges faced by the phasor plot. This technique makes use of the attributes of glyph objects (phasor arrows) which are geographically-distributed on the one-line diagram to capture mode activities at all the individual buses. As a layering option, contour plots which encode other bus or area information (e.g. the direction of swing) are set in the background to provide more system dynamics that might not be fully captured by the mode vectors. Fig. 7.7(a) and (b) show the mode shape information for an inter-area mode (0.541 Hz) and a local mode (3.576 Hz) using the frequency measurements obtained from the 2,000-bus (Cases 4 and 5 respectively). All signal amplitudes have been scaled by their standard deviation values.

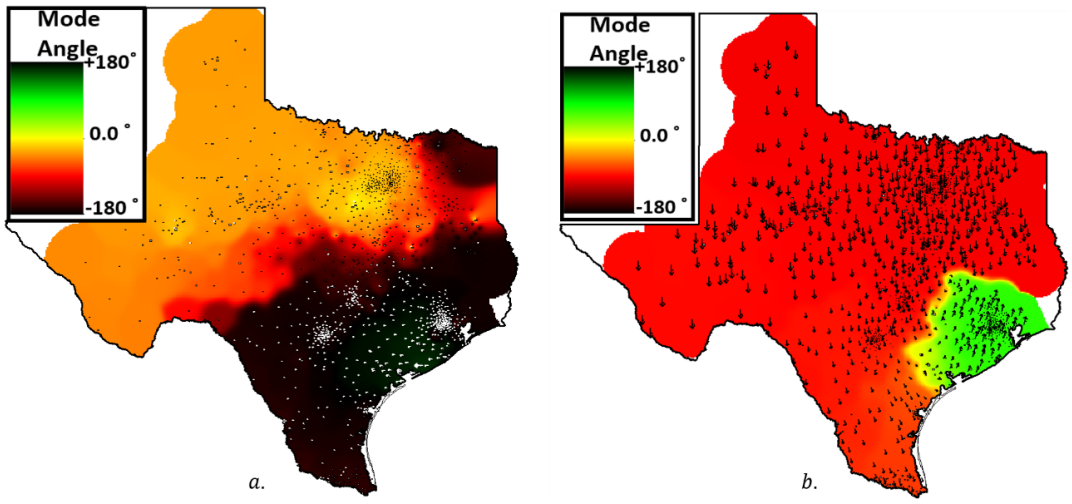


Figure 7.7 Frequency mode shape for (a) local, and (b) inter-area modes

Abrupt change in colored contours between angle limits, -180˚ and 180˚, are often misleading since in actual geometry both angles are exact. To avoid this sudden color change, a circular (or cyclic) color map, which assigns the same or similar colors to angular values close to these limits, has been used. The color map used to highlight the mode angles at different signal bus locations provides a user with a wide-area summary of the swing direction at the different buses. Individual bus signal amplitude and angle are encoded in the size and orientation of the phasor arrow relative to the positive x-axis respectively. The geographical information of each bus is used to set the location of its GDV-based arrow. The inter-area oscillation in Fig. 7.7(b) shows two marked distinct areas, such that buses in these regions have a similar direction of swing for the oscillation mode. A comparison of the arrow lengths indicates the lower level of mode activity in the local mode of Fig. 7.7(a) than the inter-area oscillation in Fig. 6.8(b).

7.2.3 Bus Coherency

In understanding the key dynamic stability behavior of large interconnected systems, the user is often interested in the system bus coherent groups (e.g. for controlled islanding) [136-138]. However, depending on the color map used in Figs. 7.7 (a) and (b), the use of a wide color spectrum for the mode shapes at different bus locations could conceal the actual system, global dynamic behavior. Hence, the need for bus aggregation, and use of fewer chromatic colors to represent the different groups [131]. The application of machine learning or any other similar techniques, with the ability to intelligently aggregate unidirectional bus mode shapes into smaller coherent group formations, uncovers the dynamic behavior of the system and provides further insights to the users.

Given two signals with phase angles ϕ_1 and ϕ_2, the angular distance (d_{12}) between them is computed from,

$$d_{12} = 1 - \cos(\phi_1 - \phi_2) \tag{3}$$

As much as possible, the goal is to identify large groups which capture major, modal coherent groups whose individual buses are in the same swing direction. To accomplish this, quality threshold clustering technique [139-141] has been applied.

Quality Threshold (QT) Clustering

This technique searches for the largest-sized cluster, containing the most similar set of objects, at every iteration until all objects have been grouped. It requires a pre-specified threshold distance (d_{th}), and which in this case, is based on the angular distance separation in (7.7). Every bus is initialized as a candidate cluster center, and computes the number of buses that are within a d_{th} distance. The pseudo-code for this procedure is stated in Fig. 7.8.

Input: d_{th}; $\mathcal{S}$=Set of all buses; ClusNum = 0;
while $\mathcal{S} \neq \emptyset$
  ClusNum = ClusNum+1;
  for i = 1 to $n(\mathcal{S})$
    $Bus_{cluster,i} = \{Bus_j\}\, s.t.\ \ d_{candidate(i),j} \leq d_{th}$
  end
  Cluster_ ClusNum =$Bus_{cluster,\hat{\imath}}\ s.t.\ , \hat{\imath} = \ ^{arg\,max}_{i}\{Bus_{cluster,i}\}$
end

Figure 7.8 QT clustering

Fig. 7.9(a) and (b), respectively show the system-wide, inter-area (0.48 Hz) and local mode (1.71 Hz) information for the 10,000-bus involving the deactivation of 19 stabilizers and outage of 2 system generators (Cases 2 and 1 respectively).

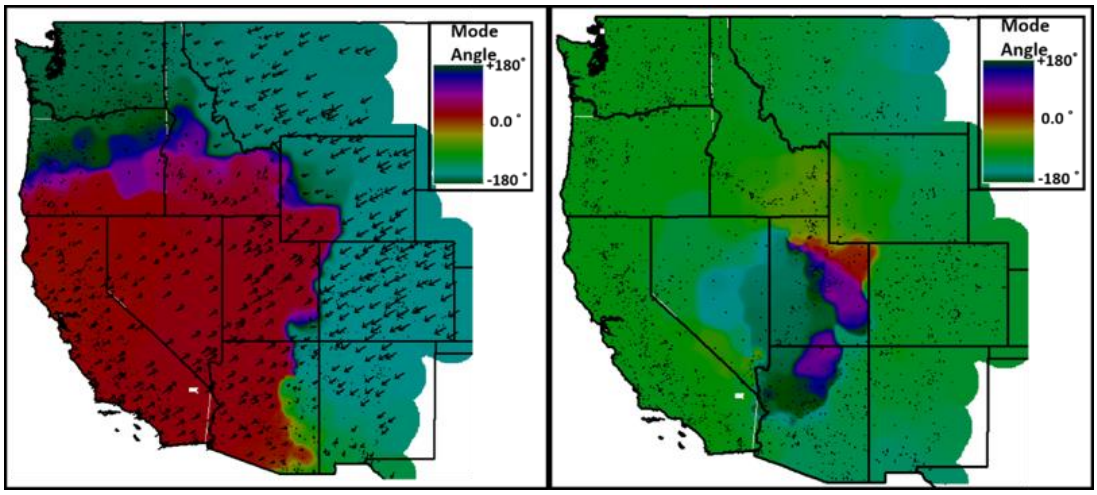


Figure 7.9 Frequency mode shape for (a) inter-area mode (0.48 Hz), and (b) local mode (1.71 Hz). Reprinted with permission from [133]

The variation in the mode shapes are similar to those observed in Fig. 7.7, and are also prone to wrong interpretation when a wide color spectrum is used. Fig. 7.8 (a) and (b) show contour plots of the identified coherent modal groups after clustering.

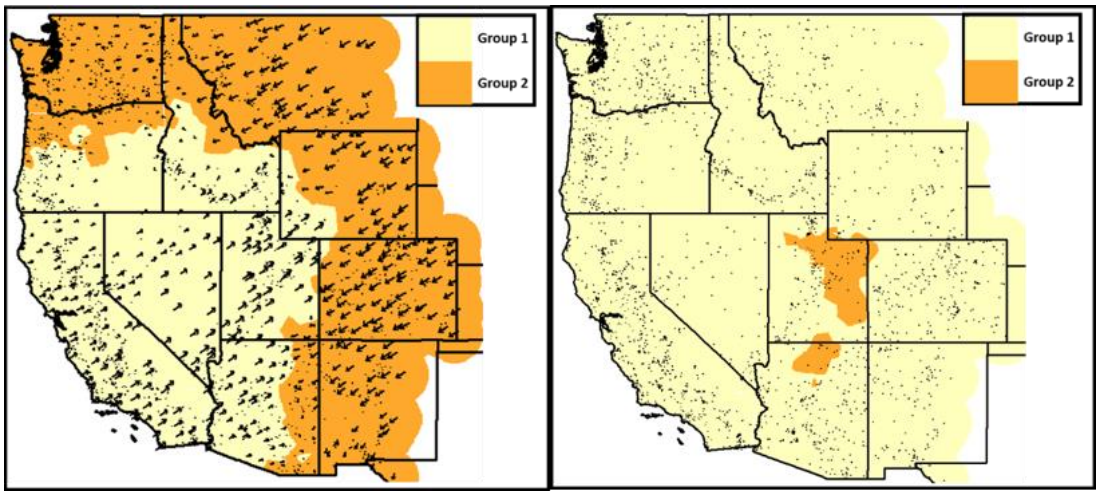


Figure 7.10 Frequency coherent groups for (a) inter-area mode (0.48 Hz), and (b) local mode (1.71 Hz). Reprinted with permission from [133]

Fig. 7.10(a) shows a dominant western-eastern, inter-area oscillation as observed by the group formations and opposing directions of signal phasor arrows in both groups. Local mode oscillation is indicated by the formation of the smaller group 2 in Fig. 7.10(b). The same can be done likewise for the frequency modes previously obtained for the 2,000-bus network in Fig. 7.7.

In summary, key dynamic system behavior and extent of mode activities can be easily captured for all identified modes in the system by clustering mode shapes and visualizing them similar to Fig. 6.8.

7.3 Visualization of Oscillation Sources

Sustained oscillations pose a threat to the safe and secure state of the system, and it is important to identify oscillation sources in order to eliminate them. An energy-based method [142-144] is used for locating the source of oscillation by computing several dissipating energy (DE) coefficients associated with oscillation energy flowing across different transmission lines in the network. In large, inter-connected systems, a wide-area visualization of branch energy flows becomes critical for users to reliably point to disturbance sources by tracking the directions and magnitudes of DE flow arrows.

7.3.1 Oscillation Energy Flow

Given any branch ij, the branch dissipating energy computed in [142] is given by (4)

$$W_{ij}^{D}(t) = \int (\Delta P_{ij} d\Delta\theta_i + \Delta Q_{ij} d(\Delta \ln V_i))$$
$$= \int (2\pi \Delta P_{ij} \Delta f_i dt + \Delta Q_{ij} d(\Delta \ln V_i)) \quad (4)$$

ΔP_{ij}, ΔQ_{ij} are deviations from the steady-state active and reactive flows of branch ij; $\Delta\theta_i$ and Δf_i are deviations in the bus angle and frequency at bus i; and $\Delta \ln V_i = \ln V_i - \ln V_{i,s}$, where $V_{i,s}$ is the steady-

state voltage magnitude. However, due to the sign constraint imposed on the $\ln V_i$ term in (4) for filtered signals, [143] replaces the term $\Delta \ln V_i$ with an approximation term given by $d(\Delta V_i)/V_i^*$. The branch oscillation energy is then computed as

$$W_{ij}^D(t) \approx \int \left(2\pi \Delta P_{ij} \Delta f_i dt + \Delta Q_{ij} \frac{d(\Delta V_i)}{V_i^*}\right) \quad (5)$$

Furthermore, a dissipating energy coefficient, DE is obtained by fitting $W_{ij}^D(t)$ using a linear model, $DE_{ij}.t + b_{ij}$. Regardless of the type of oscillation, it is observed that the ratio of branch DE is relative constant, hence a normalization of all branch DEs is often used to preserve all branch DE relationships.

$$DE_{ij}{}^* = DE_{ij} / {}^{max}_{i,j}\{|DE_{ij}|\} \quad (6)$$

The direction of oscillation is dictated by the sign of DE_{ij} – negative sign indicating energy production from a source to the network element dissipating the energy. In a wide-area visualization sense, computed DE coefficients can then be aggregated at each bus to show the net contribution to oscillation energy in the network.

7.3.2 Source of Oscillations

Implementation of the oscillation energy flow is carried out for the 2,000-bus (Case 4) where one of the system generators is set to negative damping and a 500-kV line is outaged. A 3.576 Hz local mode, with a negative damping of -0.06, is identified in the system, which indicates a sustained system oscillation. Using the measurements (frequency and voltage at all buses, and real and imaginary power flows across all transmission lines) obtained for this mode, the energy-based approach is used to track the source of disturbance in the system. Fig. 7.11 shows the time-evolution oscillation energy and computed DE coefficient for all branches in the network.

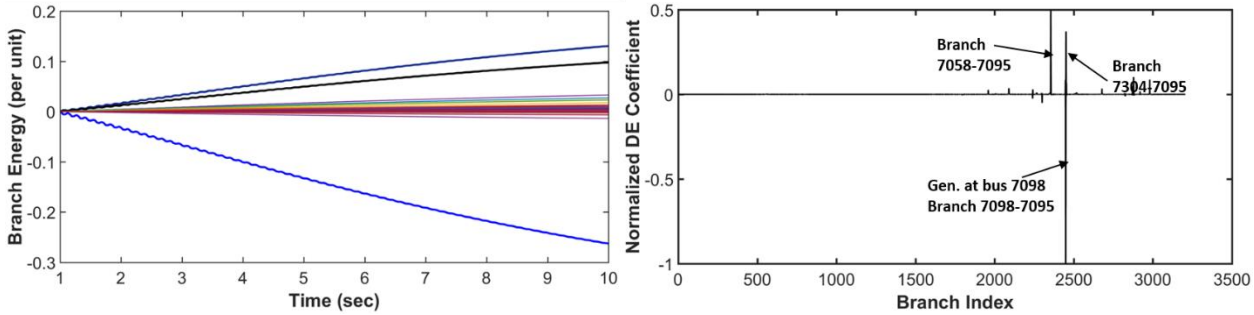


Figure 7.11 All branch oscillation energies and dissipating energy (DE) coefficients

The increasing, outward flow of oscillation energy on the branch connected to bus 7098 is due to the negative damped response of the generator at the node, which is supported by the computed DE value. Relatively few transmission lines are involved in the flow of oscillation energy. Using size and color attributes of GDV-based, ovals to encode bus DE magnitude and direction of flow of the oscillation energy respectively, Fig. 7.12 is able to quickly convey to a user the source of oscillation. Constant generation of oscillation energy is a result of the negative damping which was set on the generator machine. An informed, control decision (e.g. disconnect the generator from the system) can then be taken.

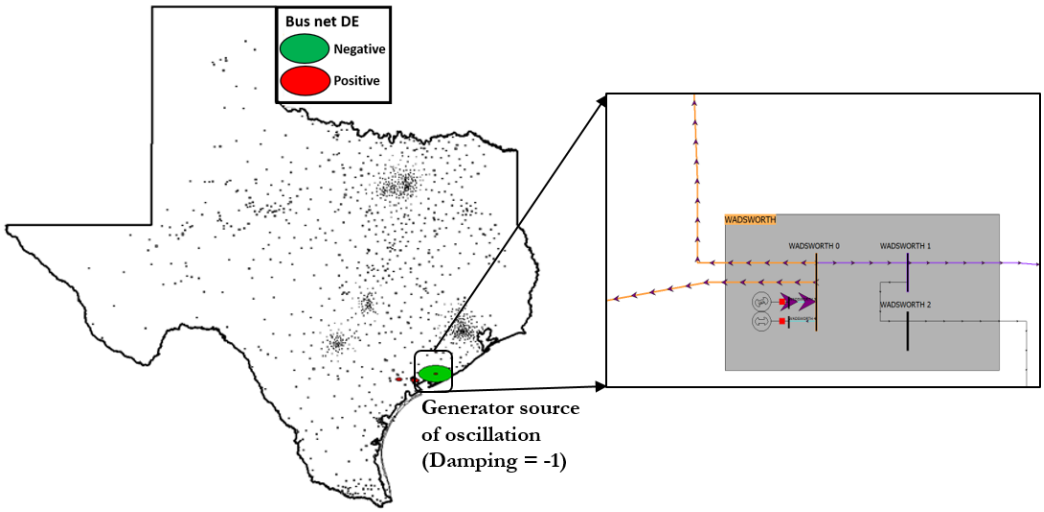


Figure 7.12 Oscillation source and branch *DE* flow

Inter-area oscillations are more complex than local oscillations, as it involves a higher participation of majority of the system transmission lines, buses and substations. Several research works are still being performed to understand this type of oscillation. An example wide-area visualization of an inter-area mode is shown in Fig. 7.13, and is based on the computed oscillation energies and branch DE values in Fig. 7.14.

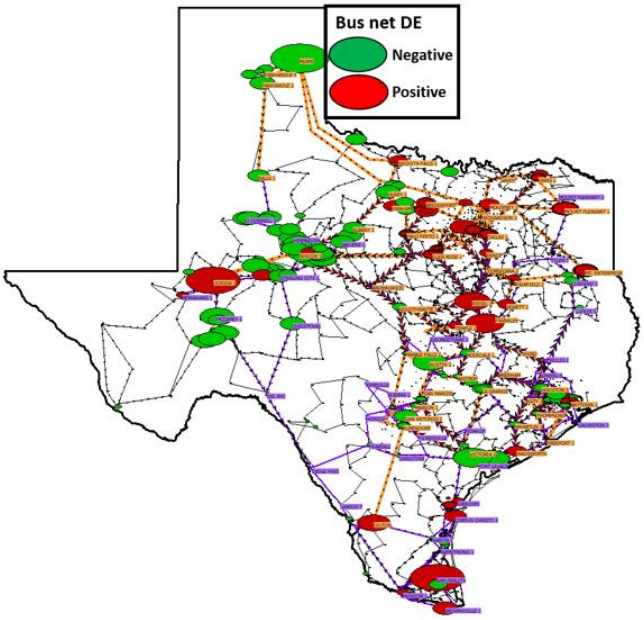


Figure 7.13 Oscillation source and branch *DE* flow

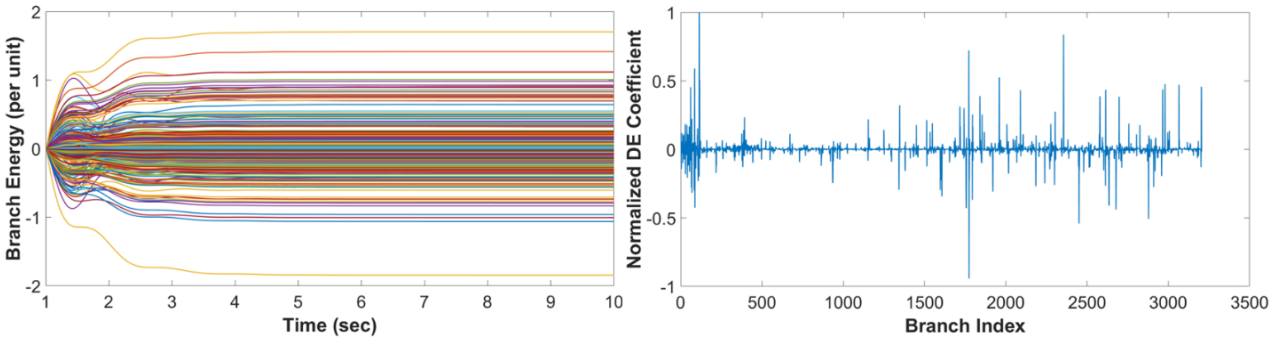


Figure 7.14 All branch oscillation energies and dissipating energy (*DE*) coefficients

7.4 Summary

In this chapter, a visualization of electric oscillation modes, and the corresponding sources, in large-scale, interconnected grids is proposed to unravel the underlying dynamics of these systems. It ensures that operators can have a better comprehensive understanding of their systems.

8 CONCLUSION

8.1 Summary

In this work, techniques for the analysis and presentation of large-scale, grid information embedded in phasor measurement unit (PMU), or alternatively synchrophasor, datasets were presented. First, leveraging on the features of real, power systems PMU measurements, and other phenomena which cause variation levels in industry-grade data as studied in the literature, we generated realistic, research-based, synthetic dataset from pre-existing synthetic grids. Initial validation results indicated the retention of the underlying, electrical behavior of the power system used, while ensuring a close semblance with real PMU data. Secondly, we proposed techniques for the reliability assessment of PMUs through the detection and discrimination of measurement errors from actual system events. Here, an application of a dynamic time warping, pattern-learning technique was used to identify a clock time error in PMU based solely on reported measurements. Finally, we implemented wide-area techniques to condense and visualize hidden dynamics of large, interconnected systems.

8.2 Future Direction

The initial development of synthetic dataset in this work reveals opportunities for advancing the creation of more realistic, PMU measurements for use in data-driven, power systems research. Some of the available opportunities for improving these synthetic datasets include:

1. An expansion of the phenomena causing variations in PMU data. For example, the inclusion of multiple contingencies (such as frequent line-switching activities, transformer tap control and saturation, consideration of nominal voltage ratings, and a better distribution of different load types/models across the grid) to simulate longer-duration, real grid operations.

2. An assessment of variabilities in actual PMU measurements obtained from different, real sources. The goal will be to establish a common reference for which the introduced variabilities in realistic synthetic datasets are tuned.
3. An evaluation of the techniques and metrics for validating the extent of realism of the generated, synthetic datasets.
4. Currently, variability factors (for examples, those due to system randomness and noise) have been mostly applied uniformly across the system. However, and as expected, preliminary investigations carried out on the real data revealed the existence of varying levels of correlations among PMU measurements depending on their nominal voltages. The use of multiple variability factors, each assigned to specific nominal voltage levels will incorporate more realism to the synthetic dataset.

Beyond event detection is the problem of grid disturbance identification and classification. Given realistic, synthetic datasets can be generated, a dictionary comprising of various patterns of disturbances can be developed to distinguish among events detected on the grid. For example, a generator trip pattern is observed by the sudden drop in frequency measurements within a few number of cycles. We propose that the implementation of techniques, such as dynamic time warping and moving-window, principal component signatures can aid the classification of events occurring on the power system.

In furtherance to the visualization work presented in chapter 7 is the subject of reactive power reserves monitoring in a bid to avoid most of the voltage instability issues associated with dangerously, low reserve levels. Current techniques utilize markers and dashboard visualizations to locate the ancillary, reactive power sources present in the system. Taking advantage of the visualization techniques discussed, one can present wide-area information on the available levels

of reactive power assets, and their corresponding, most-influential regions (often known as voltage basins), across the grid. The motivation for an enhanced, reactive power visualization method will be to better inform Engineers on the available, effective voltage control actions in the event of grid contingencies.

REFERENCES

[1] N. S. Group, "Technical analysis of the August 14, 2003, blackout: What happened, why, and what did we learn," *report to the NERC Board of Trustees,* 2004.

[2] E. H. Allen, R. B. Stuart, and T. E. Wiedman, "No Light in August: Power System Restoration Following the 2003 North American Blackout," *IEEE Power and Energy Magazine,* vol. 12, no. 1, pp. 24-33, 2014.

[3] D. Novosel, V. Madani, B. Bhargave, K. Vu, and J. Cole, "Dawn of the grid synchronization," *IEEE Power and Energy Magazine,* vol. 6, no. 1, pp. 49-60, 2008.

[4] G. Andersson *et al.*, "Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance," *IEEE Transactions on Power Systems,* vol. 20, no. 4, pp. 1922-1928, 2005.

[5] P. Kundur, C. Taylor, and P. Pourbeik, "Blackout experiences and lessons, best practices for system dynamic performance, and the role of new technologies," *IEEE Task Force Report,* 2007.

[6] B. Liscouski and W. Elliot, "Final report on the august 14, 2003 blackout in the united states and canada: Causes and recommendations," *A report to US Department of Energy,* vol. 40, no. 4, p. 86, 2004.

[7] A. G. Phadke and J. S. Thorp, "HISTORY AND APPLICATIONS OF PHASOR MEASUREMENTS," in *2006 IEEE PES Power Systems Conference and Exposition,* 2006, pp. 331-335.

[8] A. G. Phadke, "The Wide World of Wide-area Measurement," *IEEE Power and Energy Magazine,* vol. 6, no. 5, pp. 52-65, 2008.

[9] A. Silverstein. Electric Power Systems and GPS [Online]. Available: https://www.gps.gov/cgsic/meetings/2016/silverstein.pdf

[10] A. Silverstein. Synchrophasors and the grid [Online]. Available: https://www.energy.gov/sites/prod/files/2017/09/f36/2_Modern%20Grid-networked%20Measurement%20and%20Monitoring%20Panel%20-%20Alison%20Silverstein%2C%20NASPI.pdf

[11] J. Zhu, E. Zhuang, C. Ivanov, and Z. Yao, "A Data-Driven Approach to Interactive Visualization of Power Systems," *IEEE Transactions on Power Systems,* vol. 26, no. 4, pp. 2539-2546, 2011.

[12] T. J. Overbye and J. D. Weber, "Visualization of power system data," in *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*, 2000, p. 7 pp.: IEEE.

[13] J. Gronquist, W. Sethares, F. Alvarado, and R. Lasseter, "Animated vectors for visualization of power system phenomena," in *Proceedings of Power Industry Computer Applications Conference*, 1995, pp. 121-127.

[14] R. A. Becker, S. G. Eick, and A. R. Wilks, "Visualizing network data," *IEEE Transactions on Visualization and Computer Graphics,* vol. 1, no. 1, pp. 16-28, 1995.

[15] M. J. Laufenberg, "Visualization approaches integrating real-time market data," in *IEEE PES Power Systems Conference and Exposition, 2004.*, 2004, pp. 1550-1555 vol.3.

[16] T. J. Overbye, D. A. Wiegmann, A. M. Rich, and S. Yan, "Human factors aspects of power system voltage contour visualizations," *IEEE Transactions on Power Systems,* vol. 18, no. 1, pp. 76-82, 2003.

[17] S. J. S. Tsai, Z. Jian, Z. Yi, and L. Yilu, "Frequency visualization in large electric power systems," in *IEEE Power Engineering Society General Meeting, 2005*, 2005, pp. 1467-1473 Vol. 2.

[18] E. National Academies of Sciences and Medicine, *Enhancing the Resilience of the Nation's Electricity System*. National Academies Press, 2017.

[19] E. National Academies of Sciences and Medicine, *Analytic research foundations for the next-generation electric grid*. National Academies Press, 2016.

[20] A. G. Phadke, "Synchronized phasor measurements in power systems," *IEEE Computer Applications in Power*, vol. 6, no. 2, pp. 10-15, 1993.

[21] "IEEE Standard for Synchrophasor Data Transfer for Power Systems," *IEEE Std C37.118.2-2011 (Revision of IEEE Std C37.118-2005)*, pp. 1-53, 2011.

[22] P. NASPI, "PMU Data Quality: A Framework for the Attributes of PMU Data Quality and Quality Impacts to Synchrophasor Applications," 2017.

[23] A. R. Goldstein, D. Anand, and Y. Li-Baboud, "Investigation of PMU Response to Leap Second: 2015," 2015.

[24] J. Guo, "Data analytics and application developments based on synchrophasor measurements," 2016.

[25] D. Trudnowski. Properties of the Dominant Inter-Area Modes in the WECC Interconnect [Online]. Available: https://www.wecc.biz/Reliability/WECCmodesPaper130113Trudnowski.pdf

[26] G. Rogers, "Demystifying power system oscillations," *IEEE Computer Applications in Power*, vol. 9, no. 3, pp. 30-35, 1996.

[27] NERC, "Reliability guideline - forced oscillation monitoring & mitigation," 2017.

[28] Y. Zhang *et al.*, "Visualization of wide area measurement information from the FNET system," in *2011 IEEE Power and Energy Society General Meeting*, 2011, pp. 1-8.

[29] J. N. Bank, O. A. Omitaomu, S. J. Fernandez, and Y. Liu, "Extraction and visualization of power system interarea oscillatory modes," in *IEEE PES General Meeting*, 2010, pp. 1-7.

[30] R. M. Gardner, G. B. Jordan, and Y. Liu, "Wide-Area mode visualization strategy based on FNET measurements," in *2009 IEEE Power & Energy Society General Meeting*, 2009, pp. 1-6.

[31] A. G. Phadke and J. S. Thorp, *Synchronized phasor measurements and their applications*. Springer, 2008.

[32] A. G. Phadke, "Synchronized phasor measurements-a historical overview," in *IEEE/PES Transmission and Distribution Conference and Exhibition*, 2002, vol. 1, pp. 476-479 vol.1.

[33] J. Zhao, L. Zhan, Y. Liu, H. Qi, J. R. Garcia, and P. D. Ewing, "Measurement accuracy limitation analysis on synchrophasors," in *2015 IEEE Power & Energy Society General Meeting*, 2015, pp. 1-5.

[34] M. Brown, M. Biswal, S. Brahma, S. J. Ranade, and H. Cao, "Characterizing and quantifying noise in PMU data," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1-5.

[35] D. Macii, D. Fontanelli, G. Barchi, and D. Petri, "Impact of Acquisition Wideband Noise on Synchrophasor Measurements: A Design Perspective," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 10, pp. 2244-2253, 2016.

[36] Q. Zhang, V. Vittal, G. Heydt, Y. Chakhchoukh, N. Logic, and S. Sturgill, "The time skew problem in PMU measurements," in *2012 IEEE Power and Energy Society General Meeting*, 2012, pp. 1-6.

[37] Q. F. Zhang and V. M. Venkatasubramanian, "Synchrophasor time skew: Formulation, detection and correction," in *2014 North American Power Symposium (NAPS)*, 2014, pp. 1-6.

[38] T. Bi, J. Guo, K. Xu, L. Zhang, and Q. Yang, "The Impact of Time Synchronization Deviation on the Performance of Synchrophasor Measurements and Wide Area Damping Control," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1545-1552, 2017.

[39] C. Huang *et al.*, "Data quality issues for synchrophasor applications Part I: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 4, no. 3, pp. 342-352, 2016.

[40] D. P. Shepard, T. E. Humphreys, and A. A. Fansler, "Evaluation of the vulnerability of phasor measurement units to GPS spoofing attacks," *International Journal of Critical Infrastructure Protection*, vol. 5, no. 3-4, pp. 146-153, 2012.

[41] X. Jiang, J. Zhang, B. J. Harding, J. J. Makela, and A. D. Domı´nguez-Garcı´a, "Spoofing GPS Receiver Clock Offset of Phasor Measurement Units," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 3253-3262, 2013.

[42] W. Yao *et al.*, "A novel method for phasor measurement unit sampling time error compensation," in *2017 IEEE Power & Energy Society General Meeting*, 2017, pp. 1-1.

[43] P. Kansal and A. Bose, "Bandwidth and latency requirements for smart transmission grid applications," in *2013 IEEE Power & Energy Society General Meeting*, 2013, pp. 1-1.

[44] M. Asprou and E. Kyriakides, "The effect of time-delayed measurements on a PMU-based state estimator," in *2015 IEEE Eindhoven PowerTech*, 2015, pp. 1-6.

[45] B. Naduvathuparambil, M. C. Valenti, and A. Feliachi, "Communication delays in wide area measurement systems," in *System Theory, 2002. Proceedings of the Thirty-Fourth Southeastern Symposium on*, 2002, pp. 118-122: IEEE.

[46] J. D. Taft, "Grid architecture 2," Pacific Northwest National Laboratory (PNNL), Richland, WA (United States)2016.

[47] I. Idehen, Z. Mao, and T. Overbye, "An emulation environment for prototyping PMU data errors," in *2016 North American Power Symposium (NAPS)*, 2016, pp. 1-6.

[48] A. B. Birchfield, T. Xu, and T. J. Overbye, "Power Flow Convergence and Reactive Power Planning in the Creation of Large Synthetic Grids," *IEEE Transactions on Power Systems,* pp. 1-1, 2018.

[49] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid Structural Characteristics as Validation Criteria for Synthetic Networks," *IEEE Transactions on Power Systems,* vol. 32, no. 4, pp. 3258-3265, 2017.

[50] T. Xu, A. B. Birchfield, K. S. Shetye, and T. J. Overbye, "Creation of synthetic electric grid models for transient stability studies," in *The 10th Bulk Power Systems Dynamics and Control Symposium (IREP 2017)*, 2017.

[51] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating Statistically Correct Random Topologies for Testing Smart Grid Communication and Control Networks," *IEEE Transactions on Smart Grid,* vol. 1, no. 1, pp. 28-39, 2010.

[52] E. Cotilla-Sanchez, P. D. H. Hines, C. Barrows, and S. Blumsack, "Comparing the Topological and Electrical Structure of the North American Electric Power Infrastructure," *IEEE Systems Journal,* vol. 6, no. 4, pp. 616-626, 2012.

[53] G. A. Pagani and M. Aiello, "The power grid as a complex network: a survey," *Physica A: Statistical Mechanics and its Applications,* vol. 392, no. 11, pp. 2688-2700, 2013.

[54] Z. Wang, S. H. Elyas, and R. J. Thomas, "A novel measure to characterize bus type assignments of realistic power grids," in *2015 IEEE Eindhoven PowerTech*, 2015, pp. 1-6.

[55] T. ECEN. Electric Grid Test Case Repository [Online]. Available: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/

[56] P. Kundur *et al.*, "Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions," *IEEE Transactions on Power Systems,* vol. 19, no. 3, pp. 1387-1401, 2004.

[57] S. M. Amin and B. F. Wollenberg, "Toward a smart grid: power delivery for the 21st century," *IEEE Power and Energy Magazine,* vol. 3, no. 5, pp. 34-41, 2005.

[58] L. Xie, Y. Chen, and P. R. Kumar, "Dimensionality Reduction of Synchrophasor Data for Early Event Detection: Linearized Analysis," *IEEE Transactions on Power Systems,* vol. 29, no. 6, pp. 2784-2794, 2014.

[59] A. Tarali and A. Abur, "Bad data detection in two-stage state estimation using phasor measurements," in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, 2012, pp. 1-8.

[60] S. P. Teeuwsen and I. Erlich, "Neural network based multi-dimensional feature forecasting for bad data detection and feature restoration in power systems," in *2006 IEEE Power Engineering Society General Meeting*, 2006, p. 6 pp.

[61] N. H. Abbasy and W. El-Hassawy, "Power system state estimation: ANN application to bad data detection and identification," in *AFRICON, 1996., IEEE AFRICON 4th*, 1996, vol. 2, pp. 611-615: IEEE.

[62] NERC. 1996 System Disturbances [Online]. Available: https://www.nerc.com/pa/rrm/ea/System%20Disturbance%20Reports%20DL/1996SystemDisturbance.pdf

[63] J. F. Hauer, "Application of Prony analysis to the determination of modal content and equivalent models for measured power system response," *IEEE Transactions on Power Systems*, vol. 6, no. 3, pp. 1062-1068, 1991.

[64] J. F. Hauer, C. J. Demeure, and L. L. Scharf, "Initial results in Prony analysis of power system response signals," *IEEE Transactions on Power Systems*, vol. 5, no. 1, pp. 80-89, 1990.

[65] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 5, pp. 814-824, 1990.

[66] T. K. Sarkar and O. Pereira, "Using the matrix pencil method to estimate the parameters of a sum of complex exponentials," *IEEE Antennas and Propagation Magazine*, vol. 37, no. 1, pp. 48-55, 1995.

[67] L. L. Grant and M. L. Crow, "Comparison of Matrix Pencil and Prony methods for power system modal analysis of noisy signals," in *2011 North American Power Symposium*, 2011, pp. 1-7.

[68] A. R. Borden and B. C. Lesieutre, "Variable Projection Method for Power System Modal Identification," *IEEE Transactions on Power Systems*, vol. 29, no. 6, pp. 2613-2620, 2014.

[69] A. B. Birchfield and T. J. Overbye, "Convergence characteristics of the variable projection method for mode extraction," in *2017 IEEE Texas Power and Energy Conference (TPEC)*, 2017, pp. 1-6.

[70] S. Mohapatra and T. J. Overbye, "Fast modal identification, monitoring, and visualization for large-scale power systems using Dynamic Mode Decomposition," in *2016 Power Systems Computation Conference (PSCC)*, 2016, pp. 1-7.

[71] S. Yan and T. J. Overbye, "Visualizations for power system contingency analysis data," *IEEE Transactions on Power Systems,* vol. 19, no. 4, pp. 1859-1866, 2004.

[72] T. J. Overbye, J. D. Weber, and M. Laufenberg, "Visualization of flows and transfer capability in electric networks," *Urbana,* vol. 51, p. 61801, 1999.

[73] I. Idehen and T. J. Overbye, "PMU Time Error Detection Using Second-Order Phase Angle Derivative Measurements," in *2019 IEEE Texas Power and Energy Conference (TPEC)*, 2019, pp. 1-6.

[74] U. S. D. o. Energy, "Maintaining Reliability in the Modern Power System," 2016, Available: https://www.energy.gov/sites/prod/files/2017/01/f34/Maintaining%20Reliability%20in%20the%20Modern%20Power%20System.pdf.

[75] S. Wang, J. Zhao, Z. Huang, and R. Diao, "Assessing Gaussian Assumption of PMU Measurement Error Using Field Data," *IEEE Transactions on Power Delivery,* vol. 33, no. 6, pp. 3233-3236, 2018.

[76] J. E. Tate and T. J. Overbye, "Extracting steady state values from phasor measurement unit data using FIR and median filters," in *2009 IEEE/PES Power Systems Conference and Exposition*, 2009, pp. 1-8.

[77] M. Klein, G. J. Rogers, and P. Kundur, "A fundamental study of inter-area oscillations in power systems," *IEEE Transactions on Power Systems,* vol. 6, no. 3, pp. 914-921, 1991.

[78] G. Ghanavati, P. D. H. Hines, and T. I. Lakoba, "Identifying Useful Statistical Indicators of Proximity to Instability in Stochastic Power Systems," *IEEE Transactions on Power Systems,* vol. 31, no. 2, pp. 1360-1368, 2016.

[79] M. Wu and L. Xie, "Online Detection of Low-Quality Synchrophasor Measurements: A Data-Driven Approach," *IEEE Transactions on Power Systems,* vol. 32, no. 4, pp. 2817-2827, 2017.

[80] D. Shi, D. J. Tylavsky, and N. Logic, "An Adaptive Method for Detection and Correction of Errors in PMU Measurements," *IEEE Transactions on Smart Grid,* vol. 3, no. 4, pp. 1575-1583, 2012.

[81] L. Vanfretti, S. Bengtsson, and J. O. Gjerde, "Preprocessing synchronized phasor measurement data for spectral analysis of electromechanical oscillations in the Nordic Grid," *International Transactions on Electrical Energy Systems,* vol. 25, no. 2, pp. 348-358, 2015.

[82] P. Gao, M. Wang, S. G. Ghiocel, and J. H. Chow, "Modeless reconstruction of missing synchrophasor measurements," in *2014 IEEE PES General Meeting | Conference & Exposition*, 2014, pp. 1-5.

[83] R. M. Gardner, "Conditioning of FNET data and triangulation of generator trips in the eastern interconnected system," 2005.

[84] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[85] P. J. Brockwell, R. A. Davis, and M. V. Calder, *Introduction to time series and forecasting*. Springer, 2002.

[86] Y. Pei and O. Zaïane, "A synthetic data generator for clustering and outlier analysis," *Department of Computing science, University of Alberta, edmonton, AB, Canada,* 2006.

[87] D. R. Jeske *et al.*, "Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems," in *Proceedings of the eleventh ACM SIGKDD*

international conference on Knowledge discovery in data mining, 2005, pp. 756-762: ACM.

[88] G. Albuquerque, T. Lowe, and M. Magnor, "Synthetic Generation of High-Dimensional Datasets," *IEEE Transactions on Visualization and Computer Graphics,* vol. 17, no. 12, pp. 2317-2324, 2011.

[89] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach," *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 30-39, 2004.

[90] J. Verzani, *Using R for introductory statistics*. Chapman and Hall/CRC, 2014.

[91] X. Zheng, B. Wang, and L. Xie, "Synthetic Dynamic PMU Data Generation: A Generative Adversarial Network Approach," *arXiv preprint arXiv:1812.03203,* 2018.

[92] P. W. Sauer and M. A. Pai, *Power system dynamics and stability*. Prentice hall Upper Saddle River, NJ, 1998.

[93] D. K. Ranaweera, G. G. Karady, and R. G. Farmer, "Economic impact analysis of load forecasting," *IEEE Transactions on Power Systems,* vol. 12, no. 3, pp. 1388-1392, 1997.

[94] P. S. M. T. Dept. (2014). *Fundamentals of Transmission Operations*. Available: https://www.pjm.com/-/media/training/nerc-certifications/trans-exam-materials/foto/foto-lesson4-load-forecasting-and-weather.ashx?la=en

[95] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," *IEEE Transactions on Power Systems,* vol. 16, no. 3, pp. 498-505, 2001.

[96] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," in *2018 IEEE Texas Power and Energy Conference (TPEC)*, 2018, pp. 1-6.

[97] Y. Qiu, J. Zhao, and H. D. Chiang, "Effects of the stochastic load model on power system voltage stability based on bifurcation theory," in *2008 IEEE/PES Transmission and Distribution Conference and Exposition*, 2008, pp. 1-6.

[98] V. S. Perić and L. Vanfretti, "Power-System Ambient-Mode Estimation Considering Spectral Load Properties," *IEEE Transactions on Power Systems*, vol. 29, no. 3, pp. 1133-1143, 2014.

[99] H. Banakar, C. Luo, and B. T. Ooi, "Impacts of Wind Power Minute-to-Minute Variations on Power System Operation," *IEEE Transactions on Power Systems*, vol. 23, no. 1, pp. 150-160, 2008.

[100] E. A. DeMeo, W. Grant, M. R. Milligan, and M. J. Schuerger, "Wind plant integration [wind power plants]," *IEEE Power and Energy Magazine*, vol. 3, no. 6, pp. 38-46, 2005.

[101] D. O. Koval and A. A. Chowdhury, "Assessment of transmission line common mode, station originated and fault types forced outage rates," in *Conference Record 2009 IEEE Industrial & Commercial Power Systems Technical Conference*, 2009, pp. 1-7.

[102] B. P. Administration. (2019, 02-Feb-2019). *Outage and Reliability Reports*. Available: https://transmission.bpa.gov/business/operations/outages/

[103] P. Simulator, "PowerWorld Corporation," ed: October, 2005.

[104] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[105] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.

[106] P. M. Ashton, G. A. Taylor, M. R. Irving, I. Pisica, A. M. Carter, and M. E. Bradley, "Novel application of detrended fluctuation analysis for state estimation using synchrophasor measurements," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1930-1938, 2013.

[107] M. Khan, M. Li, P. Ashton, G. Taylor, and J. Liu, "Big data analytics on PMU measurements," in *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2014, pp. 715-719.

[108] M. Gupta, A. B. Sharma, H. Chen, and G. Jiang, "Context-aware time series anomaly detection for complex systems," in *Workshop Notes*, 2013, vol. 14.

[109] J. W. Pierre, D. J. Trudnowski, and M. K. Donnelly, "Initial results in electromechanical mode identification from ambient data," *IEEE Transactions on Power Systems*, vol. 12, no. 3, pp. 1245-1251, 1997.

[110] I. Idehen and T. Overbye, "A similarity-based PMU error detection technique," in *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*, 2017, pp. 1-6.

[111] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *ACM sigmod record*, 2000, vol. 29, no. 2, pp. 93-104: ACM.

[112] Z. Jin, E. Nobile, A. Bose, and K. Bhattacharya, "Localized reactive power markets using the concept of voltage control areas," in *2006 IEEE Power Engineering Society General Meeting*, 2006, p. 1 pp.

[113] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14, pp. 281-297: Oakland, CA, USA.

[114] D. Rafiei and A. Mendelzon, "Similarity-based queries for time series data," in *ACM SIGMOD Record*, 1997, vol. 26, no. 2, pp. 13-25: ACM.

[115] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems,* vol. 7, no. 3, pp. 358-386, 2005.

[116] M. Müller, *Information retrieval for music and motion*. Springer, 2007.

[117] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis,* vol. 11, no. 5, pp. 561-580, 2007.

[118] B. C. Giao and D. T. Anh, "Similarity search in multiple high speed time series streams under Dynamic Time Warping," in *Information and Computer Science (NICS), 2015 2nd National Foundation for Science and Technology Development Conference on*, 2015, pp. 82-87: IEEE.

[119] "IEEE Standard for Synchrophasor Measurements for Power Systems," *IEEE Std C37.118.1-2011 (Revision of IEEE Std C37.118-2005)*, pp. 1-61, 2011.

[120] F. Belmudes, D. Ernst, and L. Wehenkel, "Pseudo-Geographical Representations of Power System Buses by Multidimensional Scaling," in *2009 15th International Conference on Intelligent System Applications to Power Systems*, 2009, pp. 1-6.

[121] P. Cuffe and A. Keane, "Visualizing the Electrical Structure of Power Systems," *IEEE Systems Journal,* vol. 11, no. 3, pp. 1810-1821, 2017.

[122] A. K. Barnes and J. C. Balda, "Placement of distributed energy storage via multidimensional scaling and clustering," in *2014 International Conference on Renewable Energy Research and Application (ICRERA)*, 2014, pp. 69-74.

[123] R. Hoffmann, F. Promel, F. Capitanescu, G. Krost, and L. Wehenkel, "Situation adapted display of information for operating very large interconnected grids," in *2011 IEEE Trondheim PowerTech*, 2011, pp. 1-7.

[124] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[125] M. Togerson's Classical, "Derivation," ed.

[126] C.-h. Chen, W. K. Härdle, and A. Unwin, *Handbook of data visualization*. Springer Science & Business Media, 2007.

[127] S.-S. Choi, S.-H. Cha, and C. C. Tappert, "A survey of binary similarity and distance measures," *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43-48, 2010.

[128] F. Lourenco, V. Lobo, and F. Bacao, "Binary-based similarity measures for categorical data and their application in Self-Organizing Maps," 2004.

[129] W. Trinh, K. Shetye, I. Idehen, and T. Overbye, "Iterative Matrix Pencil Method for Power System Modal Analysis," in *2019 52nd Hawaii International Conference on System Sciences (HICSS)*, Maui, HI, 2019.

[130] T. J. Overbye, E. M. Rantanen, and S. Judd, "Electric power control center visualization using Geographic Data Views," in *2007 iREP Symposium - Bulk Power System Dynamics and Control - VII. Revitalizing Operational Reliability*, 2007, pp. 1-8.

[131] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.

[132] E. R. Tufte, *Envisioning information*. Graphics Press, 1990.

[133] I. Idehen, B. Wang, K. Shetye, T. Overbye, and J. Weber, "Visualization of Large-Scale Electric Grid Oscillation Modes," in *2018 North American Power Symposium (NAPS)*, 2018, pp. 1-6.

[134] D. Weiskopf, *Vector Field Visualization*. Springer, 2007.

[135] D. H. Laidlaw *et al.*, "Comparing 2D vector field visualization methods: a user study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, pp. 59-70, 2005.

[136] R. Podmore, "Identification of Coherent Generators for Dynamic Equivalents," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-97, no. 4, pp. 1344-1354, 1978.

[137] A. Gallai and R. Thomas, "Coherency identification for large electric power systems," *IEEE Transactions on Circuits and Systems*, vol. 29, no. 11, pp. 777-782, 1982.

[138] S. B. Yusof, G. J. Rogers, and R. T. H. Alden, "Slow coherency based network partitioning including load buses," *IEEE Transactions on Power Systems*, vol. 8, no. 3, pp. 1375-1382, 1993.

[139] L. J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring expression data: identification and analysis of coexpressed genes," *Genome research*, vol. 9, no. 11, pp. 1106-1115, 1999.

[140] S. Dutta and T. J. Overbye, "Feature Extraction and Visualization of Power System Transient Stability Results," *IEEE Transactions on Power Systems*, vol. 29, no. 2, pp. 966-973, 2014.

[141] Quality Threshold (QT) Clustering [Online]. Available: https://www.chem-agilent.com/cimg/qt_clustering.pdf

[142] L. Chen, Y. Min, and W. Hu, "An energy-based method for location of power system oscillation source," *IEEE Transactions on Power Systems,* vol. 28, no. 2, pp. 828-836, 2013.

[143] S. Maslennikov, B. Wang, and E. Litvinov, "Locating the source of sustained oscillations by using PMU measurements," in *2017 IEEE Power & Energy Society General Meeting*, 2017, pp. 1-5.

[144] W. Bin and S. Kai, "Location methods of oscillation sources in power systems: a survey," *Journal of Modern Power Systems and Clean Energy,* vol. 5, no. 2, pp. 151-159, 2017.

[145] PJM. (2018, 08-Nov-2018). *Private generic PMU data*. Available: https://www.pjm.com/markets-and-operations/advanced-tech-pilots/private-generic-pmu-data.aspx

APPENDIX A

SYNTHETIC 2,000-BUS AND 10,000-BUS NETWORKS

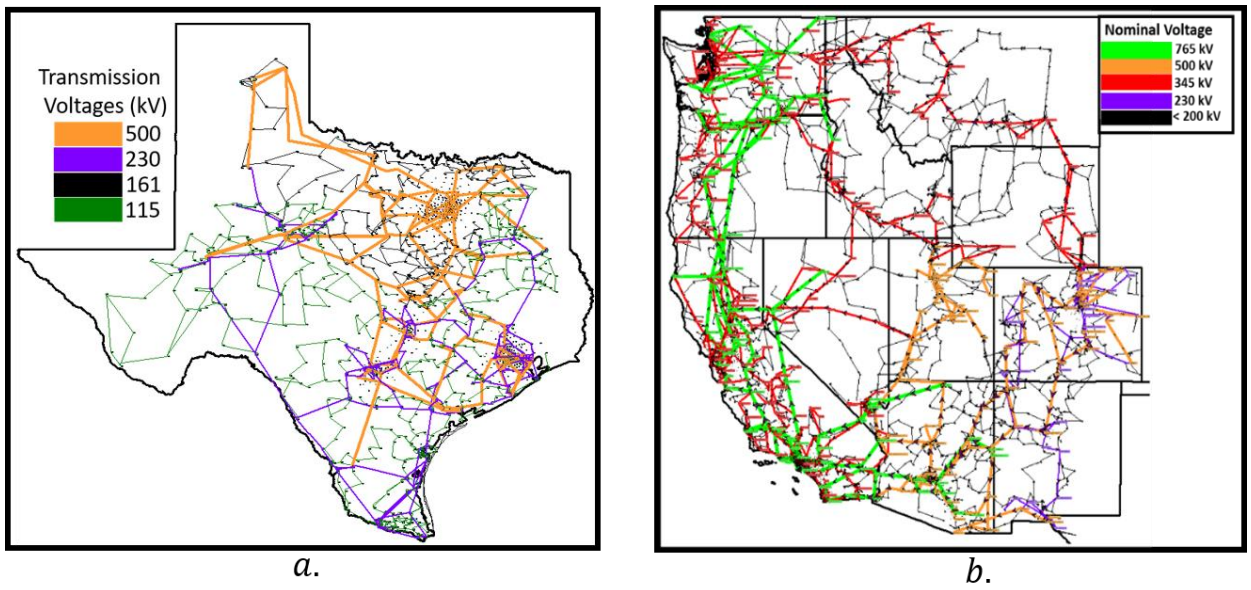


Fig. A-1. (a) 2,000-bus network (b) 10,000-bus network

The 2,000- and 10,000-bus networks are artificially-created grids covering the geographical space of the state of Texas, and the interconnected grid of Western U.S. respectively. Relevant details used for this work are shown in Table A-1.

Table A-1. Network Information

	2,000-bus	10,000-bus
# of substations	1,250	4,762
# of generators	432	1,937
# of transmission lines	3,209	12,706
Operating frequency	60 Hz	
PMU report rate	30 samples/ second	

The contingency cases that have been used in this work are defined in Table A-2.

Table A-2. Contingency Cases

Synthetic Network		
Case	**2,000-bus**	**10,000-bus**
1	1. Outage of 230-kV line. 2. Duration is 3 seconds.	1. Five outaged generators 2. Duration is 10 seconds
2	1. Outage of one 115-kV line. 2. Duration is 10 seconds.	1. Five outaged generators 2. 19 deactivated stabiizers 3. Duration is 10 seconds
3	1. Two outaged generators. 2. Duration is 10 seconds.	
4	1. One of the system largest generators whose model is changed to classic mode (GENCLS). 2. Generator damping is set to -1. 3. Outage of 500-kV line with large MVA flow. 4. Duration is 10 seconds.	
5	1. Inactive stabilizers. 2. Two outaged generator.	

APPENDIX B

REAL PMU DATA AND SAMPLES OF SYNTHETIC DATASETS

Real Dataset

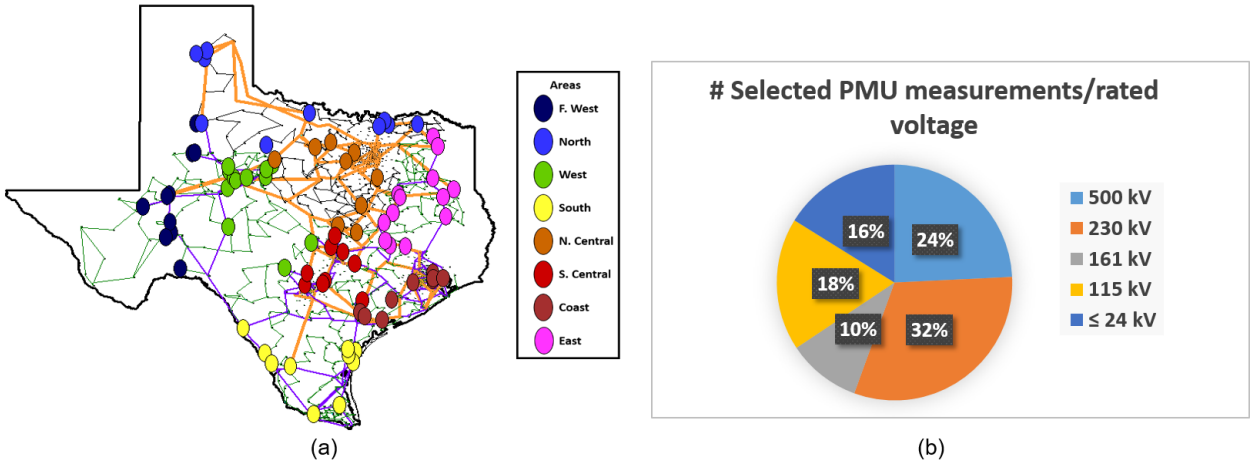
The industry-grade synchrophasor dataset used for the analysis was obtained from a public repository [145] managed by the Pennsylvania-New Jersey-Maryland (PJM) Interconnection in November 2018. No other information was provided. The dataset comprises of 30-minute duration of voltage magnitude, angle and frequency measurements obtained from over 133 PMUs with report rates of 30 samples per second. However, only 123 measurements have been used

Synthetic Dataset

Real, large-scale, power systems consist of thousands of buses, and as a result, it is unrealistic to individually monitor these different buses. However, to ensure full system observability, PMUs will often be distributed across different portions of the grid. To ensure a realistic analysis, only a proportion of the total measurements in the synthetic 2,000-bus case has been used. A total of ninety-nine PMU measurements, each with a report rate of 30 samples per second, have been selected according to the listed criteria thus,

1. All source measurement locations span the eight pre-defined geographical areas of the grid.
2. All measurements cover the spectrum of the different nominal grid voltages.
3. As much as possible, measurements from each area has been distributed among all nominal voltages available in that region.

Fig. B-1 (a) and (b) shows the distribution of the selected grid locations, and a statistics of their nominal voltage values.



(a) (b)

Figure B-1. Distribution of selected source PMUs and their nominal voltages

APPENDIX C

Given objects $p, o \in D, o' \in D\backslash\{p\}$; D= Total space; p = Object of interest ; $k, MinPts > 0$

Step 1. Compute k-distance, $k\text{-}dist(p)$

$$k\text{-}dist(p) = d(p,o); \begin{cases} d(p,o') \leq d(p,o); for\ at\ least\ k\ objects \\ d(p,o') \leq d(p,o);\ for\ at\ most\ (k-1)\ objects \end{cases}$$

Step 2. Compile k-nearest neighbors, $N_{k-dist}(p)$

$$N_{k-dist}(p) = \ o \in D\backslash\{p\}; d(p,o) \leq \ k\text{-}dist(p)$$

Step 3. Compute reachability distance, $reach\text{-}dist_k(p,o)$

$$reach\text{-}dist_k(p,o) = \max(k - dist(o), d(p,o))$$

Step 4. Compute reachability density , $lrd_{MinPts}(p)$

$$lrd_{MinPts}(p) = 1/\left[\frac{\sum_{o \in N_{MinPts(p)}} reach - dist_{MinPts}(p,o)}{|N_{MinPts}(p)|}\right]$$

Step 5. Compute local outlier factor, $LOF_{MinPts}(p)$

$$lrd_{MinPts}(p) = 1/\left[\frac{\sum_{o \in N_{MinPts(p)}} reach - dist_{MinPts}(p,o)}{|N_{MinPts}(p)|}\right]$$

A small LOF value (~ 1.0) is indicative of a high density neighborhood around an object. Large LOF values are associated with a sparse neighborhood, and typical of an outlier object [10]. For this work, k and $MinPts$ parameters are set to both equal 50% of the total measurements (or objects) in the dataset being considered.

APPENDIX D

Classical MDS

Given n objects, the pairwise dissimilarity between any i^{th} and j^{th} object (computed as the distance between them, d_{ij}) set in a proximity matrix, $\boldsymbol{D}$

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{bmatrix}$$

The classical MDS finds the coordinates of each object in a τ-geometrical space such that the pairwise distance between objects is preserved

Step 1. Compute matrix of squared proximity, $\boldsymbol{D^2}$

Step 2: Apply double centering, $\boldsymbol{B} = -\frac{1}{2}\boldsymbol{P}\boldsymbol{D^2}\boldsymbol{P}$ **;**

$$\boldsymbol{P} = \boldsymbol{I} - \tfrac{1}{n}(\mathbf{1}\mathbf{1}^T)$$

$\boldsymbol{P}$ **=** center matrix; $\boldsymbol{I}$ = identity matrix; **1** = row vector, $(\in R^n)$

Step 3: Perform eigen-decomposition of $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = I$$

Step 4: Compute coordinates $\boldsymbol{X} = \mathbf{Q}_+\boldsymbol{\Lambda}_+^{1/2}$

$\boldsymbol{\Lambda}_+$=Matrix of first m eigenvalues greater than zero; $\mathbf{Q}_+$=Corresponding first m columns

APPENDIX E

Given an actual measurement $y(t)$ comprising of α equally spaced samples, and with the dc offset removed, MPA fits a set of (un)damped sinusoidal signals to yield $\hat{y}(t)$ which is an approximate of $y(t)$.

$$\hat{y}(t) = \sum_{j=1}^{q} A_j e^{\sigma_j t} \cos(\omega_j t + \phi_j) \qquad \text{(A.1)}$$

The MPA performs a singular value decomposition technique on a Hankel matrix which comprises of the data points in $y(t)$. A user-provided threshold is used to determine the number of signal modes (q) by retaining singular values greater than the threshold value. Afterwards, a generalized eigenvalue solution is applied to obtain q discrete-time poles, and then used to compute σ_j.

Step 1. Generate the Hankel matrix:

$$[\boldsymbol{H}] = \begin{bmatrix} y(0) & y(1) & \cdots & y(L) \\ y(1) & y(2) & \cdots & y(L+1) \\ \vdots & \vdots & \ddots & \vdots \\ y(\alpha - L - 1) & y(\alpha - L) & \cdots & y(\alpha - 1) \end{bmatrix} \qquad (A.2)$$

L is a pencil parameter which is used to eliminate the effects of noise in the data

Step 2. Perform a singular value decomposition on $\boldsymbol{H}$

$$[\boldsymbol{H}] = [\boldsymbol{U}][\mathbf{E}][\boldsymbol{V}]^T \qquad (A.3)$$

$[\boldsymbol{U}]$ and $[\boldsymbol{V}]$ are unitary matrices comprising of the eigenvectors of $[\boldsymbol{H}][\boldsymbol{H}]^T$ and $[\boldsymbol{H}]^T[\boldsymbol{H}]$ respectively, and $[\mathbf{E}]$ is a diagonal matrix containing the singular values of $[\boldsymbol{H}]$ in descending order.

Step 3. Decide on a choice of q

The ratio of each singular value (σ_c) to the largest one (σ_{max}) compared to a threshold value determines the retained and eliminated signal modes. Consider a singular value σ_c, then

$$\frac{\sigma_c}{\sigma_{max}} \sim 10^{-p} \qquad (A.4)$$

where p is the number of significant decimal digits in the data. Assuming p is set to to be accurate to 3 significant digits, then the singular values for which the ratio in (A.3) is below 10^{-3} are assumed to be part of the signal noise, and not included in the reconstructed signal.

Step 4. Extract matrices for the generalized eigenvalue process

From the filtered matrix, $[\boldsymbol{V}] = [v_1, v_2 \dots v_q]$, delete last and first columns in $[\boldsymbol{V}]$ to obtain $[\boldsymbol{V_1}] = [v_1, v_2 \dots v_{q-1}]$ and $[\boldsymbol{V_2}] = [v_2, v_3 \dots v_q]$ respectively. Define two matrices $\boldsymbol{Y_1}$ **and** $\boldsymbol{Y_2}$ such that,

$$\boldsymbol{Y_1} = [\boldsymbol{U}][\boldsymbol{E'}][\boldsymbol{V_1}]^T, \qquad \boldsymbol{Y_2} = [\boldsymbol{U}][\boldsymbol{E'}][\boldsymbol{V_2}]^T$$

$\boldsymbol{E'}$ comprises of the first q columns of $[\mathbf{E}]$, and corresponds to the q dominant singular values.

Step 5. Compute all complex eigenvalues σ_j from the generalized eigenvalue solutions for the matrix pair $(\boldsymbol{Y_1}, \boldsymbol{Y_2})$ i.e. $|\boldsymbol{Y_2} - \boldsymbol{\lambda Y_1}| = \mathbf{0}$

Step 6. Compute the mode shape amplitudes from the residue vector

$$\begin{bmatrix} {\sigma_1}^0 & \dots & {\sigma_q}^0 \\ \vdots & \ddots & \vdots \\ {\sigma_1}^{\alpha-1} & \dots & {\sigma_q}^{\alpha-1} \end{bmatrix} \begin{bmatrix} A_1 \\ \vdots \\ A_q \end{bmatrix} = \begin{bmatrix} y(0) \\ \vdots \\ y(\alpha-1) \end{bmatrix}$$