LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO -
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Technical Note	LIGO-T2300147–v	2023/08/23

Interim Report 2: LIGO Seismic State Characterization using Machine Learning Techniques

Isaac Kelly

California Institute of Technology
LIGO Project, MS 18-34
Pasadena, CA 91125
Phone (626) 395-2129
Fax (626) 304-9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Project, Room NW22-295
Cambridge, MA 02139
Phone (617) 253-4824
Fax (617) 253-7014
E-mail: info@ligo.mit.edu

LIGO Hanford Observatory
Route 10, Mile Marker 2
Richland, WA 99352
Phone (509) 372-8106
Fax (509) 372-8137
E-mail: info@ligo.caltech.edu

LIGO Livingston Observatory
19100 LIGO Lane
Livingston, LA 70754
Phone (225) 686-3100
Fax (225) 686-7189
E-mail: info@ligo.caltech.edu

http://www.ligo.caltech.edu/

1 Introduction

The Laser Interferometer Gravitational-wave Observatory (LIGO) is a massive laser interferometer constructed specifically to detect gravitational waves. These distortions in spacetime appear as changes in the relative lengths of the interferometer arms, which causes a phase shift in the light reflected by the test masses. The detector signal indicates the current strain on spacetime. Analysis of the strain over time allows extraction of signals from individual events, and these events provide insight on astrophysical and relativistic phenomenae. Current technological limitations constrain the observable emissions to those from inspiraling compact binary objects. Other sources are predicted by models, but no detection has been made yet.

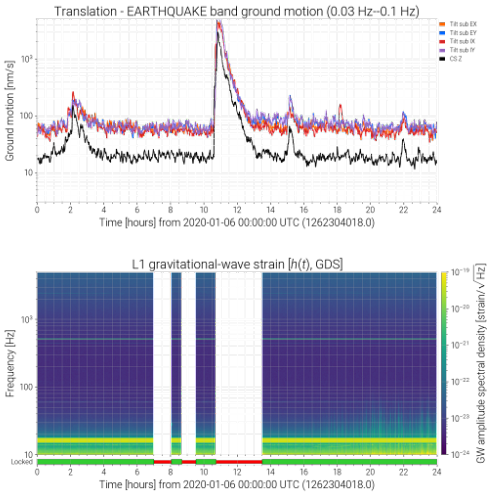

Figure 1: Earthquake event and corresponding data loss

Besides gravitational radiation, terrestrial detectors are subject to a range of other strains, all of which can interfere with the correct operation of the detectors. A significant part of this interference is ground motion; the 4-km interferometer arms are susceptible to distortion caused by movements in the earth beneath them. Even the strongest gravitational waves require a detector sensitivity of approximately $1*10^{-21}/\sqrt{Hz}$ to be evaluated with scientific significance.[1] One common type of ground motion, called the 'secondary microseism', is over 10 orders of magnitude stronger than the real signal at 10 Hz. [1] Earthquakes and human-caused (or anthropogenic) noise also cause distortions or loss of the strain signal. Ground motion is a significant contributor to noise and glitches in the detector (see figure 1) with both active and passive isolation utilized in the system to reduce its effects. [1] However, this isolation cannot completely nullify its effects, and so it is necessary to determine when the detector is being affected by this interference in order to find noise sources, test new isolation methods, and avoid misinterpretation of such noise as a gravitational wave event.

In order to prevent ground motion from propagating to the interferometer system, LIGO includes multiple seismic isolation devices. One of these is the Internal Seismic Isolation (ISI) system built into each optical chamber. One part of the ISI is a seismometer mounted on the ground outside the vacuum chamber in order to provide feed-forward correction. [2] These seismometers (one for each vacuum chamber) also allow the measurement of the respective ground motion at each chamber. Data from these seismometers is recorded and stored as time-series, and is available through the LIGO Network Data Service.

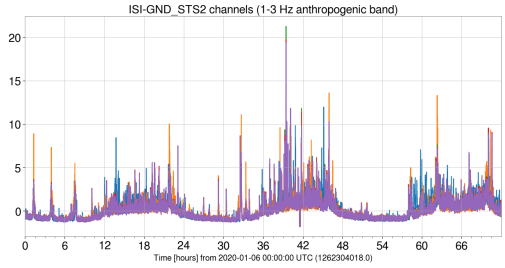

Figure 2: Periodic elevation in noise floor in the anthropogenic band during daytime hours, followed by reduction during nighttime hours.

We use traditional clustering methods to analyze this

data in order to evaluate the seismic state of the detector. Time-series data from multiple seismometers is acquired and divided into segments of equal lengths. A feature extraction process is applied to these segments. Finally, the K-means clustering algorithm is applied to the resulting features, and the clusters compared to known seismic states and changes in glitch rates. This pipeline successfully identifies known seismic states.

2 Objectives

Given environmental data from ISI ground motion sensors located at each vacuum enclosure, we plan to determine the seismic state of the LIGO detector using clustering algorithms. The specific objectives are summarized below.

- **Objective 1: Dataset creation.**

  Time-series ground motion data will be acquired from sensors in multiple locations at LIGO Livingston. Fixed-length segments of time will be extracted, and a feature extraction process will transform each resulting timeseries into scalar features.

- **Objective 2: Clustering and evaluation.** While individual sensors can provide clear information about seismic states in some situations, analysis of the entire corpus of sensor data should improve the reliability of state determination. Given N time segments, K discrete clusters of time segments will be identified with K-means clustering, each describing a unique seismic state.

- **Objective 3: State identification.** Manual labeling will permit correlation of clusters with known detector states; this may allow discovery of new states, and provides room for future exploration. The final goal for this project is the automatic labeling of the detector's ground-motion state as a veto provider and data quality metric.

3 Progress

A revised detection pipeline has been constructed and evaluated. This system refines and extends the pipeline presented in the previous report. The fundamental structure is the same, as illustrated in Figure 3. Ground motion timeseries data is split into segments and run through a feature extraction process; the resulting dataset is clustered using the K-means algorithm. The clusters are evaluated for stability, mathematical goodness, and correlation with periods of increased glitch rate and known seismic states.

Ground motion timeseries data from ISI seismometers is split into segments and run through a simplified **tsfresh** feature extraction, using a minimal parameter set consisting of 9 features. The feature set previously in use, the standard comprehensive list extracted by **tsfresh**, contains over 700 feature-parameter combinations. The dimensionality reduction is significant when moving to a minimal feature set; this increases the range of inter-point distances, improving the ability of the k-means algorithm to distinguish between examples.

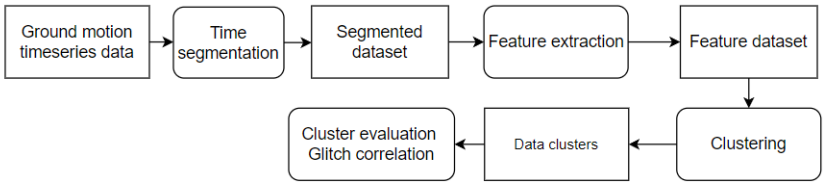

Figure 3: Pipeline structure. Raw ground motion data is processed by segmentation and feature extraction; K-means clustering of the resulting dataset groups time segments by Euclidean distance. The clusters are then evaluated for intrinsic goodness, their correlation with LIGO events of interest, and coherence with known seismic states.

The computational cost of feature extraction is significantly reduced by this smaller parameter set; while a comprehensive feature extraction on a 30-hour segment of data required 7 minutes of compute time, the minimal featureset could be extracted in 10 seconds.

The resulting features are clustered using a k-means [3] algorithm, as implemented in the **scikit** library. [4] The K-means algorithm separates the time segments into k discrete clusters, based on their Euclidean distance from certain centroids; these centroids are initially selected randomly and are then refined as the algorithm proceeds.

The K-means algorithm has already shown promise in identifying seismic states. Bernhardt et al [5] applied K-means to seismometer data, and then to microphones and accelerometers in the physical environment monitoring (PEM) system. Bernhardt's approach used raw timeseries data, taking a 2-hour segment from each point in time, while we are applying feature extraction and separating time segments altogether.

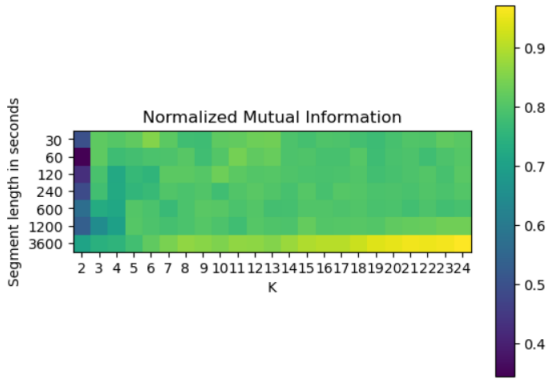

Figure 4: Normalized Mutual Information (NMI) score across 25 runs of single initialization kmeans++. We see a mean of approximately 0.8; this means that around 80% of clusters are identical between runs.

Depending on the clustering parameters, and especially the initial cluster centroids, running the same process on the same data multiple times can yield drastically different results. To address this problem, the parameters of the k-means++ initialization were adjusted to take multiple trials of centroid locations and choose the best. The initial randomness, when combined with sufficient sample size, creates greater determinism in the final result. In order to evaluate the similarity between various results, a few algorithms can be applied. In this

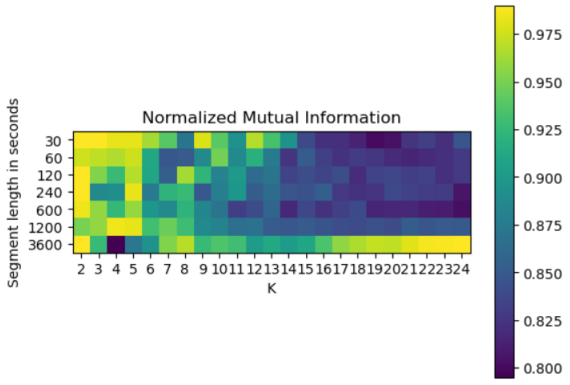

Figure 5: NMI score across 25 runs of 10-initialization kmeans++. All scores are above 0.8, with some as high as 0.98; similarity between multiple clustering runs is thus very high.

work we have utilized normalized mutual information (NMI), adjusted rand index (ARI), and Fowlkes-Mallows score; any one of the three metrics is sufficient to establish the difference in stability between different clustering parameters. In Figures 4 and 5, we compare NMI scores for a varying number of kmeans++ initialization trials. Allowing kmeans++ to use more initialization attempts clearly increases the stability of the results.

DBSCAN was explored for the previous report. As noted previously, results were not informative; we theorize that this dataset is not well suited for the density-type algorithm, as k-means was able to cluster the data effectively.

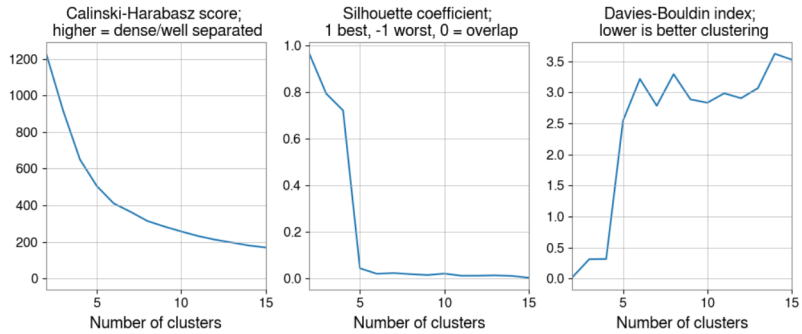

Figure 6: Cluster evaluation metrics for 30 hour dataset split into 240-second segments. Each metric addresses a different statistical characteristic of the clusters; however, they all agree that clustering is much 'worse' for k values of 5 or higher.

Intrinsic cluster evaluation metrics have been applied to our clustering results. The Davies-Bouldin, silhouette, and Calinski-Harabasz algorithms have been tested. These metrics evaluate the properties of the clusters, not their correlation to 'truth'. This is helpful for comparing 'successful' clusterings when the clusters are already clearly correlated to points of interest. However, when that correlation is weaker, the metrics fail to provide much insight. Figure 6 shows an example of these intrinsic scores. These three metrics point to k values of 4 or lower.

4 Algorithm

Intrinsic cluster metrics are helpful as a purely mathematical evaluation of cluster characteristics. However, cluster goodness and cluster usefulness are not intrinsically connected; well-separated clusters do not necessarily provide insight into a glitch in LIGO. So we have designed and implemented a metric, the Glitch Correlation Metric, whereby LIGO glitches can be correlated with clustering results; this allows objective analysis of such results. The GCM is formulated as follows:

A timeseries T is composed of segments T_i of constant length l, each of which corresponds to a segment G_i of timeseries G. Each segment G_i contains a count g_i of glitches in that time period.

A clustering algorithm is applied to T with cluster count parameter K. This creates K discrete clusters C_j, each of which contains a set S_j of segments of T, corresponding to segments of G.

Glitch count G_j for each cluster is defined as $\sum g_i$ for all i in S_j.

The mean glitch rate R_j for each cluster is then $\frac{G_j}{count(S_j)*l}$

The mean glitch rate R_T for timeseries T is similarly $\frac{\sum g_i}{count(T_i)*l}$

Now each cluster C_j is placed in a bin determined by comparison between glitch rates R_j and R_T.

$\sum_{below} = \sum^{R_j > R_T} G_j$

$\sum_{above} = \sum^{R_j < R_T} G_j$

And the final score $M = \frac{\sum_{above}}{\sum_{above} + \sum_{below}}$

This metric evaluates how well the clustering has captured periods of time with higher glitch rates; these time periods are important because one goal of the project, as stated above, is to identify exactly these periods. The score asymptotically ranges from 0 to 1, with higher results indicating a more effective capturing of high glitch rate periods with clusters. We consider every resulting cluster in this calculation, avoiding selection bias in comparisons.

A broad gridsearch has been completed with this metric in which has been used to objectively evaluate clusterings before moving to individual analysis. In Figure 7, we see two clustering with similar parameters but very different GCM results; this disparity was first identified with the illustration of GCM found in Figure 8.

Besides this mathematical work, some effort has been put into visualization of clustering results. The original approach, using another graph on a twinned Y axis, was simple to implement but not very instructive. The gwpy library supports binary flags (as seen in Figure 1), but this is only useful for a result with two clusters. To visualize three or more clusters, a new approach is necessary. Coloring of the seismic timeseries was visually informative but slow to run. So a simple linear bar graph has been implemented. This is accompanied by an illustration of the concurrent glitch rate and as many seismic channels as desired. Additionally, a flag for locklosses has been added, as glitches are not recorded during these

times.

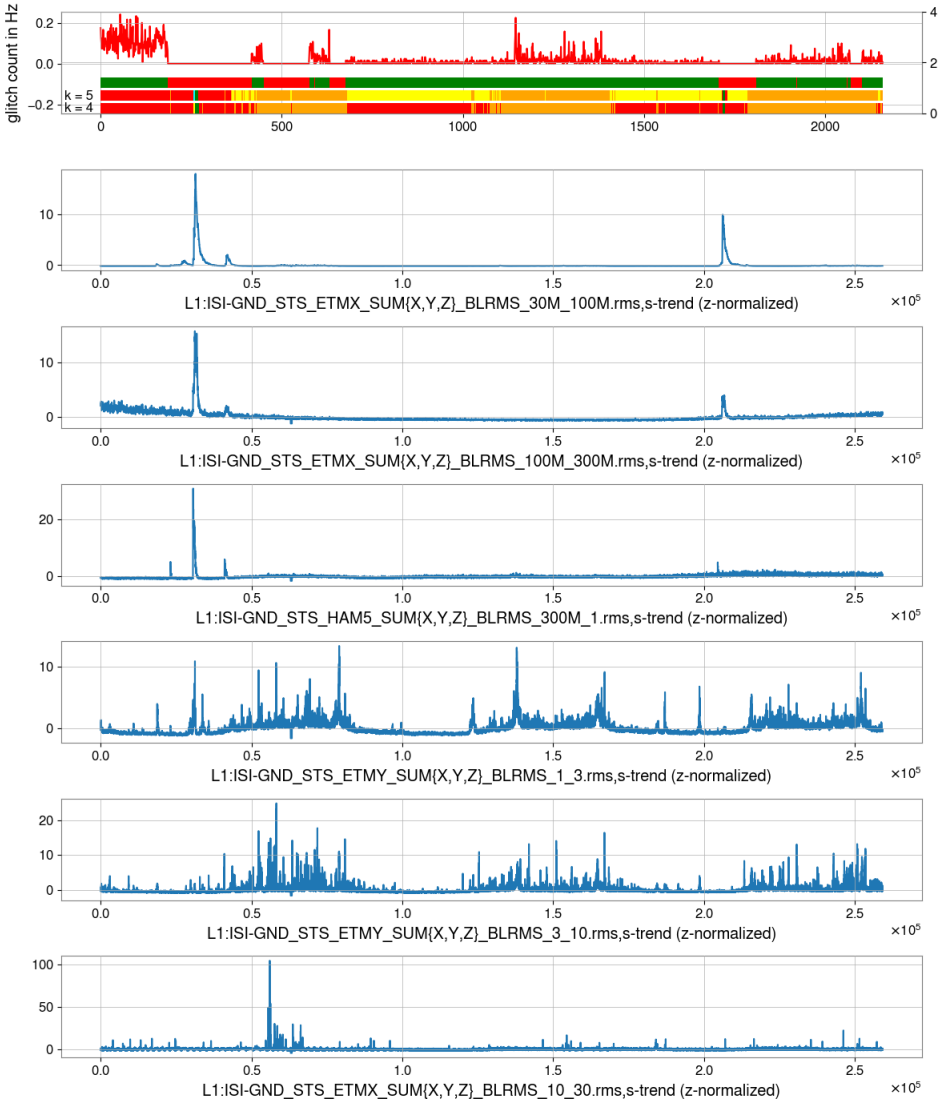

Figure 7: Clustering with 120 second segments; visualization of clusters, glitch rate, and lockloss status in the first graph, with six seismic channels of interest below. Minimal feature extraction. Here we are contrasting two clustering results, with k=4 and k=5. The additional cluster allows the isolation of more interesting seismic activity, and changes the GCM from 0.6 to 0.8.

5 Challenges and Future Work

Points of interest in LIGO strain data and the corresponding auxiliary channels include known noise events, false GW events, and other unexpected behaviors. Analysis of ground motion data from those times should provide some insight into the factors that negatively affect the detector. Comparing glitch rates to seismic channels has already proven useful, as we have shown with our novel metric.

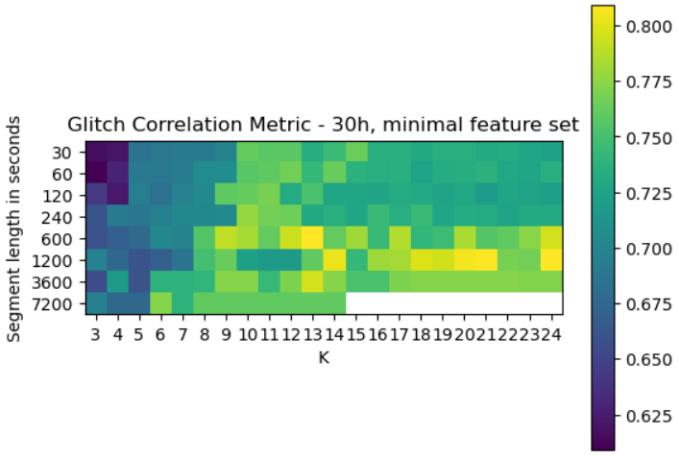

Figure 8: Heatmap of new metric – higher is better. Note the 'sweet spot' between k=6 and k=12, with slight variations based on segment length; additionally, the significant changes as k surpasses 4. These points of interest have given significant insight into clustering characteristics.

One extension of this project would be the use of clustering results as a predictive tool. If a unique state tends to precede locklosses, this could be a trigger for something akin to the current "earthquake mode" used by LIGO sites, which adjusts seismic isolation settings to handle increased ground motion while maintaining lock. Controllers and actuators could be adjusted to address the specific seismic motion that is linked to locklosses.

References

[1] Rana X. Adhikari. Gravitational radiation detection with laser interferometry. *Reviews of Modern Physics*, 86(1):121–151, feb 2014.

[2] LIGO Scientific Collaboration (August 2014 LSC author list). Advanced LIGO. *Classical and Quantum Gravity*, 32(7):074001, mar 2015.

[3] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA, 1967.

[4] scikit-learn. https://scikit-learn.org/stable/. Accessed: 2023-05-17.

[5] Jacob Bernhardt. Data clustering techniques for the correlation of environmental noise to signals in ligo detectors. 2019.