Machine Learning and Proactive Network Maintenance: Transforming Today's Plant Operations

A Technical Paper prepared for SCTE by

Brady Volpe
Founder and CEO
The VolpeFirm and NimbleThis
3000 Old Alabama Rd. Suite 119-434, Alpharetta, GA 30022
+1-404-954-1233
brady.volpe@volpefirm.com :: brady.volpe@nimblethis.com

Berk Ottlik, Intern, NimbleThis LLC

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

0. Introduction

Proactive Network Maintenance (PNM) aims to proactively determine issues in a network so higher quality service can be provided and service impairments can be fixed before subscriber's experience issues. PNM can leverage updates in Data Over Cable Service Interface Specification (DOCSIS), an international telecommunications standard that enables high-bandwidth data transfer through existing cable television systems. The introduction of many PNM related test metrics has made it possible to pinpoint the root causes of issues in an HFC network. Full band capture (FBC) data allows operators to have visibility into all downstream RF signals anywhere DOCSIS 3.0 or 3.1 modems are deployed. This eliminates the need to bring spectrum analyzers to customer homes and perform inspection. Through PNM, downstream RF signals can be monitored 24x7x365 just using the subscriber's cable modem, which leads to better performance and impairment resolution. Issues can be identified and located faster, leading to greater cost savings and improved subscriber experience.

A challenge for operators is manually analyzing the FBC data from thousands or millions of modems. Further, the cable operator must be able to determine if RF impairments in FBC data are associated with a single home or multiple homes. When an impairment impacts a single home one can usually assume sending a technician to the individual home is the correct action. However, when multiple homes see the same impairment, sending a technician to a single home is almost always the wrong answer as the impairment is in the outside plant. In this scenario, rolling a truck to a single home for an outside plant impairment wastes time, money, extends MTR and annoys the subscriber.

This is where the power of machine learning and PNM shine. Machine learning can quickly analyze the data of thousands or millions of modems in just minutes. Then it will lead the end user to determine if there are impairments and if so, where the impairments are located.

This paper will discuss the type of RF impairments observable by PNM. Next it will discuss how machine learning is used to analyze impairments using an unsupervised model. Then it will look at how machine learning is combined with CableLab's spectral impairment detector (SID) to substantially improve on SID's impairment classifiers. Finally, the paper will look at how the author is using gamification to use feedback from end users to migrate to a supervised learning model.

Enjoy.

1. Types of Impairments

It is important to provide a brief understanding of the typical types of impairments that are generally found using FBC. Rather than using the standard impairment chart kindly produced by Larry Wolcott of Comcast, this document will demonstrate the same impairments, but with new charts. These charts are taken from live cable operator plants using a PNM application.

In the next sections, we will be focusing on the following impairments: adjacencies, suckouts, resonant peaks, rolloff, standing waves, and tilt.

1.2.1 Adjacencies

Adjacencies are essentially misalignments of radio frequency (RF) channels where adjacent channels have a large delta in channel power. This can result in packet loss, video tilting, freezing, and black screens. These impairments can often affect multiple cable modems (CMs) downstream, meaning that clustering and especially localized clustering could find common impairments to better be able to identify the root cause of an adjacency [7].

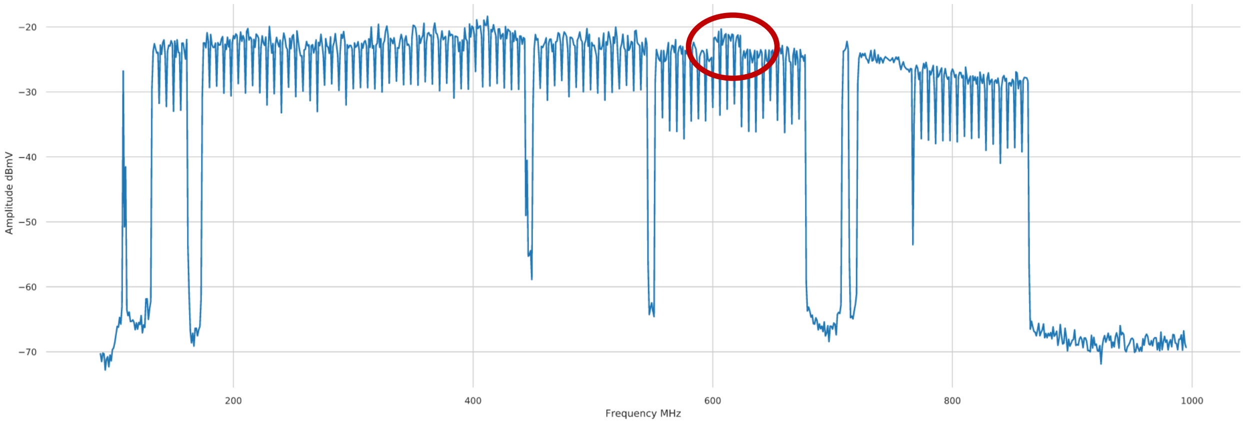


Figure 1.1 Example of Adjacency at around 600 MHz

1.2.2 Suckouts

Suckouts are another type of RF impairment that span multiple channels. They dip down to a certain depth to make a V-shape in the signal. The depth and width of these impairments determine the severity and effect and may often not have any significant impact on customer performance [7].

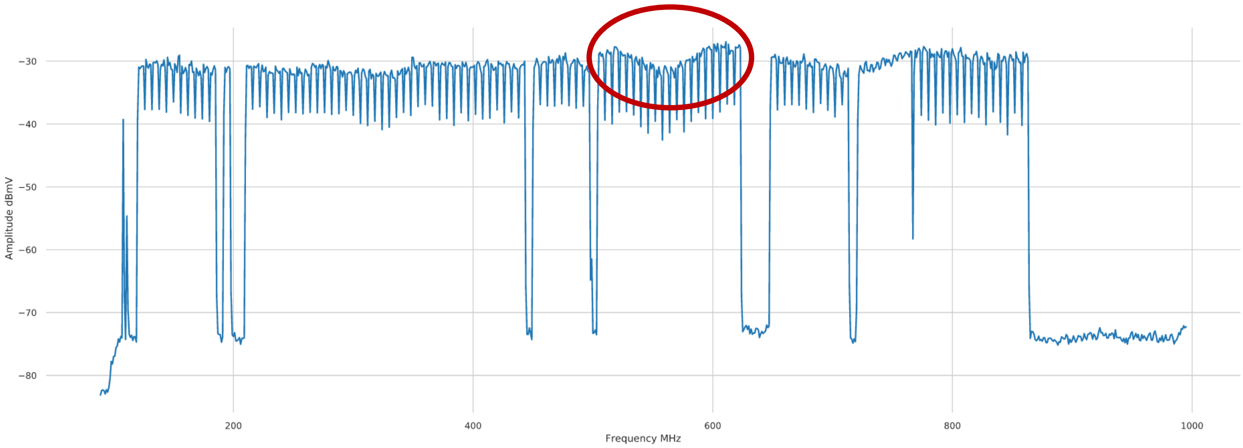


Figure 1.2 Example of a Suckout at around 550 MHz

1.2.3 Resonant Peaks

Resonant peaks are another impairment that usually spans multiple RF channels. They look like inverse suckouts, forming mountain-like peaks in the signal. They can be quite sporadic, forming and disappearing quickly, due to factors such as temperature and can have a wide range of performance impacts including packet loss, tiling, and freezing [7].

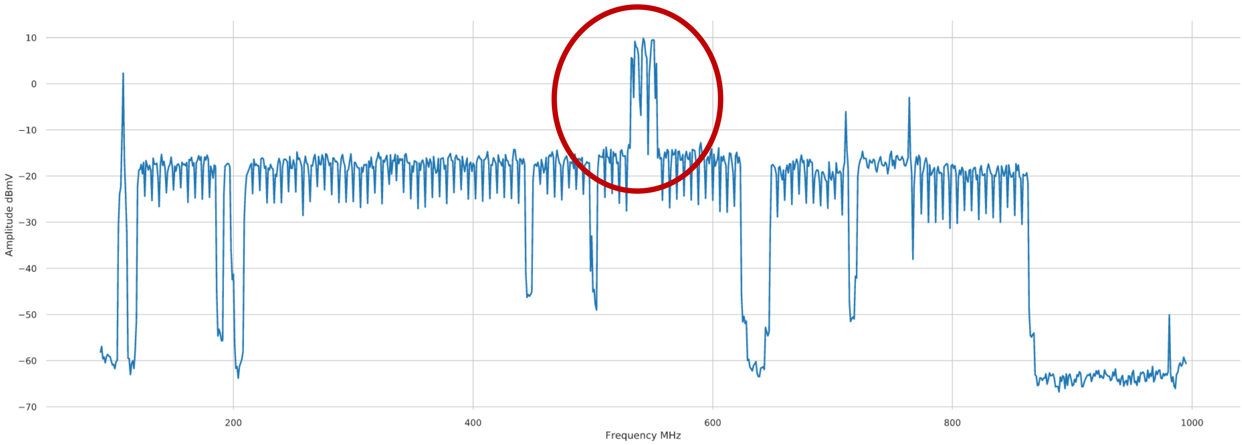


Figure 1.3 Example of a Resonant Peak at around 520 MHz

1.2.4 Roll Off

Roll off is an impairment characterized by a gradual, non-linear, exponential looking decrease in amplitude and power. Roll off can have many causes including old cables being used or individual elements in the network being configured incorrectly. It can cause freezing or tiling of video channels and is, unfortunately, one of the more common RF impairments [7].

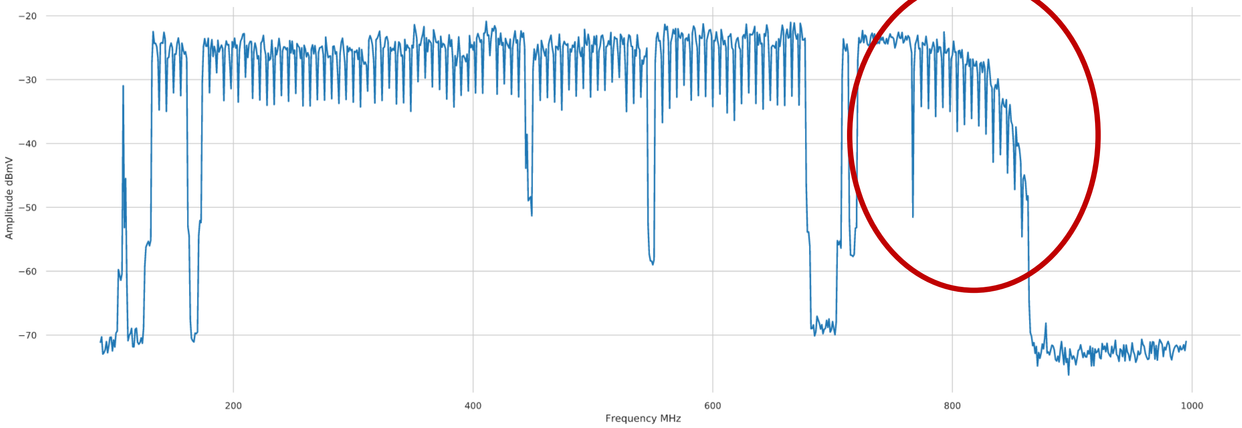


Figure 1.4 Example of Roll Off at around 820 MHz

1.2.5 Standing Waves

Standing waves are RF impairments which affect the entire spectrum. They are usually caused by an impedance mismatch in the signal and appear as waves seen at the peaks of the signals [7].

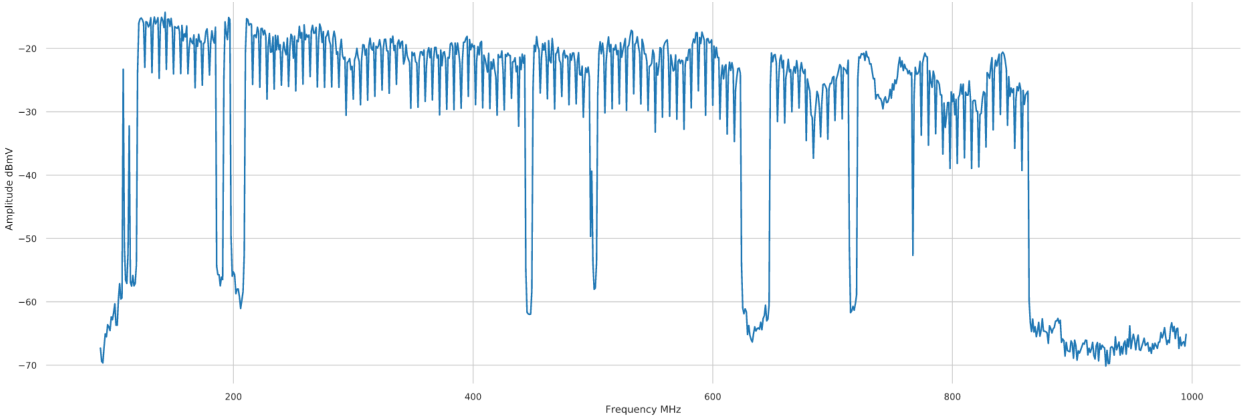


Figure 1.5 Example of a Standing Wave

1.2.6 Tilt

Finally, tilt is an impairment that is characterized by amplitude differences between higher and lower frequencies. There can be a positive or negative slope to a CM. This impairment does not always cause issues for customers [7].

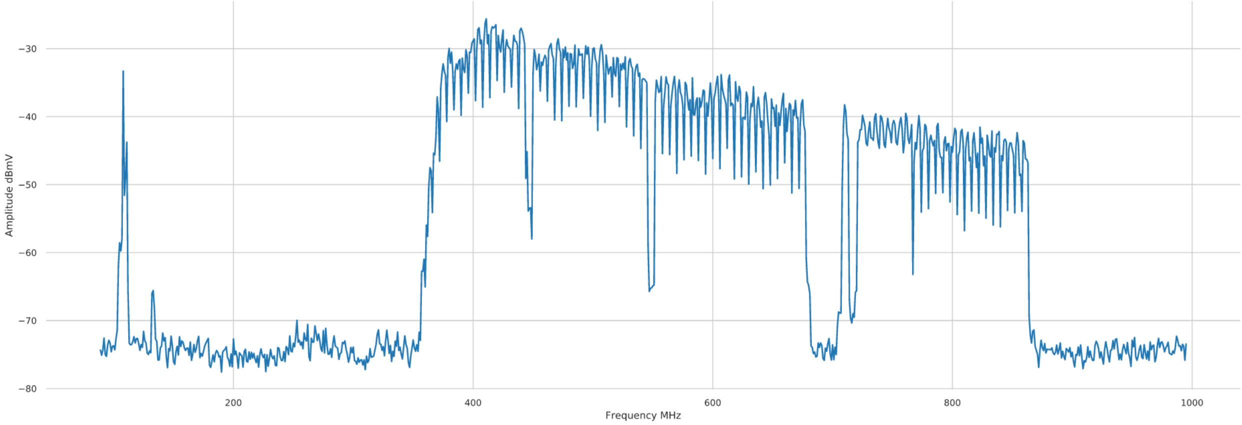


Figure 1.6 Example of Tilt

1.3 Purpose of Clustering

The purpose of clustering is to be able to find shared impairments between CMs in an automated method. This allows us to determine if an impairment affects one cable modem (CM) or if it affects multiple CMs so that cable operators can better pinpoint issues in their systems. As mentioned previously, the focus on clustering is to determine; do we roll a truck to the subscriber's home or to the outside plant. Getting this right results in immediate time and cost savings.

The objective is to find both global and local clusters of impaired modems. Global clusters are clusters where the entire signature of cable modems matches tightly while local clusters have similar signatures or impairments in localized regions. Examples of localized clusters may be things such as a shared suckout between multiple cable modems.

1.4 Purpose of SID Overlays

The final step is to overlay CableLabs spectral impairment detection (SID) impairment labels to validate SID outputs and generalize SID predictions to more CMs. SID is software that can identify and localize common RF impairments in FBC data. It was created by the CableLabs PNM working group and has many thresholds that must be tuned to optimize performance. This does mean however that it often makes mistakes in correctly identifying impairments such as suckouts, standing waves, etc.

By overlaying SID detections on existing clusters, it is possible to validate the SID impairment labels and generalize them to the rest of the CMs in the cluster. Further, if a certain percentage of CMs in a cluster share a common SID label, then these SID overlays can be applied to both global and local clusters. Once the ML model has identified a high correlation of modems having impairments detected by the ML engine and by the SID engine, it is possible to label the impairment to the end user with a high degree of confidence.

2. Technical Approach

It is assumed that the reader has some knowledge of machine learning from previous SCTE or other papers on this topic related to ML and FBC. For this reason, topics such as unsupervised learning, supervised learning, features, and general machine learning terminology will not be covered. It will be up to the reader to review currently available documents as referenced at the end of this document.[1][5][11]

The next sections discuss how FBC data must be manipulated prior to any machine learning analysis. From a development standpoint, this is where the most work occurs in a machine learning exercise. It is often said that machine learning is easy, but it’s the preparation of the data that is hard. Meaning bad data in means bad data out. To get to a good machine learning model means lots of pre-work.

While the following sections may initially appear a bit intimidating, it is important to note that the work being shown is fully automated in a PNM application. The user need not know any of the mechanics behind the machine learning. From the end user’s perspective, the result is displayed data which is easily actionable because now meaningful data is being presented. Machine learning just did the hard work of sifting through piles of data for the user.

Now, let’s look at the technical approach for giving the end user a meaningful experience.

2.1 Pre Processing

Preprocessing steps are necessary to manipulate FBC data to have an optimal performance with clustering.

2.1.1 Downscale Data

The raw FBC data has a varied sample rate and a varied number of data points per cable modem. The clustering algorithms that were used however require all the data to have the same dimensions. Also, it makes the code simpler and faster since it allows for NumPy arrays to be used for many operations [8]. The data is downscaled to one datapoint per integer frequency from 89 to 996 MHz. Simply put, the corresponding amplitude for the sampled frequency closest to each integer frequency is kept and stored to end up with 907 data points per cable modem.

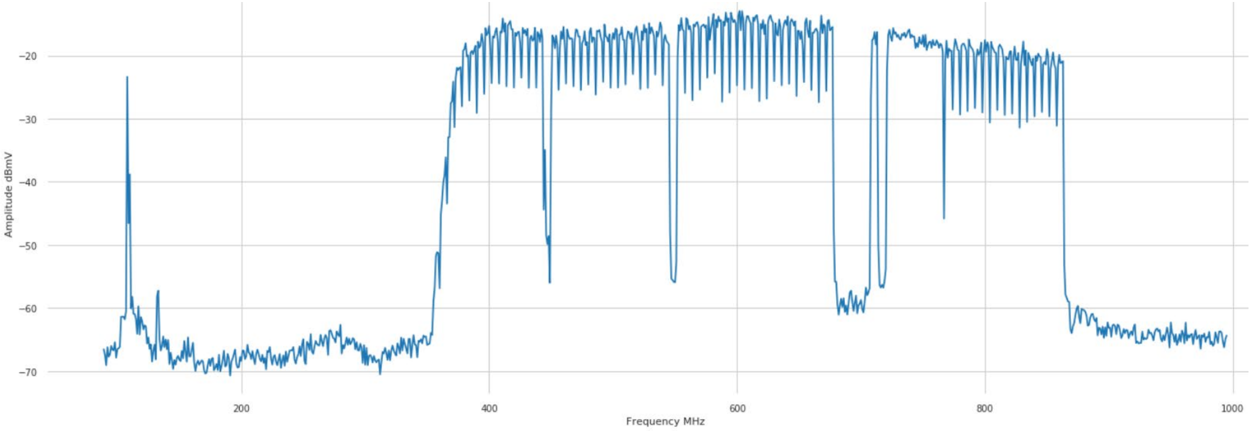


Figure 2.1 FBC Data Scaled Down

Additional downscaling was tested by using the data points closest to every other frequency from 89 to 996 MHz since most impairments did not lose any resolution and instead only some noise in the signal was eliminated.

2.1.2 Rolling Median

A rolling median is simply a median calculated for a certain window size passed over an entire signal [6]. This can remove very small and unimportant variations in the signal which would otherwise introduce unnecessary noise and produce incorrect clusters. Additionally, the advantage of a rolling median over something like a rolling mean is that it can filter out the gaps between channels. A rolling mean is not resistant to outliers meaning those dips would have a large impact on the surrounding regions. The median was calculated from the center meaning that n number of data points on the left and n number of data points on the right were used to find the median for the center point.

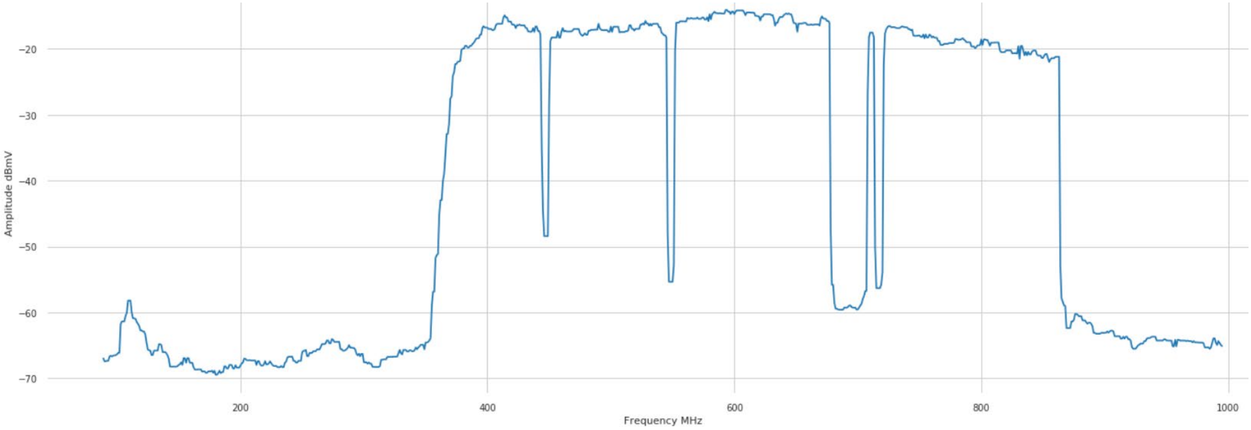


Figure 2.2 FBC Data Scaled Down with Rolling Median

2.1.3 Transforming to a New Center

When creating clusters, the power of a CM is irrelevant because impairments come from the deviations in a signal. Therefore, the signal was transformed up to 0 dBmV for the average of a certain region of the signal. The area of the signal was chosen to be 820-850 MHz because this results in transforming modems with rolloff much higher than other modems and therefore easily filtering them out (as seen with the blue CM in figure 2.3). Note that this arbitrary center frequency must be adaptive. For example, some plants may not have any signals between 820-850 MHz or there may be large impairments in this band. The transformation to a new center is used to find a clean and flat center where machine learning can be achieved. Finally, it was determined that rejecting the FM band (88-108 MHz) improved results because in nearly all cases FM ingress occurred in or near the subscriber home.

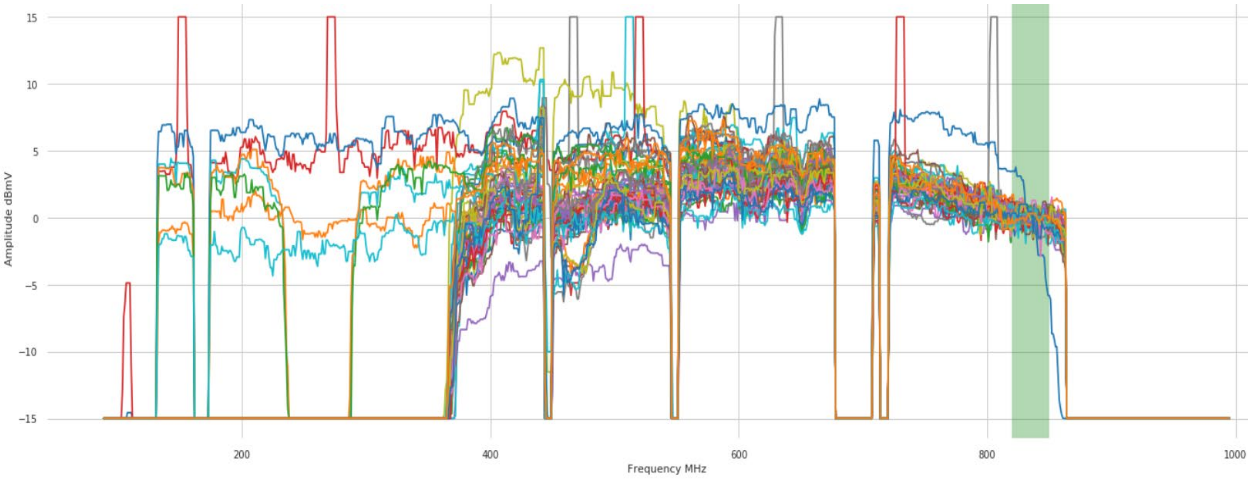


Figure 2.3 FBC Data Scaled Down with Rolling Median and Transformed and Clipped

2.1.4 Normalize Data

Data is normalized from -1 to 1 before any machine learning is performed. Since all data that is clustered on is clipped before clustering, all the data is simply divided by the clipping amplitude maintaining the same shape of the data but reducing the amplitude range to -1 to 1. The purpose of normalization is to have a standard range of the data so that any predetermined parameters work optimally [5]. Data normalization is a very common machine learning practice.

2.2 Modem Health Classifier

The first step in clustering aims to classify each modem in a node as impaired or not impaired. These classifications are done in an unsupervised manner as no accurate ground truths currently exist on which to create a supervised learning model. For this system to work properly, an assumption is made that the majority of CMs in a node are healthy. This is a bold assumption, but one which must be taken until 1) proven otherwise by inspection or 2) a supervised machine learning model is available. Note that as discussed later in this document, by using SID data, it is possible to overcome 1) above with the use of SID data. This is because SID data will provide the necessary information that all modems in a node have some type of impairment. Once a cluster is created, if all modems in the "healthy" cluster are shown to have SID impairments, then by inspection of the SID data it can be determined that the cluster is not healthy.

2.2.1 Region Identification

The first step of classifying modems as healthy or not was to identify the sections of RF spectrum that occupied by video or QAM channels. For the purpose of this document, occupied spectrum are named regions. Identifying regions is a necessary step because otherwise, classification could be made on modems with varying sections of used and unused spectrum which would be more quickly classified as an outlier and impaired than any modem which has an actual impairment. In Figure 2.5, an example of unused spectrum is the spectrum below about 375 MHz, where the blue line drops to -1.0. For someone familiar with the industry, it is apparent that a data only filter is in use in Figure 2.5. The data only trap causes all signals below 375 MHz to be attenuated. The signals that are still present below 200 MHz are FM ingress signals (88-108 MHz). As previously mentioned, the FM band is omitted from machine learning in the current model, so these low frequency signals will be ignored.

The process starts by roughly identifying all the used frequencies of every modem in a node (see figure 2.5 and 2.6).

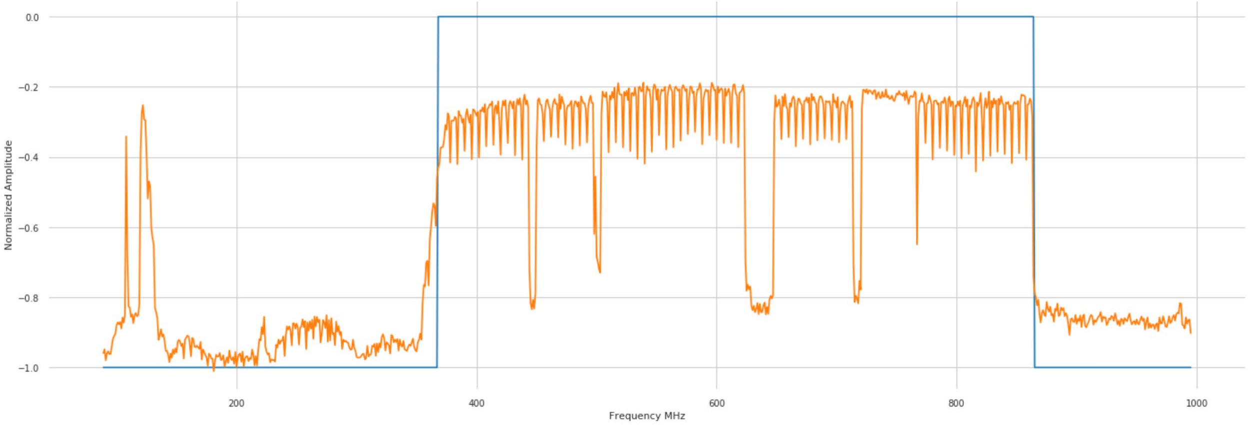


Figure 2.5 Used Frequencies Identified on Modem

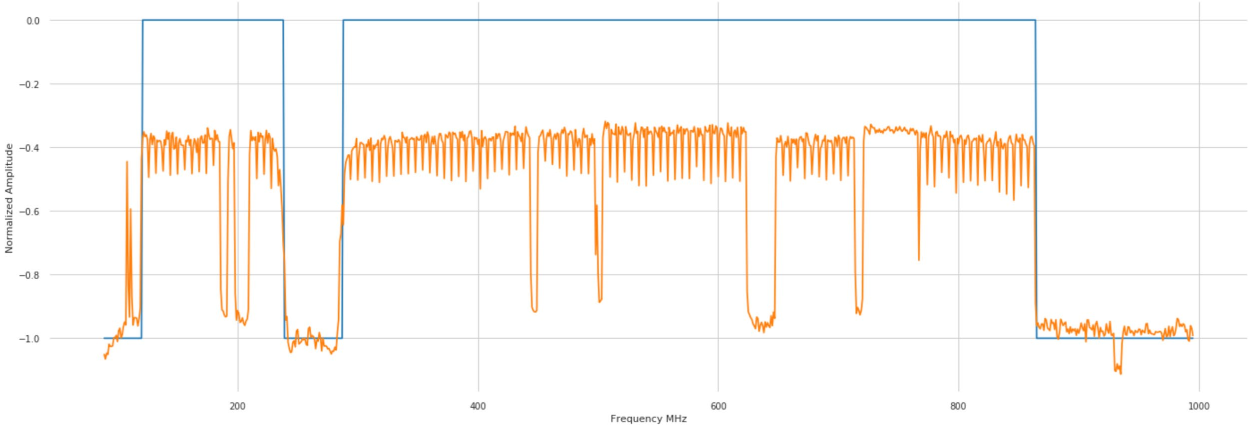


Figure 2.6 Used Frequencies Identified on Modem with Band Stop Filter

In Figure 2.6, it is evident that the pre-processing can identify unused frequencies, but this time it is not due to the presence of a data only filter. In this case, there are channels missing between 275-300 MHz. The blue line indicates pre-processing is eliminating these frequencies from the model. While there are unused frequencies higher in the band around 625 MHz, these do not the minimum width for the classification engine to exclude the band.

The used spectrum is found by first passing a rolling median with a large window size over the normalized, transformed, and clipped signals from section 2.1.3 (see figure 2.7). Then, at any frequencies where the normalized amplitude is greater than -1, those regions are labeled as used (blue line at 0) and any frequencies where the normalized amplitude is -1, those regions are labeled as unused (blue line at -1).

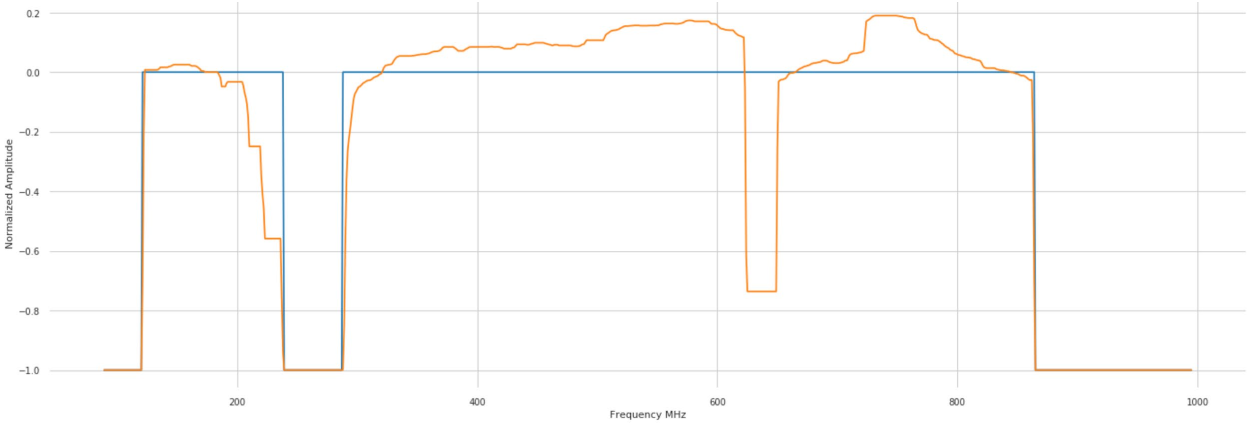


Figure 2.7 Used Frequencies Identified on Smoothed Modem with Band Stop Filter

Once all the modems have their used regions extracted, regions can be extracted representing the various regions of commonly used frequencies between modems. This works by first identifying the regions with the most used spectrum of all modems (see figure 2.8 for used spectrum of all modems). In the case of this node, that region is from around 380-850 MHz. Then, health classification is done on this region as described in region 2.2.2. Following this, all modems will be identified as impaired or not impaired for that region. This region is then removed as being a used spectrum from all modems which have used spectrum in that region. Following this, the process repeats for the next region with the highest number of commonly used frequencies by classifying and then removing that region from the used regions as well. This process is repeated until there are no more used regions in the spectrum. At this point all modems have been classified as impaired or not impaired based on regions where modems have the same used and/or unused spectrum.

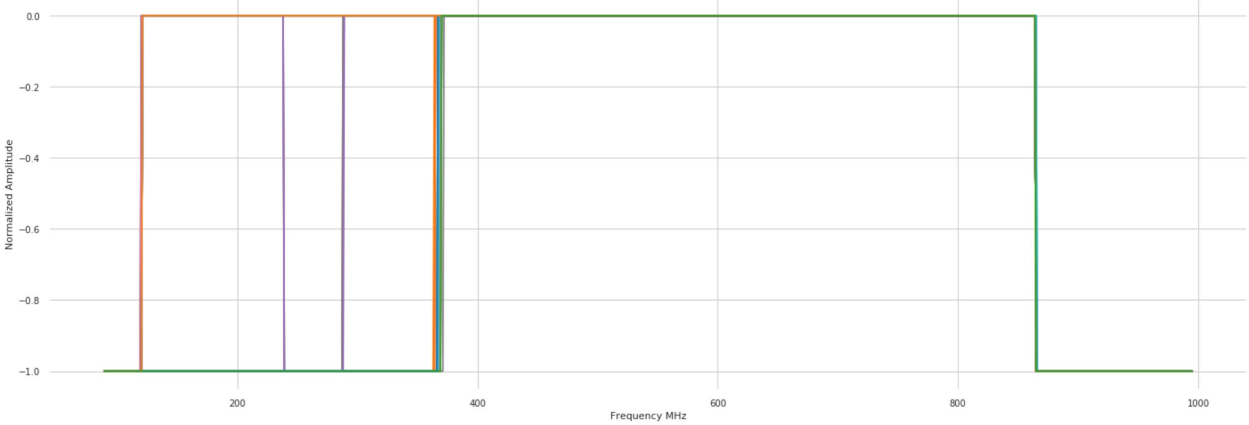


Figure 2.8 Overlapped Used Frequency Regions

2.2.2 Health Classification on Extracted Regions using Local Outlier Factor

Once common regions of used frequencies are extracted, outlier detection is done on modems which have the regions. This outlier detection utilized the local outlier factor (LOF). LOF is an unsupervised (well, semi-supervised) machine learning algorithm that uses the density of data points in the distribution as a key factor to detect outliers, LOF roughly works by calculating a standardized distance to n number of neighbors and labeling data points with a large distance as outliers [1]. In turn, this can filter out modems which deviate from the rest of the modems due to impairments.

2.3 Global Clustering

The next part of the clustering is to cluster together the impaired modems (outliers from clustering) to determine if there are any similar global impairments between cable modems. Global clustering requires that signatures of CMs match throughout the entire frequency range, meaning that it cannot cluster together smaller local clusters, but it can find modems with common signatures.

Global clustering is useful in finding modems that have large impairments such as standing waves or tilt, which impact the entire spectrum.

2.3.1 DBSCAN Clustering

In machine learning there are many algorithms for grouping or clustering common sets of data together. One of the most used is K-Means. K-Means clustering works very well, however it is non-optimal for noisy data, such as FBC data. The current implementation in this paper is using a model called DBSCAN.

DBSCAN stands for density-based spatial clustering of applications with noise. The algorithm works by first selecting a random point in the data. Then, if there are minimum points (a specified parameter) number of data points within the radius of Epsilon (EPS, a specified parameter) distance or the Euclidean distance, straight-line distance to the original point, it is labeled as a cluster. Then this process repeats for every point that was in the original cluster, if the points have at least minimum points number of points within their EPS distance, then the points are labeled as core points. However, if a data point does not have the minimum points number of data points, it is labeled an outlier, unless it is within the EPS of a core point. If there are no more data points nearby, then a new random point is chosen until all the data has been clustered [3].

2.3.1 Global Impairment Clustering

Global impairment clustering was implemented purely using DBSCAN. The minimum points parameter is lowered to two to allow for very small clusters and the EPS is slightly lowered also to ensure clusters are tight. Figure 2.8 demonstrates the use of global clustering but also highlights some of the weaknesses.

Global clustering can cluster these modems and conclude that they are all impaired and have both a common impairment and rest of signature. What the impairment is or where the impairment is located cannot be determined purely using global impairment clustering (this is where SID overlays are needed).

The issues that arise from this approach is that if that same suckout were to be seen around 470 MHz on a different modem with used spectrum below 350 MHz, then they would not be clustered together. This is when local clustering is effective.

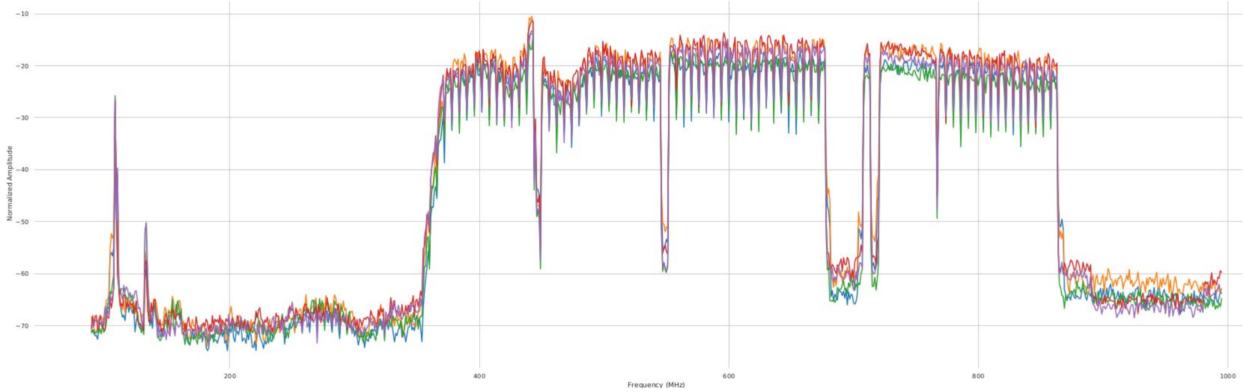


Figure 2.9 Global Cluster on Impaired Modems

2.3.2 Global Impairment SID Overlay

SID labels are overlain on global clusters to determine the location and type of impairments seen in a cluster. If a certain percentage of modems have an impairment, then that impairment can be generalized to the rest of the modems. Additionally, if that impairment is found in a local region for multiple CMs, then it can be generalized that those modems all have that impairment at that specific location. This is beneficial for two reasons:

1. The data being used is unsupervised data, which means it is not known if the FBC data is impaired or not impaired, but overlaying SID data, it is now possible to determine not only if the FBC data is impaired, but also the type of impairment (i.e., suckout, standing wave, etc.).
2. SID impairments are often inaccurate. For example, suckouts and adjacencies are often mis-labeled by the SID engine. By classifying many modems with the same SID overlay, it is possible to improve the accuracy of SID classifications through scale. If many modems show the same SID impairment at the same frequency, then the probability that SID is accurate is high.

SID overlays also allow the algorithm to discard SID impairments that are rarely found in the cluster. In figure 2.10 it is apparent that only 20% of the CMs in the cluster were labeled as having a resonant peak by SID. From this one can conclude that the modems in the cluster most likely do not have a resonant peak at around 620 MHz.

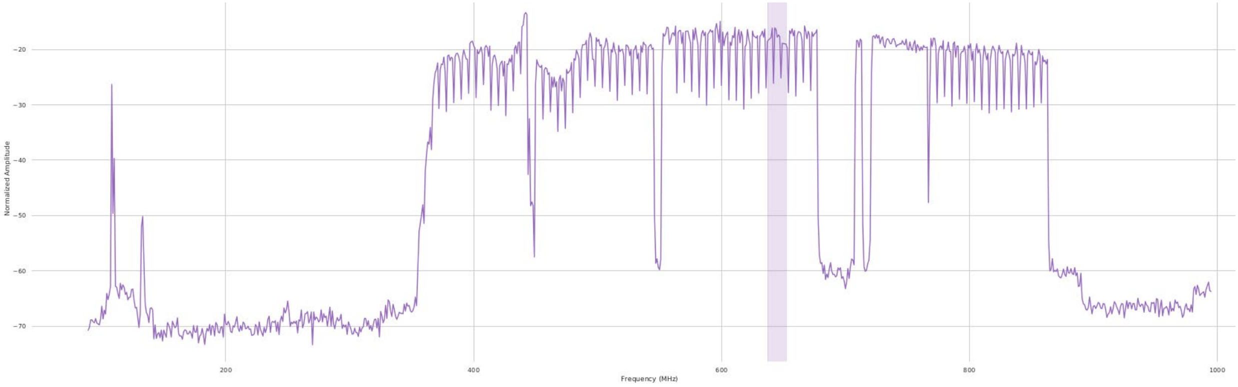


Figure 2.10 Resonant Peak Identified at 20% in Global Cluster shown by vertical purple bar

On the other hand, SID labels can be verified based on the percentages that they occur at and if the location that they occur at overlap. In Figure 2.11, it is visible that 100% of modems are labeled as having a standing wave from around 100-900 MHz. Therefore, one can accept this label as being most likely true. This can also apply to other labels which are not seen in 100% of modems however such as the adjacency labeled in 40% of modems for this cluster around 650 MHz. This can be generalized to all modems in the cluster if the threshold of the percentage of modems in a cluster that need to have a SID label for a certain region is met.

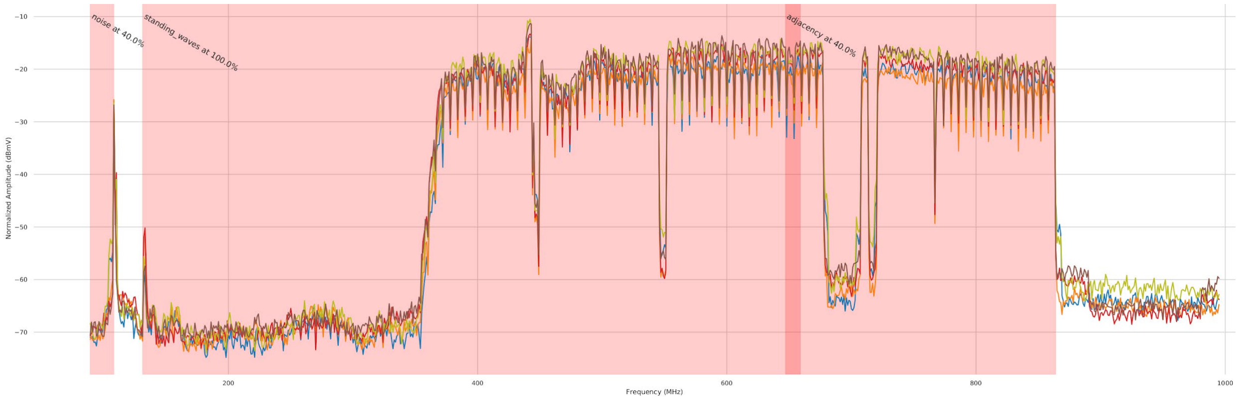


Figure 2.11 SID overlay with 40% Threshold

2.4 Local Clustering

To cluster together similar local impairments, different clustering techniques need to be used which only look at local regions. This is because CMs may share one common impairment, while not sharing a whole separate range of impairments and signatures which result in them being placed in different global clusters. Again, this is very important for narrow impairments such as suckouts and adjacencies.

2.4.1 Local Impairment Clustering

To find local impairments, clustering was done using DBSCAN on a certain window throughout the spectrum. This window slid over the entire FBC spectrum with a certain step size and clustered together all the modems in the node using DBSCAN [10]. The FBC spectrum was preprocessed the same way as with variability-based outlier removal (section 2.3.2) so that the FBC data was straightened and centered at 0. This was important because DBSCAN would not work properly unless the data was in the same shape and located at the same amplitude because that is the only way for the Euclidean distance between sections of the FBC data to be small and therefore clustered together.

Then if one of the clusters formed only contains impaired modems, it is a localized impaired cluster in that region. If a cluster contains modems that are not impaired, it means that the cluster found similarities that are not an impairment, and the cluster is therefore not considered a localized impairment cluster. Additionally, if a cluster contains more than a certain percentage of the modems in a node, it was removed from the local impairment clusters because it most likely is clustering on something that is not an impairment.

In figure 2.11 we can see that the localized impairment clustering was able to find local clusters due to the preprocessing steps taken. Preprocessing manipulated the data to be in the shape seen in figure 2.12. Here DBSCAN can easily find clusters in certain windows even though the orange modem has many used channels under 400 MHz that the blue modem doesn't and that the blue CM has tilt but the orange one doesn't. This resulted in a important breakthrough, which was the ability to detect similar impairments on modems with radically different spectrum usage due to bandpass and band stop filters.

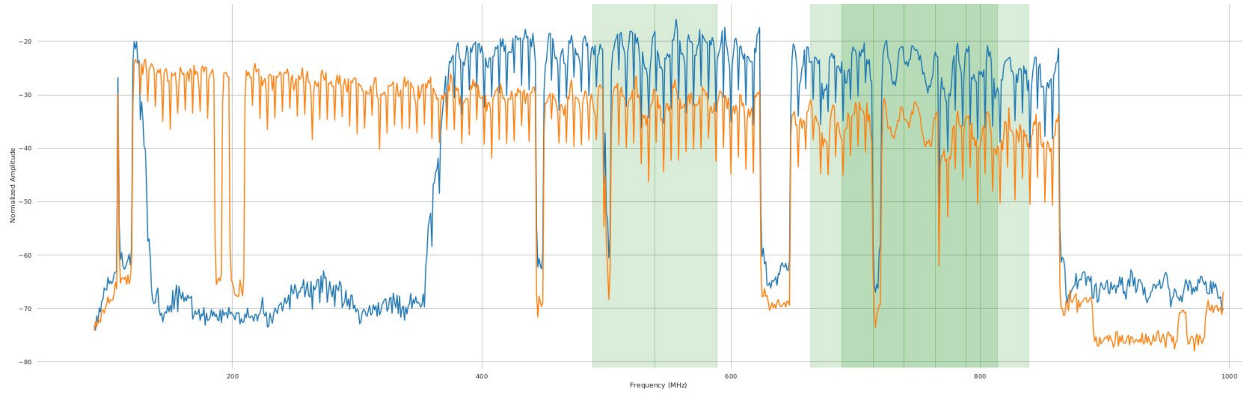


Figure 2.12 Localized Impairment Clusters on Impairment Modems

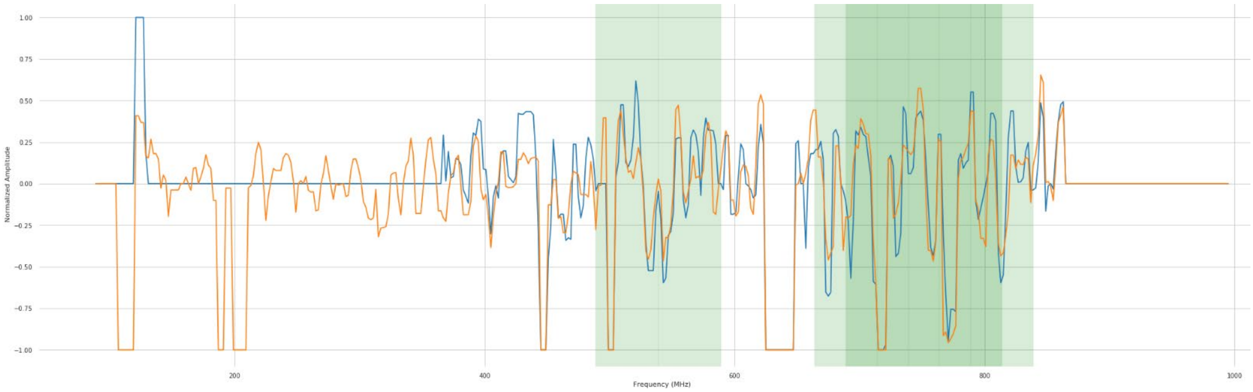


Figure 2.13 Localized Impairment Clusters on Impairment Modems Pre-Processed

2.4.2 Local Impairment SID Overlay

The last step with the localized clustering was to overlay the SID impairment labels for localized impairments (i.e., suckouts and adjacencies). This allowed us to draw conclusions about the accuracy of the SID labels and see if the labels intersected with the local cluster regions. In all the figures below, the green highlights indicate local cluster regions while the red indicates the global cluster regions.

In figure 2.13, we can see that SID labeled that both modems in the same local cluster had standing waves and that the labels intersected with regions of local clusters, meaning that we can conclude that both these modems have standing waves in the regions of overlap between the SID label and local cluster regions. We cannot generalize and say that both modems have standing waves on all regions from 100-850 MHz however because since this is local clustering, the regions could have nothing to do with each other.

As seen in figure 2.13, local impairment clustering can also be used to generalize impairments that are not detected in all modems such as the resonant peak around 530 MHz. Even though only 50% of the modems in the cluster had this label, it can be generalized to apply to both modems because it falls under a local clustering region. Figure 2.14 contains more visuals of a different local impairment cluster.

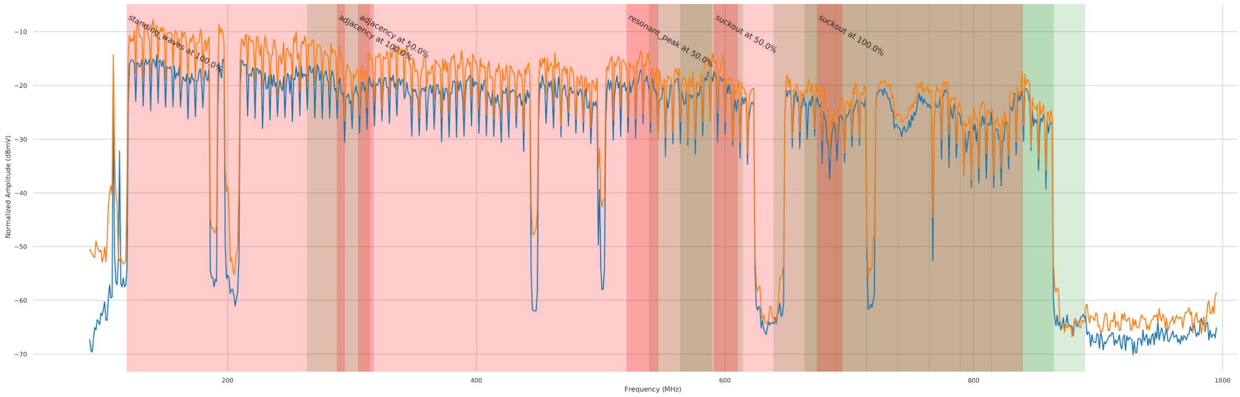


Figure 2.14 SID Overlay on a Local Cluster with Threshold of 40%

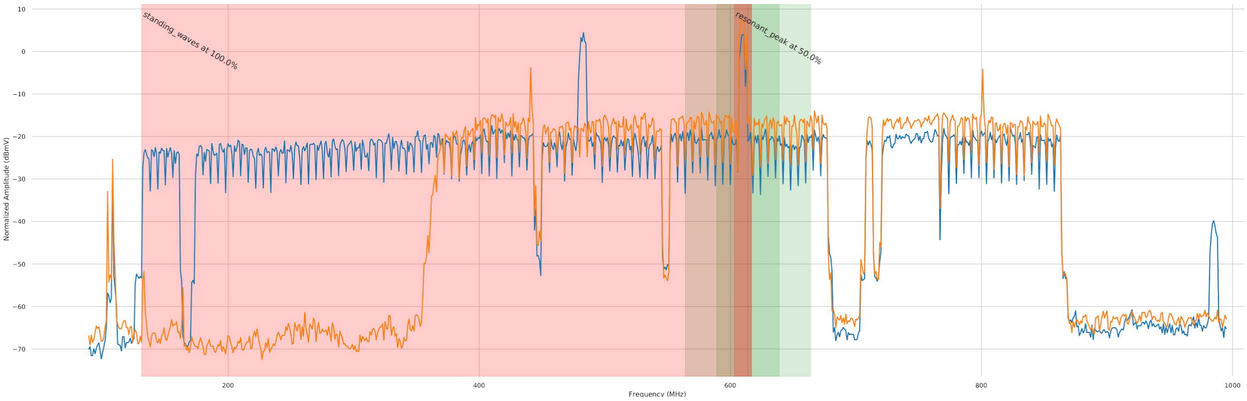


Figure 2.15 Resonant Peak SID Overlay

3. Results and Discussion

This section will analyze some of the performance, nuances, optimizes and observations identified during the use of machine with full band capture data. As this technology is continuously being improved upon but not only the author of this paper, but others in the industry, it is the hope that some of the findings in this paper will be used by others to build upon this and a collaboratively shared for the betterment of the industry.

3.1 Modem Health Classifier

3.1.1 Pre Processing Performance

The preprocessing steps were vital to the clustering. By clipping and centering the signals, small local impairments such as resonant peaks and suckouts become much more influential in the data and are therefore easily labeled as outliers only using DBSCAN.

Preprocessing does occasionally run into issues, however. Occasionally, minor impairments such as the red spike around 600 MHz in figure 3.1 are run over in preprocessing as seen in figure 3.2. This is most likely a result of the spike being very thin and the rolling median, therefore, discarding it. The spike is most likely the result of RF interference that could impact customers and was missed.

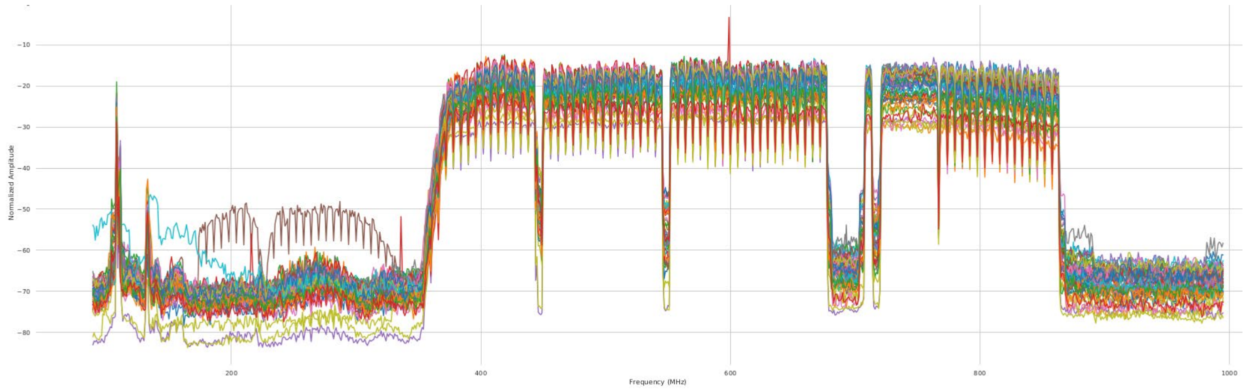


Figure 3.1 Raw FBC Cluster

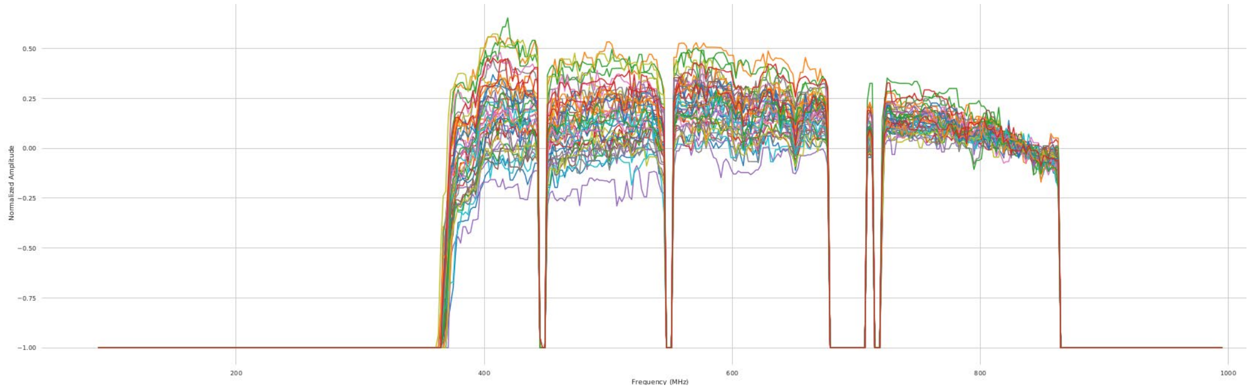


Figure 3.2 Pre-Processed FBC Cluster

3.1.2 LOF Performance

LOF proved to be optimal for this use case because it accurately and efficiently was able to identify outliers with impairments. Finding the optimal number of neighbors was a difficult task and will continue to be a difficult task when applying this software to various CMTSs and cable operators. Figure 2.14 shows LOF finding outliers on the highlighted regions while figure 2.15 shows the modems not labeled as having outliers.

Figure 3.3 shows a cluster of modems which has been continuously showing up across cable operator systems since this algorithm has deployed. Notice the spikes appearing roughly 20 dB above nominal RF spectrum. When examined more closely, these spikes are roughly 8 MHz in bandwidth. The author of this paper has visited several subscriber locations where the signal is present. The signal was not observable using traditional swept-spectrum analyzers. Further, once the subscriber modem was replaced the modem was replaced, the signals were no longer present. Suspect modems were collected and are under current evaluation.

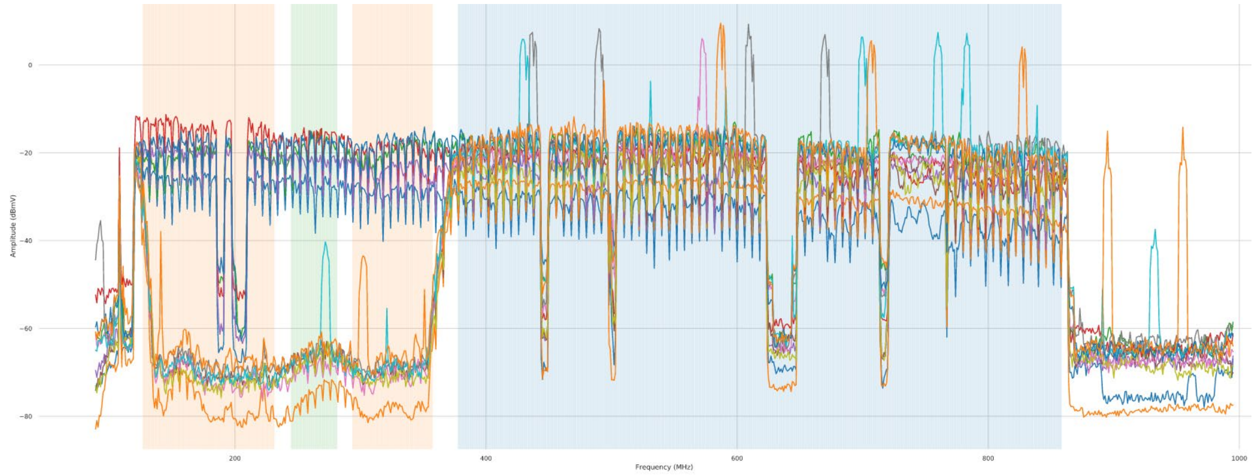


Figure 3.3 Impaired Modems with Used Regions Indicated

Figure 3.4 shows what is expected for a healthy cluster of modems on the same node. Here we see many cable modems not showing the spike in figure 3.3. The point of this example is that although root cause is not yet determined, machine learning combined with PNM was successful in identifying anonymous activity in modems which were customer impacting. Each customer with spurious activity had open tickets for downstream video or DOCSIS issues which where un-resolved.

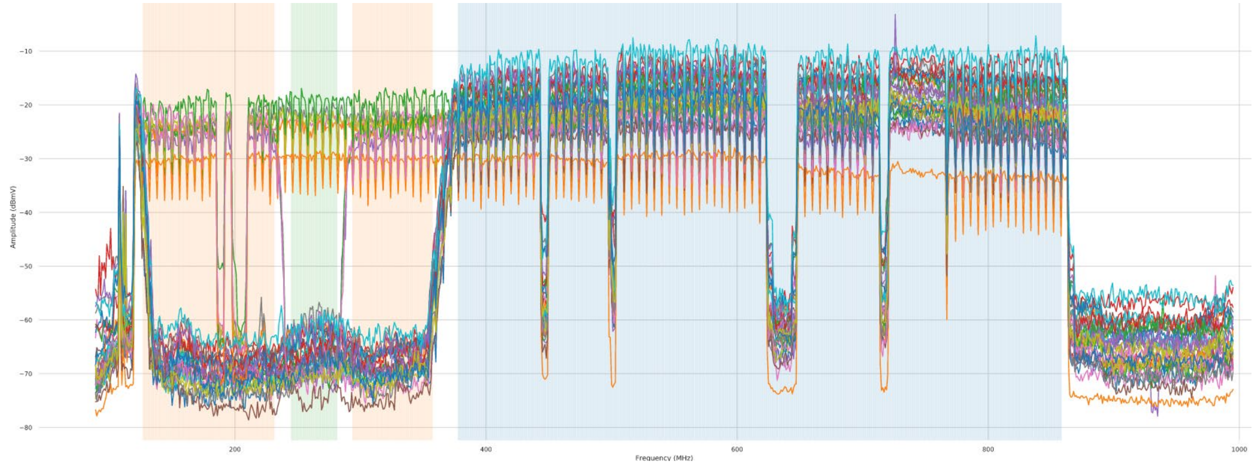


Figure 3.4 Healthy Modems with Used Regions Indicated

3.1.3 Fixes for Band Stop Filters

Band stop filters filter out ranges of frequencies in a signal [2]. They are used by cable operators to restrict certain channels from customers. In the data we used for this paper, band stop filters were occasionally found in CMs as seen in figure 3.3. By using the different regions with commonly used frequencies, modems with Band Stop filters were not automatically classified as impaired, but instead accurately classified based on the rest of the used frequencies and how they related to the rest.

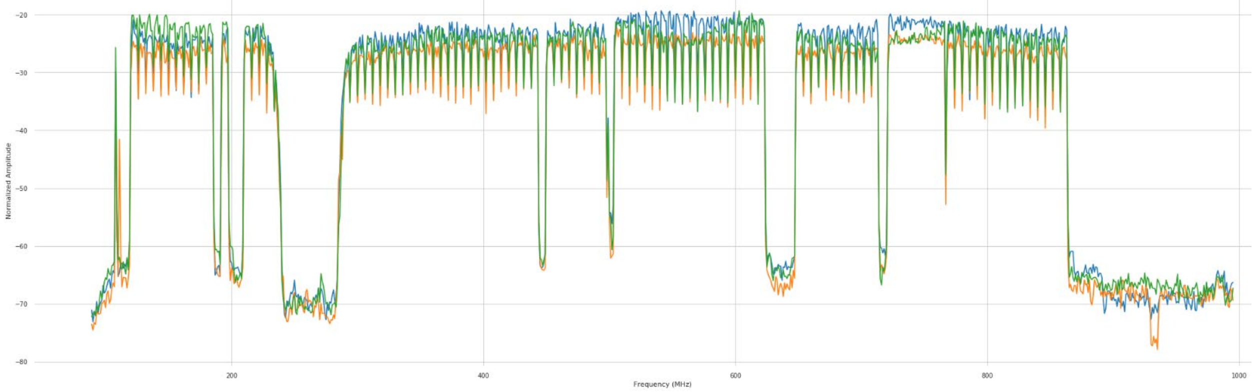


Figure 3.5 Modems with Band Stop Filters

3.2 Global Clustering

Global clustering was effective for its purpose of finding impaired modems with similar global signatures.

3.2.1 Global Clustering Shortcomings

One issue seen with both global and local clustering is when validating SID impairments when there are few modems in a cluster. If there are only 2 CMs in a cluster and one has an incorrect SID label, if the threshold for the correct SID label to be generalized is 50% then the incorrect impairment is accepted. However, if the threshold is at over 50%, correct SID labels may also be overlooked.

3.3 Local Clustering

Local clustering was effective in finding local regions with similar signatures. Occasionally regions that just happen to only be found in impaired CMs are labeled as local clusters when they are completely healthy regions of the spectrum that just have signatures not found in the rest of the healthy CMs.

3.3.1 Local Clustering Shortcomings

Shortcomings include times local clusters were identified but no SID labels were to be found in those regions and when common SID labels were found in regions and no local clusters were identified. Using both the clusters and SID labels, however, allows one to build greater confidence in the SID labels even if the system is not 100% accurate.

Local clustering also has the same issue when there are few cable modems in a cluster as seen with global clustering in section 3.2.1.

3.4 Optimization of Parameters

One factor seen across the board is that there are many parameters with this approach. Optimizing parameters also takes lots of processing and lots of time. Every cable operator with different hardware and different severities may need different parameters and perhaps even different CMTSs. There is no numerical way to find the optimal parameters as this is an unsupervised setting with no ground truths, meaning that someone must look through as much data as possible and look through many variations of parameters to find which fits the given data the best while not overfitting.

There are also many parameters that impact each other such as the window size parameter for local clustering. If the window size is changed, the EPS parameter in DBSCAN must also be changed to adjust. This makes optimization even more complex.

For this reason, real-time optimization parameters are added to the application which enable the end user to modify DBSCSAN parameters. These optimizations on performed on the cable operators' network to ensure DBSCAN is optimized based on the system's channel lineup.

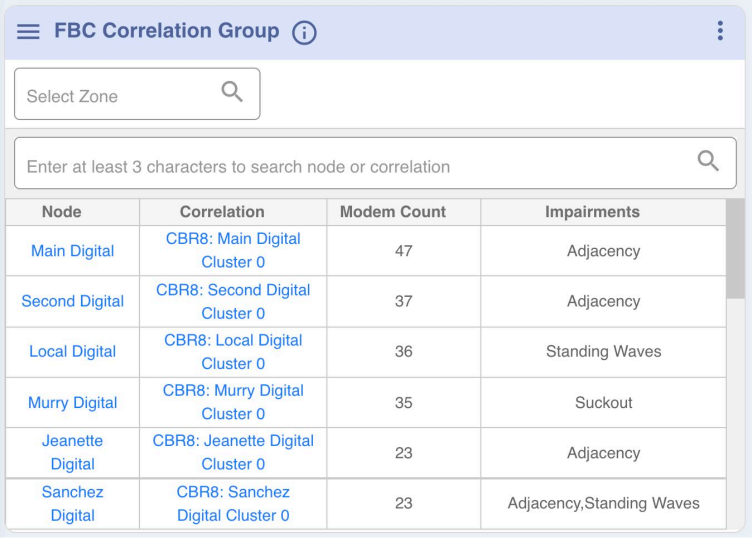
4. Conclusion

This project is effectively able to identify common impairments between CMs and is also a step towards replacing SID with a more intelligent system. Extensive preprocessing and clustering on FBC signatures were able to reveal shared impairments between CMs in a node. Additionally, SID labels were then overlaid to clusters to verify the accuracy of SID labels. This model was further modified and applied to RxMER data of OFDM channels in DOCSIS 3.1 downstreams. This had value in identifying outside plant impairments impacting multiple subscribers with DOCSIS 3.1 modems.

4.0 Operationalizing the Plant Maintenance

PNM and machine learning begin to show their complimentary value when it comes to operationalizing plant maintenance. Before applying machine learning, it would be up to the end user to manually view many fullband capture images and attempt to make mental correlations. This was a tedious process and relied on human to first do the work and second be effective at doing the job. The job being to determine if an impairment was impacting one home or many. Machine learning will automate this task by automatically clustering FBC data. It is up to the application programmer to make the data available to the end user.

One example can be seen in Figure 4.1 where a widget is made available to the end user with a list of FBC correlation groups (i.e. cluster groups), the node each correlation group, the number of subscribers impacted by the impairment and the impairment type. Clicking on the blue hyperlink takes the user to a visual representation of the impairments (shown in Figure 4.2) on a map, where action can be taken.



FBC Correlation Group ⓘ

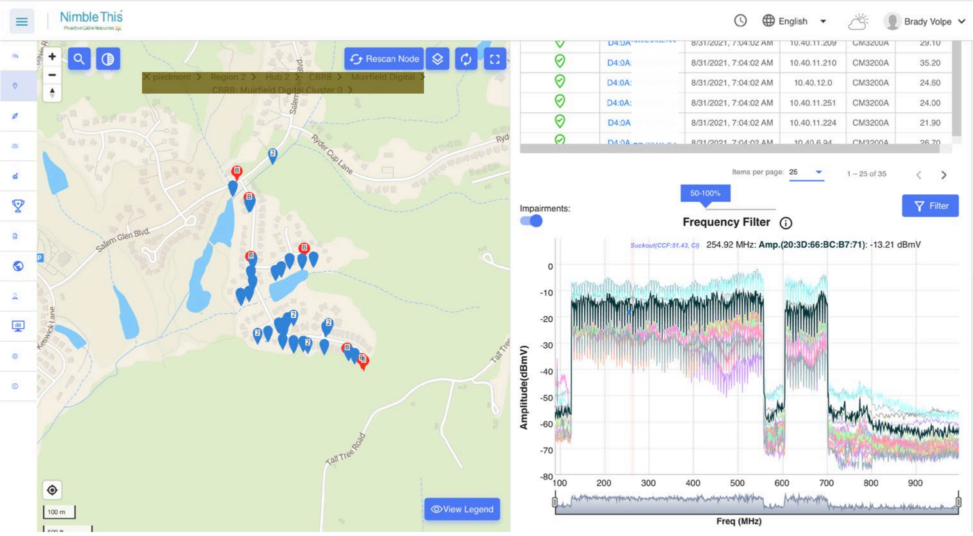
Select Zone

Enter at least 3 characters to search node or correlation

Node	Correlation	Modem Count	Impairments
Main Digital	CBR8: Main Digital Cluster 0	47	Adjacency
Second Digital	CBR8: Second Digital Cluster 0	37	Adjacency
Local Digital	CBR8: Local Digital Cluster 0	36	Standing Waves
Murry Digital	CBR8: Murry Digital Cluster 0	35	Suckout
Jeanette Digital	CBR8: Jeanette Digital Cluster 0	23	Adjacency
Sanchez Digital	CBR8: Sanchez Digital Cluster 0	23	Adjacency,Standing Waves

Figure 4.1: FBC Correlation Widget Groups by Node, Modem Count, and Impairment Type
(Image Courtesy: NimbleThis)

Figure 4.2 the machine learning-based results of FBC correlated modems as plotted on a map. The blue modems on the map are associated with the FBC data on the right-hand side. A user may select a modem on the left, a MAC address on the right, or a trace bottom right to interact with the data. The actionable data for the end user is that this section of coax plant has a system-wide standing wave. Fixing the standing wave by visiting a subscriber's home is not a good choice. This is an outside plant problem which must be addressed as such.



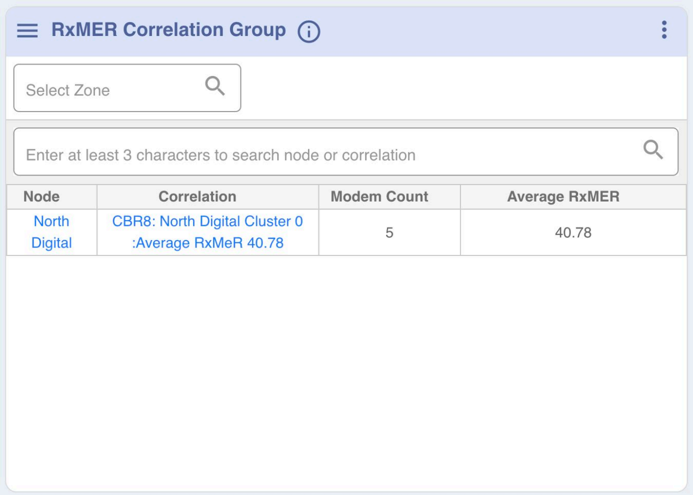
✓	D4:0A	8/31/2021, 7:04:02 AM	10.40.11.209	CM3200A	29.10
✓	D4:0A:	8/31/2021, 7:04:02 AM	10.40.11.210	CM3200A	35.20
✓	D4:0A:	8/31/2021, 7:04:02 AM	10.40.12.0	CM3200A	24.60
✓	D4:0A:	8/31/2021, 7:04:02 AM	10.40.11.251	CM3200A	24.00
✓	D4:0A	8/31/2021, 7:04:02 AM	10.40.11.224	CM3200A	21.90
✓	D4:0A	8/31/2021, 7:04:02 AM	10.40.6.94	CM3200A	26.70

Items per page: 25 1 – 25 of 35

Suckout(CCF:51.43, CI) 254.92 MHz: **Amp.(20:3D:66:BC:B7:71)**: -13.21 dBmV

Figure 4.2: Representation of FBC correlation group on map (left) with respective FBC impairments (right)
Image courtesy: NimbleThis

As indicated, this same algorithm is easily adapted for RxMER data in DOCSIS 3.1 OFDM channels. As with FBC data, it is useful to present data at a high-level first as a widget, shown in Figure 4.3.

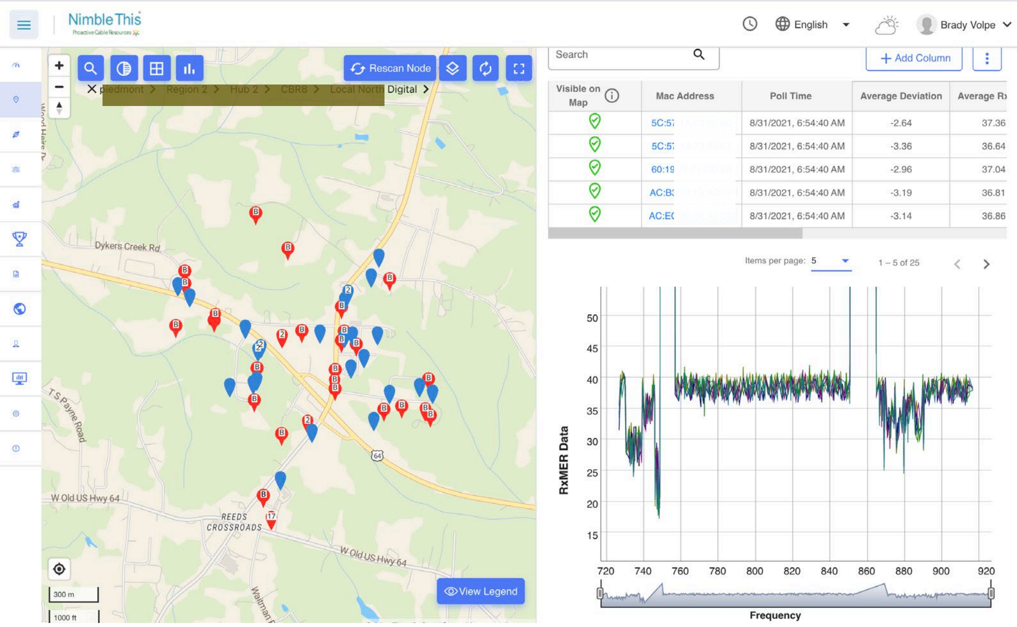


Node	Correlation	Modem Count	Average RxMER
North Digital	CBR8: North Digital Cluster 0 :Average RxMeR 40.78	5	40.78

Figure 4.3: RxMER Correlation Widget Groups by Node, Modem Count, and Average RxMER
(Image Courtesy: NimbleThis)

The operational value of the RxMER correlation widget is that an end user can quickly identify clusters with low RxMER. Clusters with low RxMER will operate at a low OFDM modulation resulting in low or no data speed to subscribers. The low RxMER in a cluster is a result of outside plant impairments, so these can be addressed by outside plant techs.

Figure 4.4 shows the mapping of the clustered data. This view results when clicking on the correlation group in the widget of Figure 4.3.



Visible on Map	Mac Address	Poll Time	Average Deviation	Average Rx
✓	5C:57[redacted]	8/31/2021, 6:54:40 AM	-2.64	37.36
✓	5C:57[redacted]	8/31/2021, 6:54:40 AM	-3.36	36.64
✓	60:19[redacted]	8/31/2021, 6:54:40 AM	-2.96	37.04
✓	AC:B3[redacted]	8/31/2021, 6:54:40 AM	-3.19	36.81
✓	AC:EC[redacted]	8/31/2021, 6:54:40 AM	-3.14	36.86

Items per page: 5 1 – 5 of 25

Figure 4.4: RxMER Clustered data on map (left) with clustered RxMER data (right)
(Image Courtesy: NimbleThis)

Figure 4.4 shows the actual RxMER clustered data on the right-hand side. As can be observed, there are many locations in the RxMER data where the MER drops below 35 dB. An ideal OFDM channel would have its RxMER data above 39 dB across every data point to support 4096-QAM.

On the left-hand side of Figure 4.4, the blue modems indicate which DOCSIS 3.1 modems are part of the cluster group and are impacted by low RxMER. In many networks today there is not 100% penetration of DOCSIS 3.1 modems, so it is quite common to have many DOCSIS 3.0 and DOCSIS 2.0 modems that are around the cluster but are not impacted because they do not use the OFDM channel.

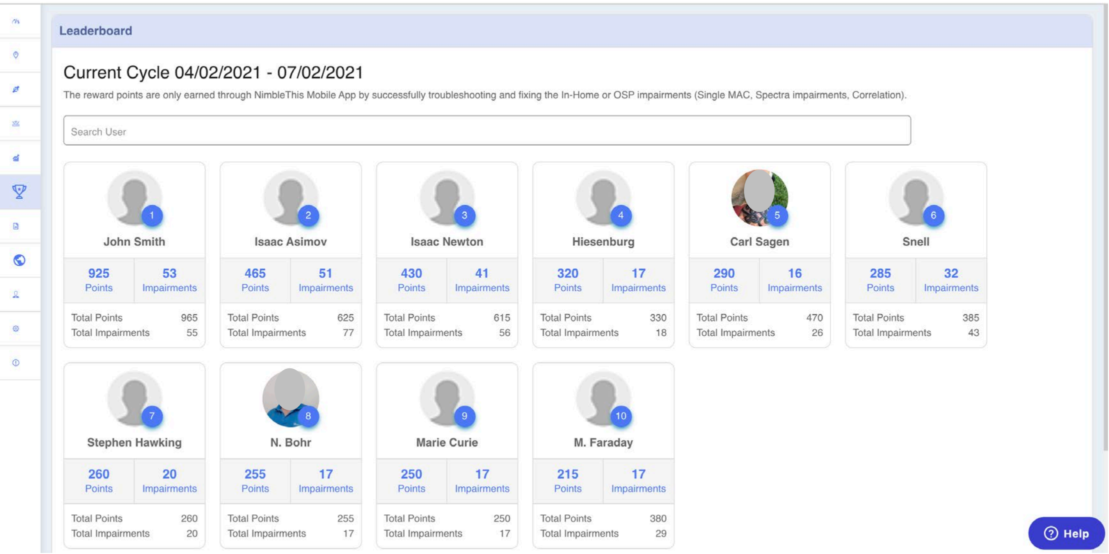
Again, the value of Figure 4.4 is that a technician can quickly observe that the downstream impairments in the OFDM channel are common to every subscriber. Visiting an individual subscriber home will not fix the impairment. This is an outside plant problem which must be addressed in the outside plant.

4.1 Future Research

4.1.1 Supervised Learning

The current implementation, while very powerful, uses unsupervised machine learning. This means the ML engine has no knowledge if the FBC data is impaired or not impaired nor does it know the impairment type once it is classified as impaired. The classification comes from SID data, which is only somewhat accurate. The next level is to achieve supervised learning, which is a machine learning engine whereby the engine already has knowledge about FBC impairment types. This requires a lot of work from end users to label these existing impairments and build a database through which the ML engine can be trained.

There are two approaches in process to do this. First, a built-in gaming feature that encourages its users to label impairments when fixing problems has been applied. Fix a problem and label an impairment and get points. Users with the most points get on the leader board. This is used by cable operators as an incentive program for their technicians. An example of this is shown in Figure 4.5.



Leaderboard

Current Cycle 04/02/2021 - 07/02/2021

The reward points are only earned through NimbleThis Mobile App by successfully troubleshooting and fixing the In-Home or OSP impairments (Single MAC, Spectra impairments, Correlation).

Search User

1 John Smith
925 Points | 53 Impairments
Total Points 965
Total Impairments 55

2 Isaac Asimov
465 Points | 51 Impairments
Total Points 625
Total Impairments 77

3 Isaac Newton
430 Points | 41 Impairments
Total Points 615
Total Impairments 56

4 Hiesenburg
320 Points | 17 Impairments
Total Points 330
Total Impairments 18

5 Carl Sagen
290 Points | 16 Impairments
Total Points 470
Total Impairments 26

6 Snell
285 Points | 32 Impairments
Total Points 385
Total Impairments 43

7 Stephen Hawking
260 Points | 20 Impairments
Total Points 260
Total Impairments 20

8 N. Bohr
255 Points | 17 Impairments
Total Points 255
Total Impairments 17

9 Marie Curie
250 Points | 17 Impairments
Total Points 250
Total Impairments 17

10 M. Faraday
215 Points | 17 Impairments
Total Points 380
Total Impairments 29

Figure 4.5: Leaderboard used for incentivizing users and collecting labeled data
(Image Courtesy: NimbleThis)

Second, CableLabs is working to have several expert users, including the author of this paper, to label a set of FBC data. It is hopeful that either or both two methods will generate a large enough dataset to be used for a true supervised learning model which can further improve upon the existing model.

4.1.2 New Impairment Detector

In the future, one could analyze ways to create a reliable way of identifying and localizing impairments to bypass the need for SID and clustering in the first place. Analysis/clustering on accurate labels of impairments would be able to find the same impairment in multiple modems but do so with greater confidence and accuracy. To be able to identify and localize all these impairments there needs to be large datasets of labeled data which are currently not accessible. Models for prediction could be made for each impairment or one large model could be made to make predictions about all impairments. Possible avenues to investigate include Convolutional Neural Networks (CNNs), perhaps some like ones seen in computer vision such as YOLO (You Only Look Once) to both classify and localize impairments [4, 9].

This starts to move into the arena of artificial intelligence (AI). Which the author of this paper chooses to use with great care. Today machine learning is being used and often times AI is used as a marketing gimmick. However, given enough data, a true AI model can be developed with the help of technicians. Lots of technicians feeding accurate data into PNM applications. Once this level is achieved, ML and/or AI models can be developed which will look at a single or multiple FBC images and not only inform the user of what the impairment is (i.e., suckout or standing wave), but further it can make very accurate suggestions of what the most probably repair for the impairment may, such as "85% probability of a bad drop cable". This technology is not years away, but something we expect to realize within the next 1-2 years and will change the technicians interact with PNM technology.

Abbreviations

CM	Cable Modem
CMTS	Cable Modem Termination System
CNN	Convolutional Neural Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DOCSIS	Data Over Cable Service Interface Specification
EPS	Epsilon parameter in DBSCAN
FBC	Full-Band Capture
FEC	forward error correction
HD	high definition
Hz	hertz
LOF	Local Outlier Factor
ISBE	International Society of Broadband Experts
PNM	Proactive Network Maintenance
RF	Radio Frequency
SCTE	Society of Cable Telecommunications Engineers
SID	Spectral Impairment Detector Released by CableLabs
YOLO	You Only Look Once

Bibliography & References

[1] Breunig, M., Kriegel, H.P., Ng, R., & Sander, J. (2000). LOF: identifying density-based local outliers. In ACM sigmod record (pp. 93–104).

[2] elktros. "Band Stop Filter Circuit Design and Applications." Electronics Hub, 26 Jan. 2019, https://www.electronicshub.org/band-stop-filter/.

[3] Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. KDD (pp. 226–231).

[4] Kiranyaz, Serkan & Avci, Onur & Abdeljaber, Osama & Ince, Turker & Gabbouj, Moncef & Inman, Daniel. (2019). 1D Convolutional Neural Networks and Applications: A Survey.

[5] Lakshmanan, Swetha. "How, When, and Why Should You Normalize / Standardize / Rescale Your Data?" Towards AI — Multidisciplinary Science Journal, 16 May 2019, https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff.

[6] Pandas.Core.Window.Rolling.Rolling.Median — Pandas 1.0.5 Documentation. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.core.window.rolling.Rolling.median.html. Accessed 7 July 2020.

[7] PNM Best Practices Primer: HFC Networks (DOCSIS 3.1). CableLabs, 6 May 2020.

[8] Python Lists vs. Numpy Arrays - What Is the Difference?: IST Advanced Topics Primer. https://webcourses.ucf.edu/courses/1249560/pages/python-lists-vs-numpy-arrays-what-is-the-difference. Accessed 7 July 2020.

[9] Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection (cite arxiv:1506.02640)

[10] "Window Sliding Technique - GeeksforGeeks." GeeksforGeeks, 16 Apr. 2017, https://www.geeksforgeeks.org/window-sliding-technique/.

[11] "Proactive Network Maintenance using Fast, Accurate Anomaly Localization and Classification on 1-D Data Series",(July 2020), Jingjie Zhu, Karthik Sundaresan, Jason Rupe CableLabs, Louisville, CO, U.S.A