Albert Nguyen
A53060342
CSE 255, Winter 2015: Assignment 1

Predicting Electrical Disturbances in a Microgrid

1. INTRODUCTION

According to the Institute of Electrical and Electric Engineers (IEEE), with the increase emphasis on online content and electrical devices, coupled with the rising middle class in India, China, and other developing countries, the world electrical supply will need to triple by 2050 to keep up with demand [1]. With higher demand, come greater strain on the current power grid, jeopardizing communities and industries that rely on stable electricity. A Lawrence Berkeley National Laboratory study in 2004 quoted the cost of blackouts in the US to be $80 billion annually, mainly in commercial and industrial losses [2]. The paradigm of centrally controlled power grids is now seen as unable to deal effectively with rapidly changing demand. The solution is to adapt the power grid into smaller smart grids, allowing more flexibility, self-monitoring and self-regulation.

Phasor Measurement Units (PMU's) are essential elements that make up the modern smart electric grid. With the ability to measure important features such as power frequency, angle, and power at a rate of 30 times a second, to total 2,592,200 data points per day per feature, PMU's provide deep insight into the health and stability of the power grid.

2. DATASET

The data came from PMU's located throughout the University of California, San Diego (UCSD) campus smart grid, collected for the whole month of October 2014, leading to a high frequency time series consisting of 8,0352,000 data points. Specifically the analysis will be studying one PMU, name MCB-1.CSSPMCB-1, located within UCSD.

The analysis will focus on the frequency data collected from the PMU, since frequency data provides the most insight into the health of the grid since it simultaneously integrates both the effects of demand and supply. Significant deviations in the frequency can indicate problems in the microgrid, such as degrading equipment and internal instabilities.

This data is collected at 30 times a second and is precise to over 4 decimal spaces. Such high precision carries the risk of noisy data, since air conditioning units, electric cars, and other high-volume electrical devices can introduce fluctuations in the PMU data, making inferences into the health of the grid difficult.

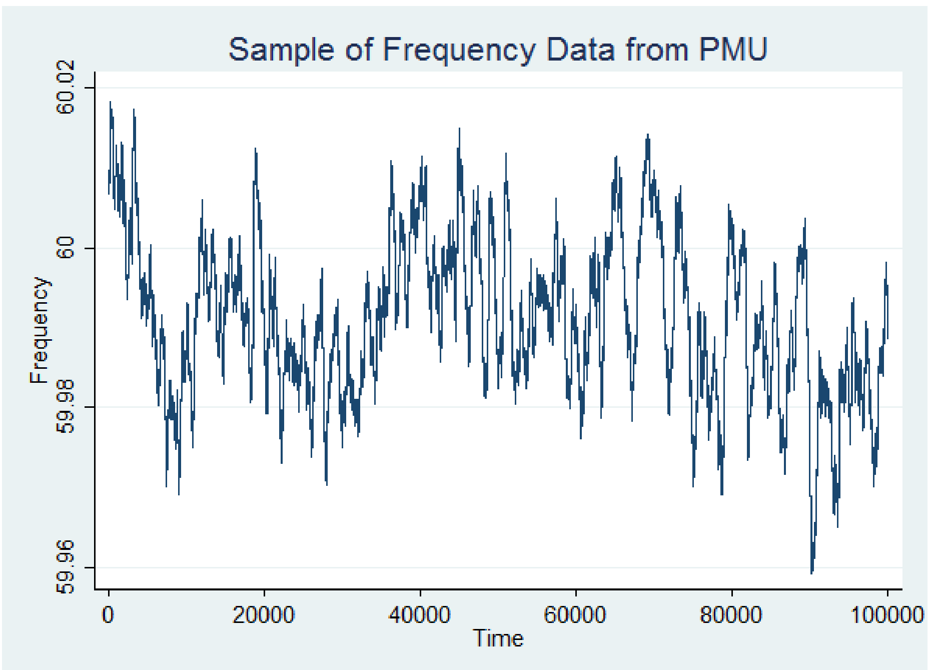

FIGURE 1: Sample of Frequency Data from PMU.

As one can see the data contain mostly ambient noise from changing demand and supply in the power grid. Besides ambient data, other researchers observed several typical behaviors, such as sustained oscillations that last for an extended amount of time and ring down oscillations that occur from the power grid self-adjusting to changing load.

Ideally, the power grids in the United States should be running at 60 Hz to provide consistent power to user. For the UCSD microgrid data, during the month of October 2014, the average frequency was at 59.99945 Hz, with a standard deviation of 0.020133 Hz. The lowest frequency was 59.88285 Hz while the highest frequency was 60.11434 Hz. At closer examination, there are a handful of instances of missing data, but given the fact that the data is collected at high speed and that adjacent values tend to be very similar to each other, these missing values can be effectively removed. Outliers, where the frequency would change drastically for one observation and return to normal, also occur at greater frequency compare to missing data and could be attributed to sensor interference. The analysis needs to take into account of these outliers and remove them from the analysis if need be. Finally, Kernel Density shows that the frequency distribution closely resembles a normal distribution. These results show a power grid which is functioning close to the standard 60 Hz.

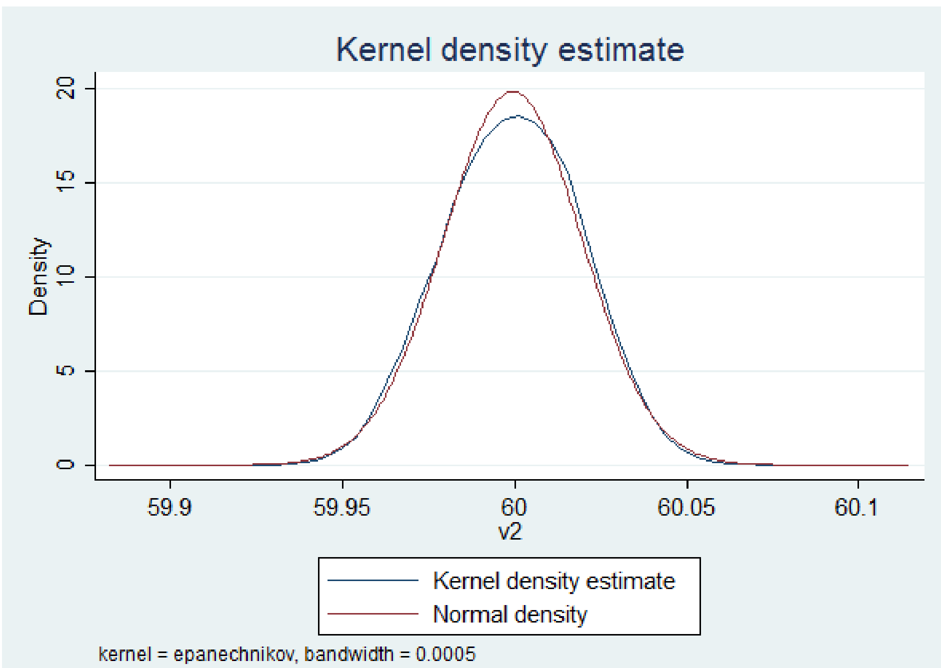

FIGURE 2: Kernel Density Estimates of Frequency Data from October 2014

Yet when one looks at the change in frequency, there are instances where the frequency change rapidly, which is not attributed to outliers effects. During the month of October 2014, there were 100 instances where the frequency change over half a standard deviation within 1/30 of a second. Looking at longer stretch of time, one can find that the changes can compound very rapidly, changing the power grid frequency over a wide range in a very short amount of time.

When the frequency is plotted over time, one will find instances where the frequency drops significantly such as below.

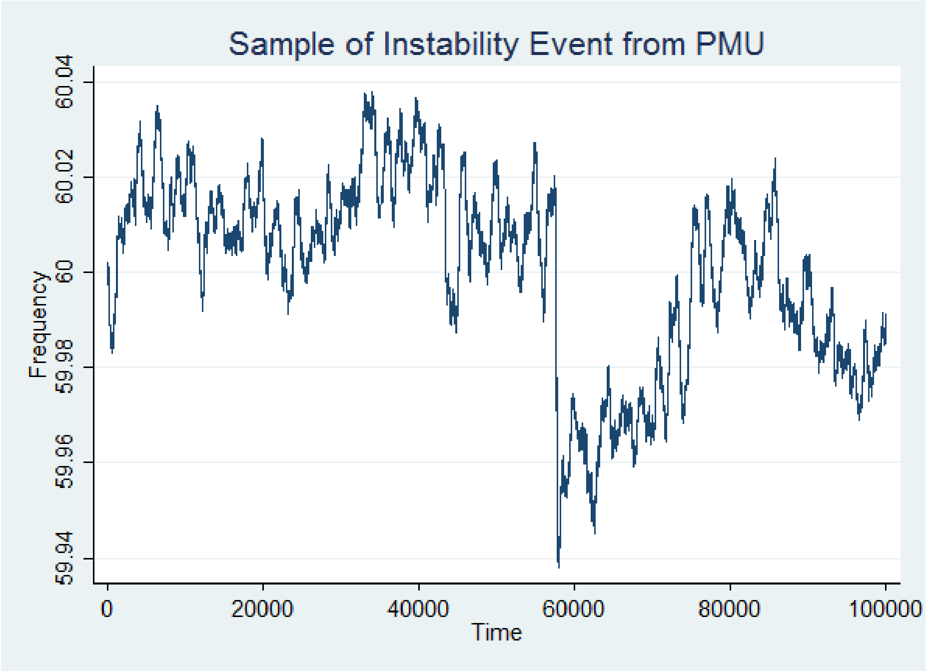

FIGURE 3: October 7, 2014 Event at 3:00 PM that last about 6 minutes

These events are an interest for utilities providers because they demonstrate instabilities within the power grid. Normally these instabilities events will not cause any perceivable effects on the power grid, such as light flicker or a black out, but they do add considerable strain to the grid as generators and other equipment ramp up to bring the grid back to 60 Hz. For many utilities companies, these instability events incur additional wear and tear onto equipment and reduce lifetime performance.

If one had the ability to be able to predict when these events would occur they would be able to mitigate the event by taking precautions. If done correctly, these precautions could help prevent blackouts from occurring, such as the August 10, 1996 blackouts, which was characterized by a cascade of minor instabilities that eventually lead to one of the largest blackouts in the United States, leading to almost 30,000 megawatts to be lost [3].

The purpose for this research project is to look into the frequency data to identify patterns that might predict these events, giving utilities the ability to dampen the effects of an occurring event in the power grid, preventing any wear and tear on equipment and also help stop instabilities cascades that can lead to power failure

3. PREDICTION AND MEASUREMENT

The goal for this analysis is to be able to predict instability event in the UCSD microgrid, by looking possible patterns in the PMU frequency data. If one is able to find a strong correlation between certain frequency patterns that occur right before an instability event, then one can develop a detection method for those certain frequency patterns. Detection of those specific frequency patterns can foreshadow an upcoming instability event.

The final efficiency of the prediction model can be measured in several ways. First the prediction model must accurately predict if an event will occur or not. In this analysis, the prediction model results will be a binary of 1 or 0, which will correspond with “an event will occur” and “no future events”, respectively. The binaries from the prediction results will then be compared to the actual raw data from the PMU. The raw data will need to be labeled with a binary itself, again corresponding with “an event will occur” and “no future events”. Using such a binary for both testing data (raw data) and the prediction data, will give the analysis a way to judge how accurate the results are. In addition, it will allow us to track and measure the proportion of false positives and true negatives, since those results might incur a cost to the power grid if utility operators intervene with false information.

Secondly, the prediction model can be measured by how early the model is able to predict instability events since a late warning will not provide utility operators the needed time to mitigate issues on the grid. It is assume that the earlier the warning, the higher chance that the results might be inaccurate, so essentially there might be a tradeoff between providing more warning time and having a lower prediction accuracy.

4. LITERATURE REVIEW

In the past few years, more and more attention been given to the analysis of PMU data, particularly in the ability to detect anomalies. This allows for a rich landscape of approaches to learn from and compare to. Generally research fall into two categories, traditional signal analysis and machine learning approaches.

For traditional signal analysis, the main tool that many researchers employ is the Fast Fourier Transformation (FFT), which is a spectral analysis which takes frequency data and decompose it down into basic sinusoid waves.

This gives researches the ability to remove noise from the data and hence be able to look at the change in frequency more accurately. One set of researchers, Zhou et al [4], from the Pacific Northwest National Laboratory, looked at FFT's that correspond to instability events. They found that instability events carry a large amount of energy at lower frequency ranges and tend to show up as spikes in the lower frequency range of an FFT. They develop a method to detect instability events by developing a threshold limit where if the FFT change and crosses the threshold, it will indicate and event most likely occurred.

Another approach to signal analysis is through an autoregressive approach. De Callafon and Wells [5], pioneered an approach to use an autoregressive moving average filter to strip out the ambient fluctuation in the data. After removing the ambient noise they determine the rate of change of the signal. If the rate of change occur outside a statistically established threshold, the data will be labeled as an instability event. De Callafon and Wells method allow for a much more define and consistent definition of what an event is and allow for filtering of ambient data and removal of the effects of outliers.

Beside signal analysis, other researchers took a different approach. Kumar Saha Roy et al [6], proposed developing indices of generator bus voltage magnitudes and generator electrical output power. From there the indices would be used to identify generalize normal behavior and instability events.

Hazra et al [7], approach the prediction of instability events by using PMU data to build a model of generator rotor positions and comparing it to the actual generator rotor positions. If there is a deviation between the model and the actual rotor position, there is a higher chance instabilities may occur.

Li et al [8], used principal component analysis to reduce the PMU frequency into a lower dimension representation. They did this for both normal ambient data and instability event and trained an adaptive model to track changes in the lower dimension representations.

Lastly, Tranchian [9] from TVA propose using Hadoop and Symbolic Aggregate Approximation to scan over petabytes of data, to identify patterns of interest.

As one can see, most of the research that been conducted focuses on the detection of instabilities, either through indexes, filtering, of threshold detection. In more recent times though, the application of data mining methods and machine learning been used to detect instability events with great promise.

5. FEATURES

The features for this analysis will be the frequency data from the PMU's. As stated earlier, the main problem with this dataset is that it contains a lot of noise since the data is very sensitive to change in demand and supply, yet provide the best insight into the health of the power grid. The frequency data needs to be transformed in such a way where the noise can be removed, leaving only the underpinning power grid frequency.

There are different ways to remove noise, one that is predominant in signal processing is using an autoregressive moving average, as demonstrated by de Callafon and Wells. Another method that is widely use is using the Fast Fourier Transformation (FFT), which breaks down a frequency into its sinusoid waves. Using a FFT, much of the noise will be retained in the high frequency ranges, while the underpinning power grid frequency will be kept in the low frequency ranges.

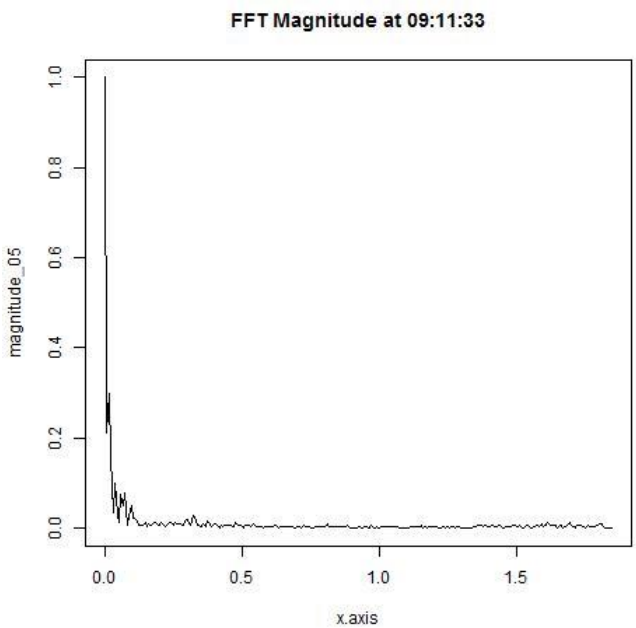

FIGURE 4: FFT Magnitude Sample. Noise and interference are located in the higher frequencies

This analysis will be using the FFT as a method to strip out the noise from the data, thus giving insight into the various changes in the power grid.

To conduct an FFT though, a proper size window must be use, which needs to be big enough to capture all needed features but also small enough to pick up small changes that would indicate that an instability event was to occur.

The best thing to do is to use multiple FFT window sizes and test out which window will provide the most variance and information but small enough that it can provide early warning. The first best guess at the proper window size would be 4096 samples, which is a 2.26 minute window, which considering the data is collected at 1/30 of a second may be rather large. If the window was cut in half to 2048, one would still be looking at a 1.13 minute window which is quite large given the speed of change in the grid. Finally at 1024, at around 34 seconds, which seems to provide a big enough window so that removal of noise is allowed while being short enough to provide details into the underpinning power grid frequency. Again, a more definitive test would be to check the various window sizes against a variance criteria yet given time constraints, this was not perused.

Another way to get around the FFT window size is to have a sliding window, where once an FFT is calculated for the 1024 samples, the window moves 1 sample ahead and takes another FFT. This allow a lot overlap between the FFT and provides a coverage to capture all possible features that may occur in the frequency data.

Before any FFT is done, the frequency sample is transformed using a Hann function, which adjust the sample to fit a distribution which emphasizes the middle of the sample, while reducing possible interference from the edges of the sample. The tradeoff of using the Hann function is that since it deemphasize the outer edges of the sample, the FFT is generally less sensitive to new disturbances until it reaches the middle of the sampling window, which is in this case should be around 17 seconds.

To summarize the first set of transformation on the PMU frequency data; first, a sample of 1024 data points are extracted, which is then run through a Hann function, which emphasize the middle of the sample and deemphasize the edges. Then the frequency data is run through an FFT and is deconstructed into the basic sinusoid waves and stored away. Once that is done, the sample window moves shifts one sample point ahead and the process is repeated until the end of the sample.

This process generates a large set of features to work with, basically every data point, a new FFT of size 1024 will be produce. Meaning in 1 second, you will produce a 30,720 data points. Keeping in mind that the FFT was utilized to cut down on noise, we are able to drop some of the FFT's high frequencies noise data from the analysis. Choosing the ranges to remove noise from the FFT requires an analysis on the effect of noise on FFT; but for the sake of brevity, this analysis concludes that the first 250 frequencies in the FFT contain a lot of information regarding the power grid health, while the rest of the FFT will be discarded.

Now the data set is the first 250 frequencies of the FFT, for every data point in the original frequency data. These features, which now shows the underpinning conditions of the power grid, with the noise effectively removed, will be used to predict power grid instability events.

6. MODEL

The analysis model uses a combination of K-means and logistical regression to label and identify possible early indicators of an instability event.

The training data that was used was extracted from October 7, 2014 from 1:00 PM to 3:00 PM. Within that time frame there was one major event that occur, which lasted 6 minutes, where the frequency dropped by 0.12869 Hz which is considered 6.39199 standard deviations of a change within 6 seconds of time. The training data was transformed to an FFT via a method described in the Features section.

With the training data transformed to an FFT, it will need to be labelled to property identify possible FFT's that can predict an instability event. To do proper labeling, K-means was used to identify unique groups of FFT's that occur before in instability event and were labeled '1', meaning "an event will occur" and all other FFT's were labeled '0' to indicate "no future events". To optimize to the proper number of clusters, the total sum of squared error was calculated for various cluster numbers. The total sum of squared errors and the numbers of clusters were then plotted to determine the number of clusters with the most amount of variation explained, in other words the cluster where the total sum of squared errors don't change so much.

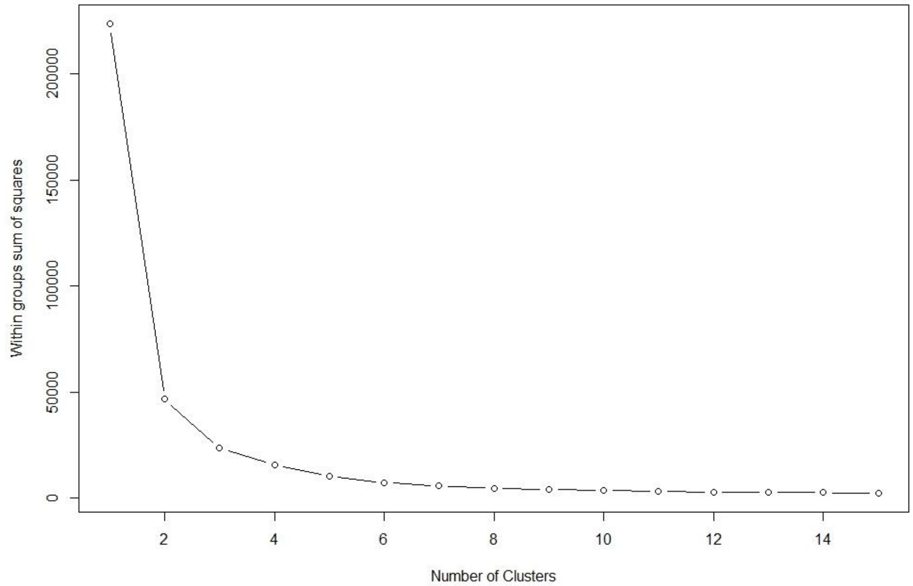

FIGURE 5: Clusters Numbers and Total Sum of Squared Deviance

But if we look at cluster assignments, the vast majority belong in one cluster, while the rest was sporadically spread out in the other 3 groups. These assignments provide no real information regarding clusters that might preclude before an instability event.

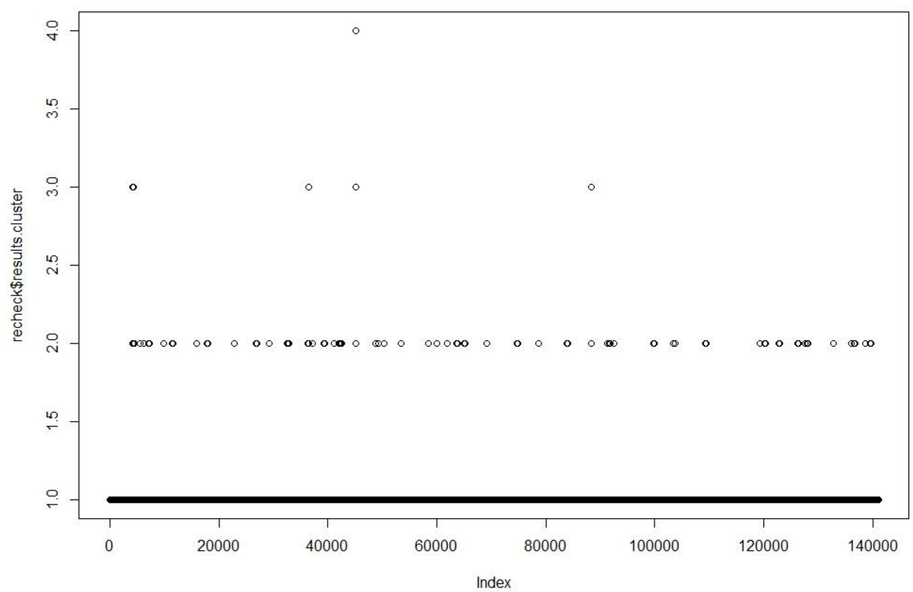

FIGURE 6: Cluster allocation using 4 clusters.

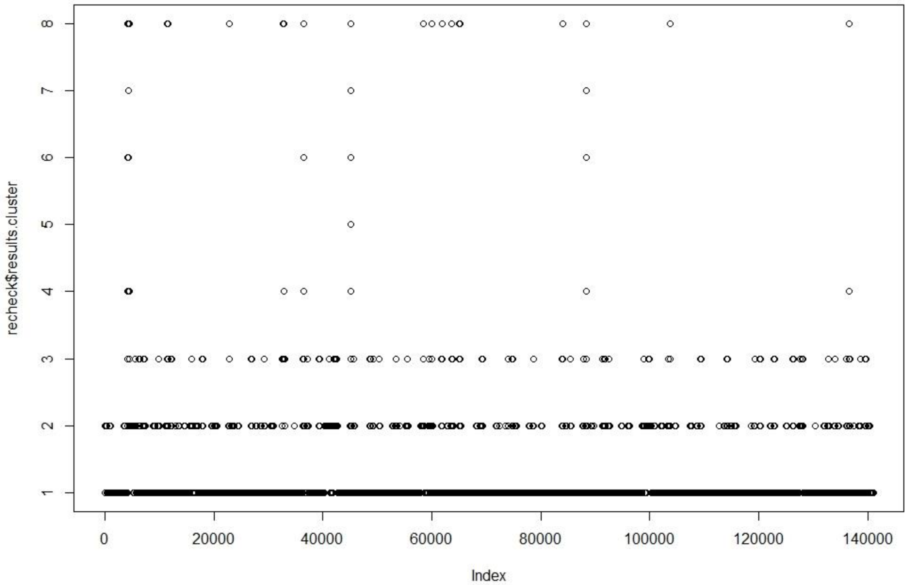

FIGURE 7: Cluster allocation using 8 clusters.

An alternative approach is to increase the number of clusters so that it forces greater resolution of clusters that occur nearby instability events. From the clusters, it is estimated that the 9000 data points before and instability event occur might be unique and hence should be labeled '1', for "an event will occur" while the rest of the data set should be labeled "0" for "no future events".

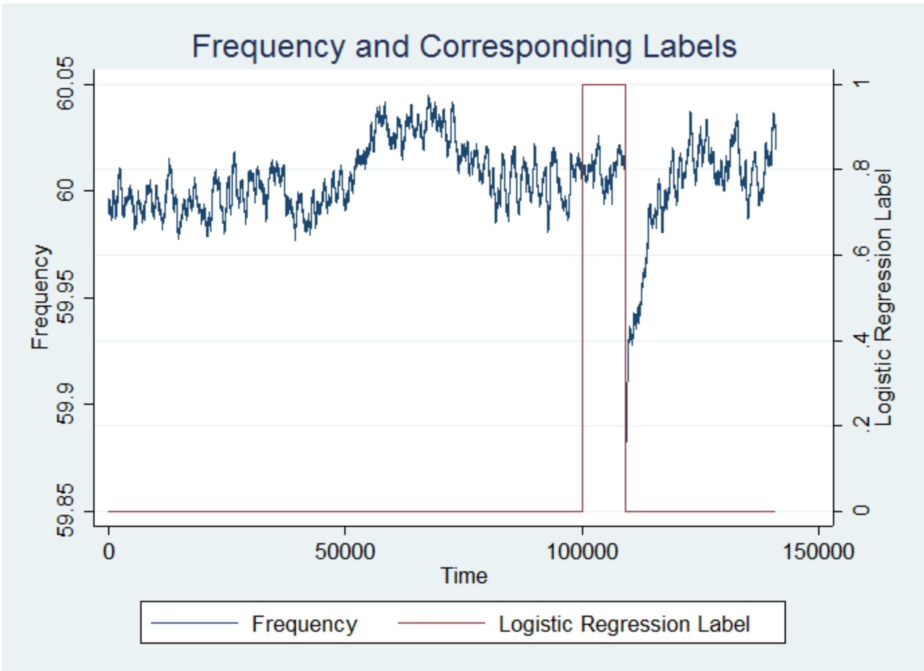

FIGURE 8: Logistic Regression training labels

The results from the labeled FFT's were then fed into a logistic regression, where the dependent variable were the binary labels and the independent variables were the FFT frequency data. Logistic regression was optimized using a regularize of

1. The resulting logistic regression coefficients were then pulled out and reused on the training data, to determine how well the logistic regression was able to pick up the labels in the training data. The following is the results

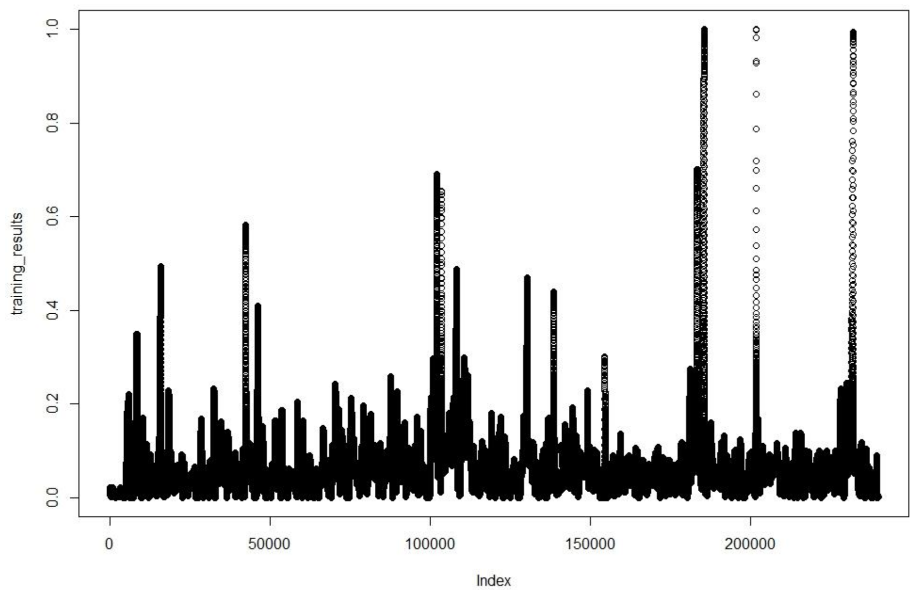

FIGURE 9: Results from Logistic Regression on Training Data

Compare to the original frequency show in Figure 8, the logistic regression is able to pick up the original labels, but at the same time mislabeled other parts of the frequency as instability events. This shows a weakness in the model where the training data for the logistic regression was too broad and hence started to mislabel other parts of the data or there is actually no correlation between the labeled FFT and instability events. The only way to understand if this is the case is to use the coefficients from the logistic regression and test it on data that never been seen before.

The testing data set consist of an composite of 12 hours data containing six 1-hour samples of event data, while the rest of the data set contain ambient data. The data set was converted to FFT using the above method and ran through the logistic model built earlier. The results from the testing data can be seen below.

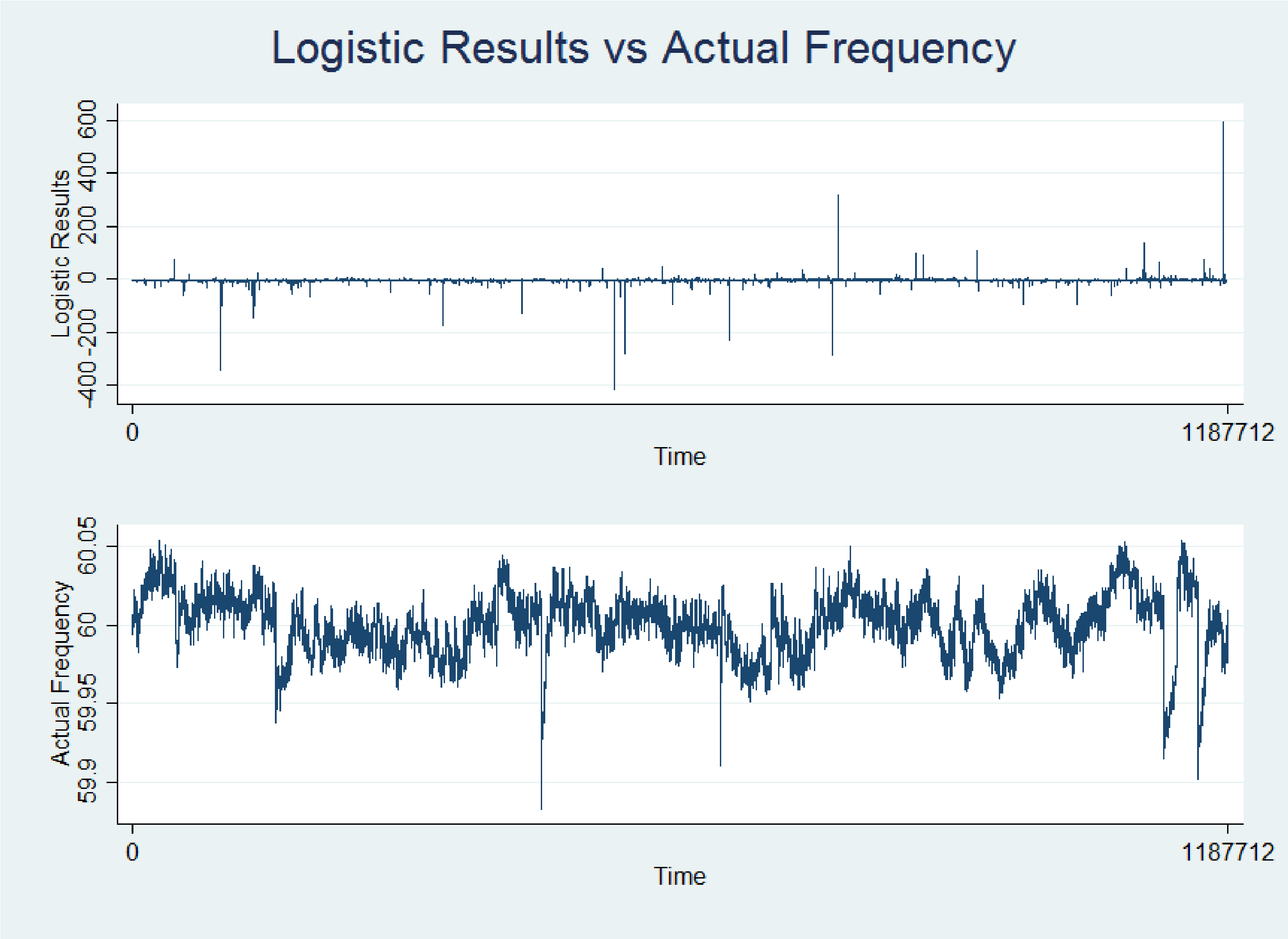

FIGURE 10: Logistic Regression Results versus actual Frequency. Higher logistic results supposedly indicate an oncoming frequency instability event. Testing sample compose of 12 different samples throughout October 2014.

7. RESULTS AND CONCLUSION

From the testing results one can conclude that the FFT/K-Means/Logistic Regression faired rather poorly in predicting if an instability event would occur. Of the five instability events in the dataset, the model only seem to pick up on two a few minutes before it occur. Yet even with two success, the model accidently mislabeled five other times where it predicted an instability to occur, but nothing happened. In short, if we were to only look at the significant logistical regression results, the following is the resulting confusion matrix.

		Actual Results	
		Event Actually Occur	No Event Occur
Predicted	Predicted Event	2	5
	Predicted No Event	3	0

TABLE 1: Confusion Matrix

Depending on the cost on the grid for unnecessary dampening, the model might incur more harm on the grid than actually preventing harm.

Looking at the coefficients themselves, there are certain ranges of frequencies, such as frequencies from 121-160 Hz, which is correlated with greater probability of an instability event occurring. On possible direction for future analysis is to look at such frequencies and see if it is the case that movements in those frequencies is indeed strongly correlated to instability events. Another set of frequencies is in the 10-30 Hz range, which demonstrate similar effects to 121-160 Hz.

At the end though the results from the model show that there is more room for growth and exploration in this field. Through this analysis, it became clear that instability events come in many different types, some of which have drastically sharp drop offs while others have a not so dramatic decent down. The next step in analysis is to use de Callafon and Well's event detection method to label all events during the month of October. Once these events are extracted then one can cluster these events into different groupings to see if there are general categories that can be used.

Once clustering of these events are done, the best thing to do is to determine the actual causes of such events. It became apparent from this analysis that there is a strong possibility that some of the events displayed are caused by external sources, such as a tree branch interfering with a power line. Such events cannot be predicted by any model since they are exogenous. These events should be removed from the training model since they will interfere with the prediction model. Instead, the analysis should identify internally caused instability events and used those for training data.

After such changes are done, a study into the correlation of FFTs vs events should be conducted to see how closely they are related to each other. From such study one can determine the proper size windows for the FFT's and length of prediction time. One can attempt to use logistic regressions again on the refine labeled data but such methods does not take into consideration the importance of time effects and general trends within the data. Time series analysis of FFT frequencies might provide a better understanding of the nature of the power grid and hence aid in prediction of instability events.

8. SOURCES AND REFERENCES

[1] Durant, Paul. "Smart Grid Consumer Benefits." IEEE.org. IDDD.org, 4 Sept. 2013. Web. 19 Feb. 2015. <http://smartgrid.ieee.org/questions-and-answers/964-smart-grid-consumer-benefits>.

[2] LBNL-55718. Understanding the Cost of Power Interruptions to U.S.. Electricity Consumers. Kristina Hamachi LaCommare and Joseph H. Eto. Energy Analysis <certs.lbl.gov/pdf/55718.pdf>.

[3] "Grid Stability." Pacific Northwest National Laboratory. Pacific Northwest National Laboratory, 1 June 2013. Web. 19 Feb. 2015. <http://eioc.pnnl.gov/research/gridstability.stm>.

[4] Zhou, Ning, Jian Yin, and Bora Akyol. "Spectral Analysis of Power Grid PMU Data." Pacific Northwest National Laboratory, 12 June 2012. Web. 19 Feb. 2015.

[5] De Callafon, Raymond, and Charles Wells. "Distributed Real-Time Electric Power Grid Event Detection and Dynamic Characterization." CIGRE US National Committee 2014 Grid of the Future Symposium (2014). Print.

[6] Saha Roy, Biman Kumar, Avinash Kumar Sinha and Ashok Kumar Pradhan. "Synchrophasor-Assisted Prediction of Stability/Instability of a Power System" International Journal of Emerging Electric Power Systems, 14.1 (2013): 1-8. Retrieved 23 Feb. 2015, from doi:10.1515/ijeeps-2013-0028

[7] Hazra, J.; Reddi, R.K.; Das, K.; Seetharam, D.P.; Sinha, A.K., "Power grid transient stability prediction using wide area synchrophasor measurements," Innovative Smart Grid Technologies (ISGT Europe), 2012 3rd IEEE PES International Conference and Exhibition on , vol., no., pp.1,8, 14-17 Oct. 2012

doi: 10.1109/ISGTEurope.2012.6465752

[8] Le Xie; Yang Chen; Kumar, P.R., "Dimensionality Reduction of Synchrophasor Data for Early Event Detection: Linearized Analysis," Power Systems, IEEE Transactions on, vol.29, no.6, pp.2784,2794, Nov. 2014

doi: 10.1109/TPWRS.2014.2316476

[9] Trachian, P., "Machine learning and windowed subsecond event detection on PMU data via Hadoop and the openPDC," Power and Energy Society General Meeting, 2010 IEEE , vol., no., pp.1,5, 25-29 July 2010

doi: 10.1109/PES.2010.5589479