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

CIGRE US National Committee
2015 Grid of the Future Symposium

# Baselining PMU Data to Find Patterns and Anomalies

**B.G. AMIDAN**
**J.D. FOLLUM**
**K.A. FREEMAN**
**J.E. DAGLE**
**Pacific Northwest National Laboratory (PNNL)**
**USA**

## SUMMARY

This paper looks at the application of situational awareness methodologies with respect to power grid data. These methodologies establish baselines that look for typical patterns and atypical behavior in the data. The objectives of the baselining analyses are to provide: real-time analytics, the capability to look at historical trends and events, and reliable predictions of the near future state of the grid.

Multivariate algorithms were created to establish normal baseline behavior and then score each moment in time according to its variance from the baseline. Detailed multivariate analytical techniques are described in this paper that produced ways to identify typical patterns and atypical behavior. In this case, atypical behavior is behavior that is unenvisioned. Visualizations were also produced to help explain the behavior that was identified mathematically. Examples are shown to help describe how to read and interpret the analyses and visualizations.

Preliminary work has been performed on PMU data sets from BPA (Bonneville Power Administration) and EI (Eastern Interconnect). Actual results are not fully shown here because of confidentiality issues. Comparisons between atypical events found mathematically and actual events showed that many of the actual events are also atypical events; however there are many atypical events that do not correlate to any actual events. Additional work needs to be done to help classify the atypical events into actual events, so that the importance of the events can be better understood.

## KEYWORDS

Situational awareness for complex systems; Multivariate statistical analyses; phasor measurement units (PMU); Baselining analyses

b.amidan@pnnl.gov

## INTRODUCTION

"Situational Awareness" is the process of understanding the elements in a complex system, discerning how they behave with changes to the system (i.e. over time), and projecting their status as these changes occur. Advanced statistical and mathematical algorithms can be applied to complex system data to help provide insight into the situational awareness of the system. This research looks at ways to build algorithms around power grid related data, to help provide the system engineers with an awareness of grid behavior. This research has been focused on two areas: 1) establishing a baseline of what is "normal" grid behavior, and 2) identifying unenvisioned anomalies within the power grid.

The recent increase in the deployment of high-speed time-synchronized measurements including Phasor Measurement Units (PMUs) provides a great challenge and opportunity to the power grid community. One significant challenge is in handling the substantial amounts of data, cleaning that data, and then creating insightful analyses and displays. There is great opportunity to provide system engineers real-time tools that increase their understanding of the current state of the grid and predictions of future grid behavior.

This paper looks at the application of situational awareness methodologies with respect to power grid data. These methodologies establish baselines that look for typical patterns and atypical behavior in the data. The objectives of these baselining analyses are to provide:

- near real-time analytics,
- capability to look at historical trends and events, and
- reliable predictions of the near future state of the grid.

This paper focuses on the first two objectives, while providing tools to start understanding how current data can help predict future behavior.

Analyses in this paper were performed using PMU data for baselining tasks for BPA (Bonneville Power Administration) and the Eastern Interconnect (EI). Over a year of PMU data were analyzed from BPA, consisting of voltages, currents, phase angles, and frequencies. This data was measured providing 60 samples per second and was taken from over 50 PMUs. Two sets of two month data were also studied from the EI. This data consisted of phase angles, so the analysis focus was concerning phase angle pairs. This data was summarized in 1 sample per second rate and was taken from over 30 locations across the Eastern Interconnect.

## METHODOLOGY

It is common within complex systems to establish rules to alert system engineers when certain behavior occurs. These rules are criteria that have been predetermined and envisioned by system engineers. An example rule would be providing an alert when power grid frequency exceeds a certain limit, like 61 Hz. These alerts are a simple way to provide a real-time check of a complex system for phenomena that has already been envisioned. While this approach has tremendous value, it also has the potential to miss abnormal phenomena or events that have not been envisioned. Advanced statistical and mathematical algorithms can provide additional situational awareness tools that look for patterns in the data and find unenvisioned phenomena. This section details the algorithms used to discover patterns and find atypical events.

The first step in analyzing data is to extract features from the data that will provide insight into the state of the system. These features are extracted from each relevant variable. These features make up a mathematical signature which provides a summary of the important aspects of the system and these signatures are used in the analyses. The signature used in these analyses was determined by fitting the following regression equation across a moving window of data for each variable:

$$y = a + bx + cx^2 + \varepsilon$$

where $y$ is the actual data within the window; $x$ is time; $a$ is the $y$-intercept, representing the mean across the window; $b$ is the slope, or rate of change; $c$ is the quadratic, or rate of rate of change; and $\varepsilon$ is the error, or lack of fit of the data to the regression equation. This equation is fit for a window of data of a certain size (one second or one minute). The window is then moved a certain amount of time forward (in this case one second) and the equation fit again. In each case, the $a$, $b$, $c$, and $\varepsilon$ are calculated and stored. This continues across all the data for each variable.

This signature calculation results in the extraction of the following features for each given variable at a specific time –

- the magnitude of the actual data values, represented by $a$;
- the rate of change of the data, or slope (first derivative), represented by $b$;
- the rate of rate of change, or acceleration (second derivative), represented by $c$; and
- the amount of error in the fit of the data to the equation, represented by $\varepsilon$.

Each of these signature elements provides insight into a different aspect of the data. These signature elements can then be *aggregated* (summarized) across a certain amount of time (in this case every minute or every hour). This aggregation is done by calculating the mean, minimum, maximum, and standard deviation of each signature element. This resulted in 16 signature elements calculated for each specified moment in time. These signature elements can then be included in all analyses concerning the data, especially those analyses looking for typical patterns in the data, or atypical events. Amidan and Ferryman (2005) provide more detailed instructions into the calculation of this signature [1].

Each signature element provides different insight into the state of the system at a given time. For example, the element $a_{mean}$ provides a view of the average magnitude of the data values. The element $a_{stdev}$ provides insight into the variability in the magnitude of the data values. The element $b_{mean}$ provides the average rate of change, while the $b_{stdev}$ provides the variability within the rate of change. When analyzing the data, it is important to note which signature elements are needed to provide the desired feedback. If an analyst is interested in exploring aspects about how the variables are changing over time, they may want to use the $b_{mean}$ element and possibly the $a_{stdev}$ element. If they are just interested in analyzing the actual values (magnitudes) of the data, they may want to use $a_{mean}$. If they want to explore all the aspects of the signature, they may include all elements.

After the mathematical signatures are calculated, then analyses can be performed. The rest of this section is focused on a multivariate approach in determining patterns and looking for anomalies, or atypical events, in the data. The first step is to decide which data to perform the analysis on. This includes selecting the time period, the variables or variable types, and the signature elements.

Once the data is selected, then a data reduction algorithm is performed. A common method to do this is principal component analysis (Rencher, 1995) [2]. The purpose of this step is to reduce the number of variables in the analysis to a set of unique, uncorrelated variables. This is necessary because many of the variables are highly correlated. Including multiple variables that are related to a certain characteristic in the data will weight that characteristic too heavily in the analysis. Principal component analysis removes this issue by creating linear combinations of the variables that result in orthogonal variables, which are not correlated. The number of components that explained at least 90% of the total variation was retained.

A non-supervised clustering algorithm is then applied to the reduced data. Non-supervised clustering is used as there are no targeted groupings of the data. There are many clustering algorithms to choose from. In this case, k-means is applied, though any clustering algorithm may be applied. Clustering uses multivariate distances within the data to determine which data points are similar. Similar data points (in this case a data point is each specific minute) form a cluster, or group. Each group represents a certain state that the system is in. These represent the common patterns that are in the selected data.

Another analysis of the data is the calculation of the global atypicality score (G). A large score indicates the data point (in this case each minute of time) is unusual, or atypical. A score closer to zero indicates a typical or normal data point. Cluster membership is one of two parts included in the global atypicality score. The other part of the global atypicality score is the distance that the data point is from the center of all the data points. These components help calculate the global atypicality score for each data point (in our case, every minute) where the score is always positive, with larger scores meaning more atypical. Global atypicality scores usually range between 0 and 25. Further detail concerning the global atypicality score calculations can be found in Amidan and Ferryman (2005) [1].

**RESULTS**

This section discusses the type of results that can be found. The actual timeframe of these results has been removed to keep the events de-identified and confidential. These results are only presented to help the reader visualize and understand what can be done. The actual events and happenings found using these analytical techniques are not the focus of this paper.

One way that clustering can be informative is by comparing what items group together during one timeframe to the groupings that occur during a different timeframe. These types of investigations can help detect shifts in the typical patterns. Figure 1 shows cluster trees from two different time periods for phase angle pairs on the Eastern Interconnect. Pairs that are connected further down the tree are more similar. It is interesting to note that in the top time period the CantonCenter-Alburtis and JacksonsFerry-Alburtis angle pairs are very similar; however in the bottom time period the angle pairs are not at all similar. In this case, clustering was helpful in identifying unusual pattern changes. Further research can be done to determine why this was occurred.

Another way clustering can be helpful is in determining possible phase angle pairs that should be studied and analyzed. With hundreds of phase angles possible in a given footprint, there can be millions of pairs that could be studied. Clustering can be used to help determine which angles are similar, so that only one exemplar angle pair is included from each cluster.

In order to find unusual events in data, the atypicality score is calculated. Figure 2 shows a sample of atypicality scores calculated over a period of time within the BPA data set. In this case, 6 different atypicality scores were calculated (each represented by a different color), each score representing a different subset of data that is being studied. As can be seen from the scores in Figure 2, there is a spike in the scores at 19:30. This indicates that something unusual happened during that time.

By looking at the atypicality score in Figure 2, it is not possible to discover what is unusual, only that something unusual has occurred. Drill down graphics are necessary to help pinpoint what issues might be occurring. The first drill down graphic is called the rationale and an example rationale is given in Figure 3. The rationale includes a listing of all the variables that are univariately unusual and an explanation of how each is unusual with respect to the atypicality score in Figure 2. The rationale is written in sentence form and makes up a paragraph. For example, in Figure 3 the first sentence is KEEL230.A2SA.freq value (mean 59.89) is very low. This is when that particular variable value is compared to all the values that this variable has had during the time period being evaluated.

In our software tool, the variable names in the rationale are links which lead to plots called performance envelopes. Figure 4 is an example of a performance envelope plot. This plot consists of a gray area that shows the likelihood that the variable of interest will have that value during that time of day during that month. This is called the performance envelope and it shows the typical values of that variable given the time of day. The orange line plot shows the actual values the variable had previous, during, and immediately after the atypical time period. In Figure 4, the frequency variable of interest shows a dramatic spike down during 19:30, the same time period in which the atypicality score was high.

Atypicality scores were calculated for all the PMU data that we analyzed from BPA and EI. Preliminary investigations have occurred comparing the atypical events found with the atypicality scores and with the actual events in the power grid system. The atypical events consistently find events that were previously known, but also time periods in which no event was recorded. Sometimes an actual event does not contain unusual data and the atypicality score calculation misses them.

Further investigation needs to be done correlating atypical events to actual events. When specific types of actual events can be mathematically characterized, then the atypical events can be classified into known categories of events. This may help attach a level of importance to each atypical event.

Another area of further investigation is prediction of events. Further research needs to be done to find and characterize pre-cursor activity that could lead to events. Time series and regression models could be instituted to help make predictions and attach probabilities of the predicted events occurring in the short term.

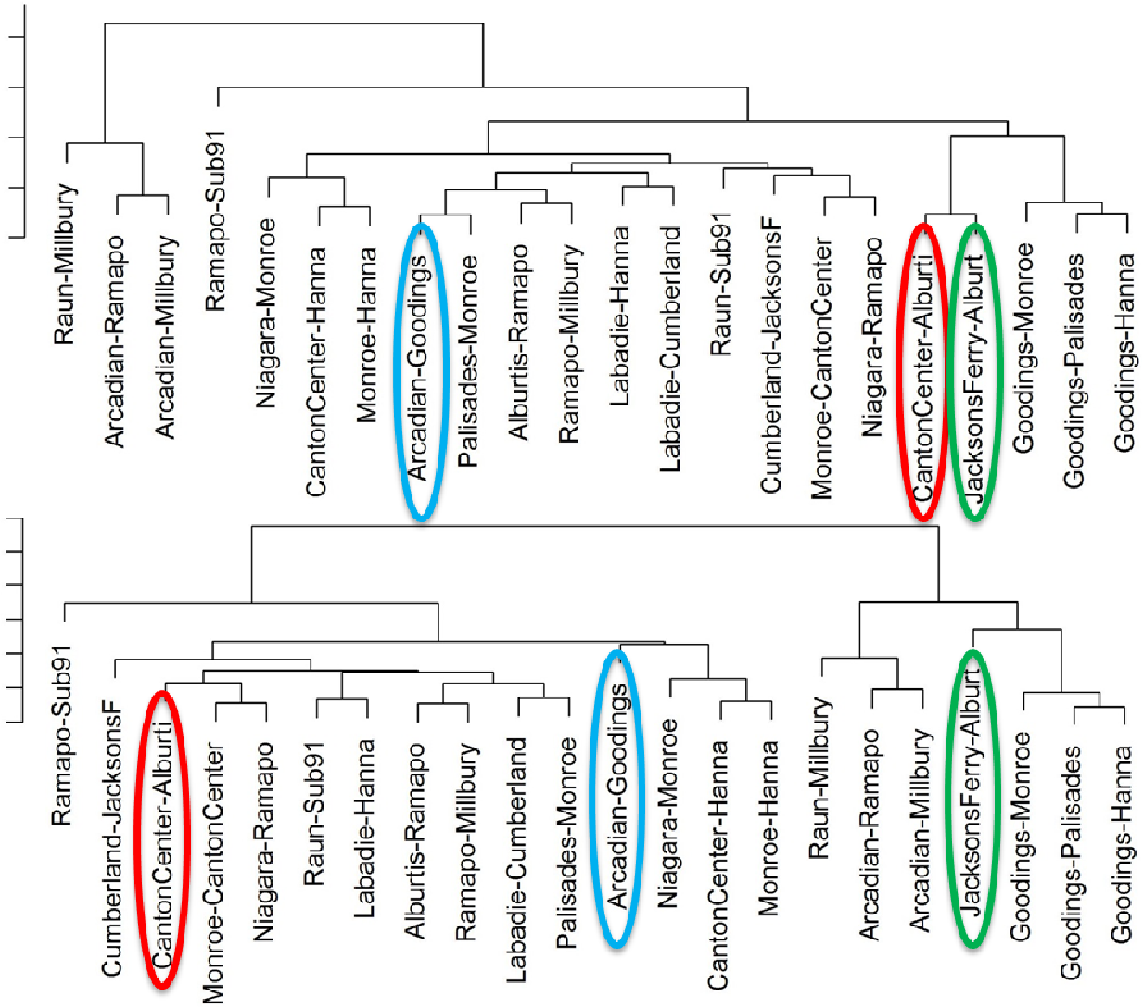


**Figure 1.** Phase Angle Pair Clustering Results for Fall (top plot) and Winter (bottom plot)

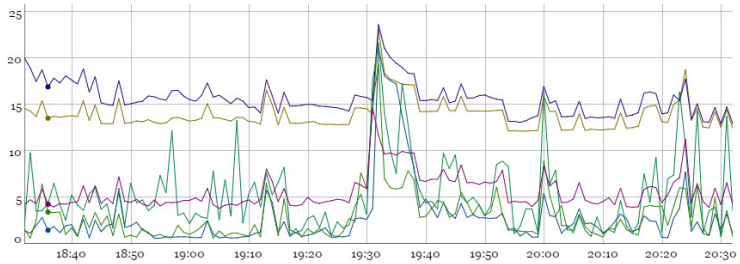


**Figure 2.** Atypicality Scores Plotted Over Time


**Figure 3.** Rationale Explaining which Variables Were Atypical and How They Were Atypical

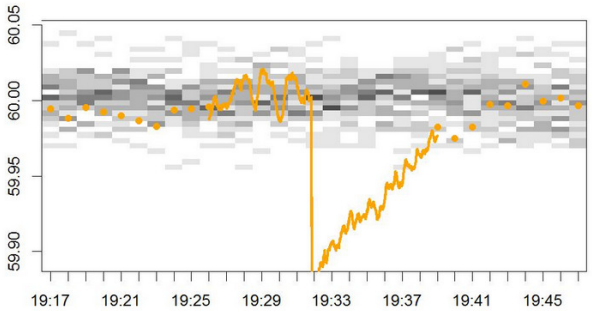


**Figure 4.** Performance Envelope Showing an Atypical Frequency Variable

## BIBLIOGRAPHY

[1] Amidan BG and TA Ferryman. 2005. "Atypical Event and Typical Pattern Detection within Complex Systems." *IEEE Aerospace Conference Proceedings, March 2005.*

[2] Rencher AC. 1995. Methods of Multivariate Analysis. John Wiley and Sons, NY.