21, rue d'Artois, F-75008 PARIS
http : //www.cigre.org

CIGRE US National Committee
2014 Grid of the Future Symposium

Assessment of Impact of Data Quality on PMU-Based Applications

S. VEDA[1] N.R. CHAUDHURI[2] C.A. BAONE[1] N. ACHARYA[1]
[1]GE Global Research, [2]North Dakota State University
USA

SUMMARY

Several wide-area applications are being proposed for protection, operation and control of electric grids. One of the major impediments towards the successful implementation of such applications is the data quality issues associated with the current synchrophasor architecture. In this paper, the state-of-the-art on synchrophasor architecture is presented and the sources of bad data in the current architecture have been identified. A generalized framework for assessing the data quality of synchrophasor measurements has been proposed. The framework presents five dimensions for measuring data quality. These dimensions have been defined so as to represent the multitude of factors that impact data quality.

The focus of the paper is to study the impact of data quality on small-signal oscillation monitoring algorithms. The effect of different characteristics of bad data samples such as the magnitude, persistence and proximity to system events on the accuracy of damping estimation using Matrix Pencil algorithm is studied. The results from an SVD-based tool and a PCA-based tool for bad data detection are also presented. Synthetic data generated by time domain simulation have been used for the studies.

KEYWORDS

Bad data detection, data quality assessment, synchrophasor, PMU data, SVD, PCA.

veda@ge.com

The authors gratefully acknowledge support provided for this work under the Technology Innovation Program (TIP-306) administered through the Bonneville Power Administration. Contract No. 00062995

1. INTRODUCTION

The concept of synchronized phasor measurements originated from the development of Symmetrical Component Distance Relaying in the 1970s. With the wider availability of GPS time signals, it became possible to provide time tags to the synchronized measurements enabling them to be transmitted across longer distances. By collating these time-synchronized phasor measurements from wide areas at a central station, a complete snapshot of the system can be discerned. [1]

The synchrophasor measuring devices called the Phasor Measurement Units (PMUs) have transitioned from being stand-alone devices to being integrated into almost every modern substation device like the digital relays and fault recorders. In recent years, various applications that rely on high speed PMU data have been proposed. These include oscillation detection, plant model validation, post-event analysis, static and dynamic state estimation, wide area visualization, wide area damping control among others. Some of these applications have already made its way to the utility control room. With increasing reliance on the PMUs for system monitoring and control, the quality of data being received at the Phasor data Concentrator (PDC) has to be managed. In this context, this paper identifies data quality issues associated with PMU measurement and provides a framework for assessing the impact of bad data quality on an oscillation monitoring application

2. STATE-OF-THE-ART AND NEED FOR BETTER DATA QUALITY

The PMU data is typically steamed at a rate of 30 to 60 samples per second. When there are multiple PMUs in a substation, a PDC collects these data, aligns them by time stamps and sends them upstream. A Super PDC collects data from several PDCs at the control center level. A synchrophasor system [2] can consist of several layers of PDCs and Super PDCs depending on the size of the monitored network. Given the large number of such measurement devices and the higher data reporting rates, the amount of data that is collected easily runs into few Terabytes a day for a typical utility.

The quality of these measurement data is impacted by the presence of atypical data like missing data, corrupted data or outliers. Missing data is associated with signal loss and are usually flagged by the PMU or the PDC. Missing data could occur at intermittent samples or for a long duration of time in a single or multiple channels depending on the root cause. Some of the causes include loss of one or more components like the communication links, network routers and synchrophasor devices and bad configuration of these devices. Outliers refer to data that is significantly different from the normal measurement. Unlike missing data, the presence of outliers cannot be easily detected and hence would require more sophisticated techniques to identify them. Since outliers are not flagged, the applications that use the measurements should incorporate outlier detection techniques into their algorithms. Outliers can be caused by temporary sensor failure or interference. Interestingly, outliers in power system can also be caused by faults or other system disturbances. Other data quality issues stem from incorrect GPS clock settings, server overload, aliasing at the PDC and latency. The sources of bad data are shown in Figure 2.1

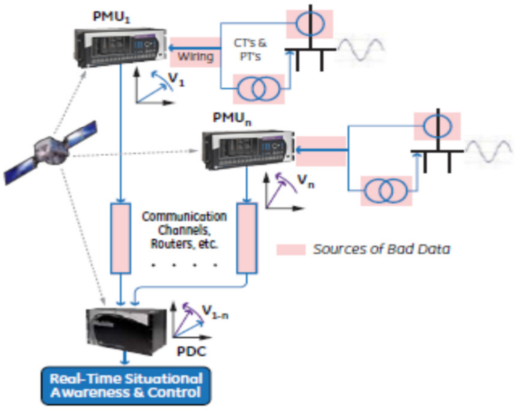
PMU1
Wiring
CT's & PT's
V1
PMUn
Vn
Communication Channels, Routers, etc.
Sources of Bad Data
V1-n
PDC
Real-Time Situational Awareness & Control

Figure 2.1 Typical WAMS Infrastructure with Sources of Bad Data

Several of the proposed techniques for identifying and characterizing data quality employ state estimation at different levels for detecting and flagging missing streams of data [3 – 6]. In [3], a data preprocessor sanitizes the data such that when a few samples are missing, linear interpolation is used and when a channel of data is missing the remaining channels are used. Recursive Least Squares (RLS) algorithms used for oscillation monitoring in [7,8] use prediction errors for flagging atypical data. Literature on the assessment and impact of data quality on synchrophasor-based applications is limited.

Data Quality is one of the concerns for utilities and Independent System Operators (ISOs) that seek to leverage synchrophasor measurements into their operations. Some utilities/ISOs report that only 91% of the data is acceptable [9]. Unlike traditional state estimators, PMU-based linear estimators measure the system state variables directly and hence there's greater emphasis on the need for accuracy of the measurements.

3. FRAMEWORK FOR ASSESSMENT OF DATA QUALITY:

Any implementation of the Wide-Area Monitoring & Control Systems requires a reliable and accurate measurement infrastructure. As the application space for synchrophasors increase, there will be a need for a framework for the assessment of data quality. Such a framework does not exist today. In [10], five dimensions of data quality are defined for Electronic Health Recording (EHR) in the field of medical instrumentation. A similar approach is proposed for evaluating synchrophasor data quality. The dimensions of data quality presented below can describe the reliability of the communication infrastructure accounting for the actual power system applications that use the measurements.

- **Completeness** refers to the characteristic of the measurements or the set of measurements to adequately define the system state. This feature would depend on the application that would actually use the data. A set of measurements that are adequate for one application may not be adequate for another application. While the completeness is accounted for during the deployment of an application, missing data streams or high noise content may render the measurements incomplete.
- **Correctness** is a measure of accuracy of the measurements. It is a characteristic that describes how well the measurements reflect the true value of the system state that is measured. While data outliers and corrupted data represent a significant departure from correctness, the errors associated with instrumentation transformers, GPS clock errors and poor calibration of devices also impact the correctness of data.
- **Concordance** is the agreement between the same data stream as it gets transferred between different devices. A typical path of a PMU data after it is generated by the PMU includes a combination of network routers and switches, communication channels, PDCs and SuperPDCs before it can be used by an application. Concordance is a measure of the reproducibility of the different devices and media of the communication architecture.
- **Plausibility** is the trustworthiness or the confidence of the measurement data. Unlike the previous dimensions where they describe the individual measurement streams, plausibility is a characteristic of the data stream in the light of information provided by the other measurements.
- **Currency** is the relevance of the measurement in describing the system state at a given time. It is also referred to as recency and timeliness. Currency requirements will depend on the nature of the application.

While the above framework is generalized for assessing data quality of such an infrastructure, the actual impact is application-specific and has to be analyzed on a case-by-case basis. The following section provides the impact of atypical data on a typical small signal oscillation algorithm, which is the focus of this paper.

4. OSCILLATION MONITORING AND DATA QUALITY

Small Signal Stability issues have been a persistent problem in large interconnected power systems. Given the wide-area nature of this phenomenon and the complex underlying dynamics, oscillation detection is a power system application that can greatly benefit from a robust WAMS architecture. With the time scale of control for such an instability problem being a few tens of seconds to minutes, the requirements on communication infrastructure is not as stringent as it is for protection and other control applications.

Power Oscillation Monitoring has been one of the key challenges facing present-day utilities. Unsurprisingly, it is one of the first application areas for PMUs. Several utilities have deployed some form of oscillation detection algorithm that helps to monitor potential small signal instabilities in their networks. Several of the control decisions made during poor damping situations, however is based on operators' knowledge and historical precedence in the form of standing dispatch orders. The accuracy of estimation of oscillations is very critical for such a control paradigm. In this section, a formulation of the matrix pencil method is presented. The matrix pencil algorithm is one of the widely used methods for oscillation detection. The objective is to analyze the impact of atypical data in the input data stream on the output of the algorithm.

4.1 Matrix Pencil Algorithm

In this study, Matrix Pencil algorithm is the chosen as the oscillation detection algorithm. This technique [11] [12] is robust to noise in the measured data and is widely used in electromagnetic applications like radar and antenna response analysis. Matrix Pencil is a block processing algorithm that uses ringdown data after a disturbance and has been used in field deployments as well.

A given time varying signal can be decomposed into a set of sinusoids that represent the modes of oscillation present in the given signal. Given an observed power systems signal, $y(t)$ with measurement noise $n(t)$, the actual signal is represented as $x(t)$. This signal $x(t)$ can be approximated to a sum of complex exponentials as $y(t) = x(t) + n(t) = \sum_{i=1}^{M} R_i \exp(S_i t) + n(t), \quad 0 \le t \le T$

Where $y(t)$ is the measured response; $x(t)$ is the signal; $n(t)$ is the measured noise; R_i represents the residues, S_i represents the complex poles, M is the order. Since the signal data is in discrete time due to sampling (say, at a rate $1/Ts$), t is replaced with kTs and the above equation is written as $y(kTs) = \sum_{i=1}^{M} R_i Z_i^k + n(kTs), \quad k = 0 \; to \; N-1$

Where Zi represents the discrete poles and N is the number of samples. The parameters of the sequence of complex exponentials described above, namely M, Zi and R_i, can be solved using matrix pencil method. Pencil of functions is formulated as, $\mathrm{f}(\mathrm{t}, \lambda) = \mathrm{g}(\mathrm{t}) + \lambda\, \mathrm{h}(\mathrm{t})$

By selecting λ, $\mathrm{g}(\mathrm{t})$ and $\mathrm{h}(\mathrm{t})$ appropriately using the given $y(t)$, modal information (from the complex poles Zi) can be extracted from $\mathrm{f}(\mathrm{t}, \lambda)$. Given the total number of samples (N) of a noisy measured signal (y), Singular Value Decomposition is used for pre-filtering the data to produce Y1 and Y2 matrices as described in [11]. Eigen decomposition is performed on the matrix pair to extract the complex poles, yielding the modal damping and frequencies. The modal estimation is performed on the given data using a 10-second moving window to enable the study the impact of bad data on modal estimation on a sample by sample basis.

4.2 Impact of Bad Data on Oscillation Monitoring

The Matrix Pencil algorithm is tested using data from the simulation of a 4-machine 2-area test system [13] widely used to study small-signal oscillation problems. The real power flow across the tie line is measured at a sampling rate of 50 samples per second. In order to simulate the system noise, the loads are injected with zero-mean white Gaussian noise during the simulation. The recorded data includes ringdown data after a self-clearing bus fault.

In order to study the impact of atypical data, another data stream is created using the same dataset as above and a few missing samples are introduced. It is assumed that in the absence of data samples, the PDC will substitute the missing samples with the most recent good sample until the data samples are available again. This assumption is based on a survey where it was found that some PDCs “will repeat old values to fill in missing data” [14]. With the continuous evolution of the synchrophasor standards, this way of handling missing data may change. Some of the options include setting the values for missing samples to a standard value and flagging the data bits accordingly. Figure 4.1 illustrates the impact of missing data on oscillation characteristics.

It can be seen that the missing data greatly influences the accuracy of modal estimation. It should be noted that bad data makes the estimated damping ratio to breach the typical 0.5% threshold for alarm indicating a small-signal instability while in reality the system is perfectly stable. Even if this estimation is discarded on account of it being from a missing data, it can be seen that the influence of the missing data on the modal results persists for a much longer duration of time (about the length of moving window – 10 seconds), rendering the estimation during this time unusable. If the measurement data of a critical signal or of a large number of signals were lost, it could greatly undermine the ability to effectively monitor the oscillatory behavior of the system.

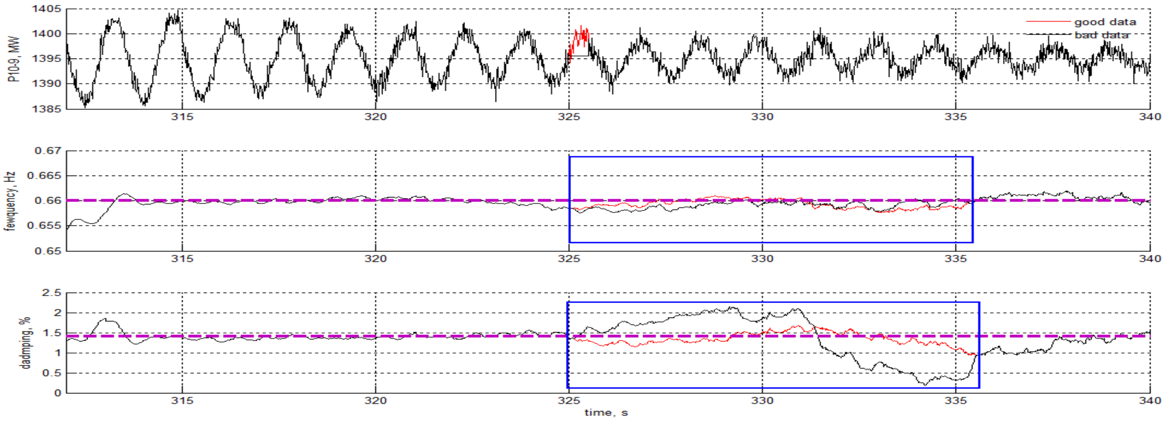

Figure 4.1: Impact of Missing Data on Oscillation Monitoring

Figure 4.2 shows the results for the same data stream with an outlier deliberately injected at 325th second. As described above, it can be seen that the influence of the bad data persists for the entire duration of moving window following its occurrence. It can also be observed that the damping estimation repeatedly hits the 0.5% damping ratio threshold for initiating alarms, while in reality the damping is much higher.

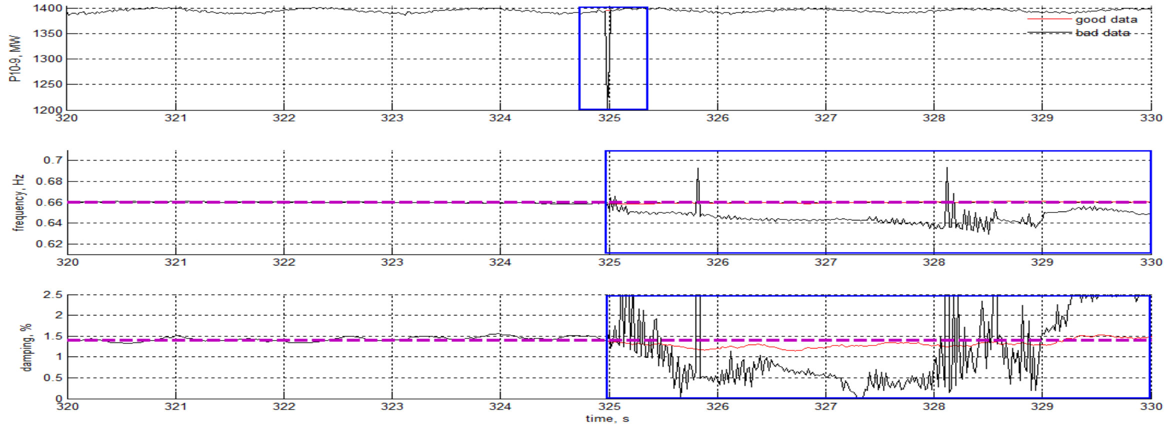

Figure 4.2: Impact of Outlier on Oscillation Monitoring

4.3 Assessment of Impact of Bad Data on Oscillation Monitoring

While the actual effect of bad data is dependent on the specific application that uses the data, the metrics and characteristics identified herein and the methodology presented can be used for studying the impact of bad data on other applications. The characteristics of bad data that have been identified as impactful from the perspective of modal estimation are (1) magnitude of bad data (2) the proximity of the bad data to a major system event like a transient and (3) the number of missing samples or outliers within the processing window of the modal estimation algorithm. These characteristics are introduced as test variables for further study.

Different data streams were generated and bad data was introduced by varying one test variable at a time, and holding the other two variables at constant values. The algorithm uses a processing window length (Tw) of 10 seconds. The impact of bad data is studied quantitatively by comparing the damping ratios estimated by the matrix pencil algorithm using the signal with bad data of different characteristics. The metrics being studied are percentage error in damping ratio over time, squared of errors over time and area under the curve for the time series of estimation errors. For modal estimation algorithms, the distortion caused by bad data has greater impact on damping ratio estimation. While the estimated frequency was also found to be affected, the impact was not as pronounced as that on damping estimation. Therefore, only the effect of bad data on damping ratio estimates is discussed.

The factor that has the greatest impact on the estimation accuracy is the number of contiguously missing samples within the moving window. Estimation errors as high as 33% were observed with data streams containing only 5% missing data within the sampling window of 10 seconds. It was also found that the magnitude of deviation of the bad sample from its true value has a direct impact on the estimation. Thus, a data stream with a continuous stream of outliers that are within the nominal value of the measurement has the potential to cause the greatest error in damping estimation. This is shown in Figure 4.3 where about 25 data samples have an offset of 50MW. The impact on damping estimation is well pronounced.

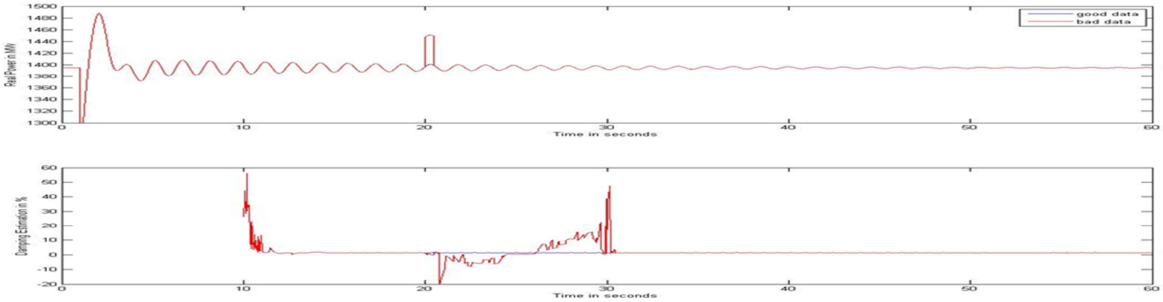

Figure 4.3 Impact of Data Offset on Damping Estimation

While the inaccuracy caused by bad data may not be operations-critical for highly damped modes, such errors are detrimental in the control of lightly damped modes since the actual control action being contemplated by the operator may depend on how worse off one mode is. During low damping scenarios, a wrong estimation may result in unwarranted or excessive control action leading to potentially destabilizing situations, while in other cases, the algorithm could be rendered ineffective or oblivious to a major oscillatory condition.

From the study, the following conclusions were made:

1. The impact of bad data is more pronounced on the damping ratio estimation than on frequency estimation.
2. The data streams with several missing samples are a cause for concern since they have the greatest impact on modal estimation. Though such data streams are discarded, they may hold valuable information especially if they originate from a critical substation.
3. While location of a bad data sample in close proximity to a system transient amplifies the estimation error, this variable may not be impactful for practical purposes since the transient data samples are not used for oscillation monitoring.

5. SVD AND PCA-BASED BAD DATA DETECTION

Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are mathematical tools that are used for linear transformation of a given dataset to provide valuable geometric intuition in the behavior of the system that is being observed. These tools provide a way to reduce high dimensions of data to a lower dimensional subspace, and it often evinces hidden and simplified structure in a large data set. This feature makes it an attractive candidate for detecting data anomaly in PMU measurements. Unlike state estimation-based techniques, the algorithm proposed herein is based only on the measurements and thus are not impacted by the accuracy of the underlying model. The algorithm also circumvents the complexities like the need for continuously validating a model and additional computational burden arising out of employing a model-based approach.

The SVD-based tool is tested with data generated from the two-area test system described earlier. An outlier is introduced at 325th second in the data stream. The output of the SVD-based tool is shown in Figure 5.1. As shown, the SVD-based tool shows a large spike when the transient data sample occurs. A smaller spike can be seen at the point where the bad data sample was introduced.

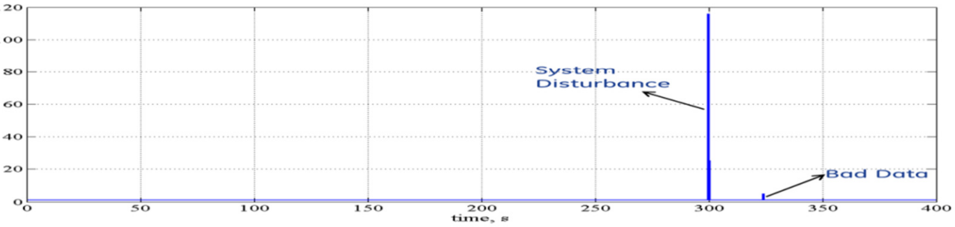

Figure 5.1 Output of SVD-based Detection Tool

The PCA-based tool with a moving window of 500 samples was tested with simulation data from the 4-machine test system. The data consists of 14 power signals, 10 voltage angles and 11 voltage magnitudes. Outliers were deliberately injected for two of signals at 325th and 340th second of the simulation. The output of the PCA-based tool for these two signals is presented in Figure 5.2. As highlighted by the red circles, subspaces 1 and 3 for the two signals show an elevated value for the duration of the time for which the outlier is present in the moving window; the subspaces of the other good signals do not show any such response. The subspaces of all the signals show similar response during the transient period, enabling the algorithm to distinguish between system events and outliers.

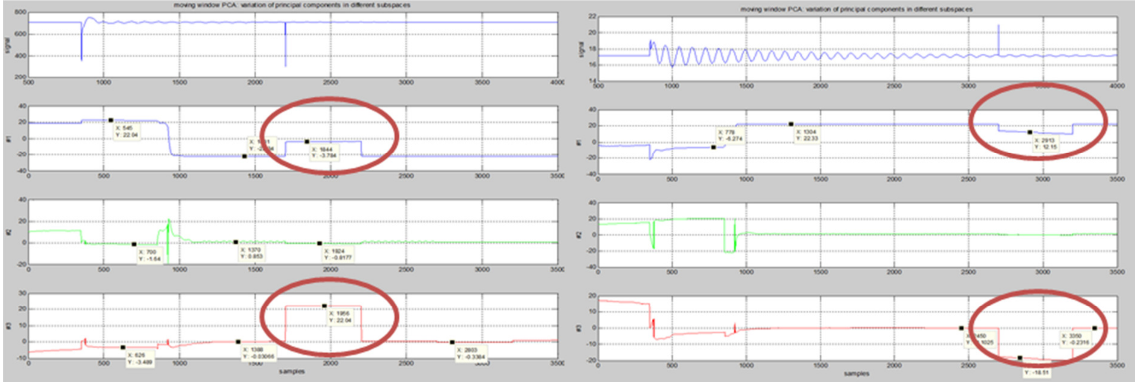

Figure 5.2 Output of PCA-based Detection Tool

6. CONCLUSION

A generalized framework for assessing the quality of synchrophasor data has been presented in this paper. The impact of poor data quality on oscillation monitoring was studied, and it was found that bad data can have a significant impact on the accuracy of such applications leading to inaccurate estimation of system stability. Finally, a data mining-based technique that can detect the bad data and distinguish such an anomaly from an actual system event was presented.

BIBLIOGRAPHY

[1] Phadke, A. G., and J. S. Thorp. "History and applications of phasor measurements." *Power Systems Conference and Exposition, 2006. PSCE'06. 2006 IEEE PES*. IEEE, 2006.

[2] Kanabar, M., M. G. Adamiak, and J. Rodrigues. "Optimizing Wide Area Measurement System architectures with advancements in Phasor Data Concentrators (PDCs)." *Power and Energy Society General Meeting (PES), 2013 IEEE*. IEEE, 2013.

[3] Yang, Tao, Hongbin Sun, and Anjan Bose. "Transition to a two-level linear state estimator—Part II: Algorithm." *Power Systems, IEEE Transactions on* 26.1 (2011): 54-62.

[4] Zhang, Liuxi, and Ali Abur. "Impact of tuning on bad data detection of PMU measurements." *Innovative Smart Grid Technologies-Asia (ISGT Asia), 2012 IEEE*. IEEE, 2012.

[5] Zhu, Jun, et al. "Enhanced state estimators." *PSerc Final Report* (2006).

[6] Farantatos, Evangelos, et al. "Advanced disturbance recording and playback enabled by a distributed dynamic state estimation including bad data detection and topology change identification." *Power and Energy Society General Meeting, 2012 IEEE*. IEEE, 2012.

[7] Zhou, Ning, et al. "Electromechanical mode online estimation using regularized robust RLS methods." *Power Systems, IEEE Transactions on* 23.4 (2008): 1670-1680.

[8] Zhou, Ning, et al. "Robust RLS methods for online estimation of power system electromechanical modes." *Power Systems, IEEE Transactions on* 22.3 (2007): 1240-1249.

[9] J. Liu and S. Fahr, "PJM Phasor Data Quality Task Force and Improvement," North American Synchrophasor Initiative (NASPI), Knoxville, 2013.

[10] N. G. Weiskopf and C. Weng. "Methods and dimensions of electronic health record data quality assessment" *Journal of the American Medical Informatics Association* 20.1 (2013): 144-151.

[11] T. K. Sarkar and O. Pereira, "Using the Matrix Pencil Method to Estimate the Parameters of a Sum of Complex Exponentials," *lEEE Antennas and Propagation Magazine,* vol. 37, no. 1,1995.

[12] Gardner, Robert Matthew. *A Wide-Area Perspective on Power System Operation and Dynamics*. Diss. Virginia Polytechnic Institute and State University, 2008.

[13] Kundur, Prabha. *Power System Stability and Control*. Vol. 7. New York: McGraw-Hill, 1994

[14] Electric Power Group, "Synchro-Phasor Data Conditioning and Validation Project Phase 1 Task 2," Office of Electricity Delivery and Energy Reliability, 2013.