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

Technical Note	**LIGO-T2300147–v**	2023/07/28

Interim Report 1: 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) uses laser interferometers 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.

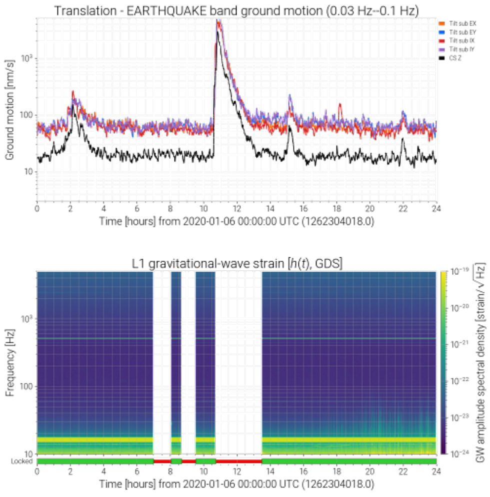

Figure 1: Earthquake event and corresponding data loss

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 monitor external noise sources, LIGO maintains many physical environmental monitors, or PEMs. These include seismometers, accelerometers, and microphones, all of which produce streams of time-series data on disturbances in the LIGO system. We will use traditional clustering methods to evaluate the seismic state of the detector via analysis of PEM sensor data. The time-series data from PEM channels will be divided into time segments. Two methods will be used to create clusters from these segments. In one method, a feature extraction process will create a reduced dataset to which the clustering algorithms will be applied. In the other approach, the algorithms will be applied directly to the raw time-series segments.

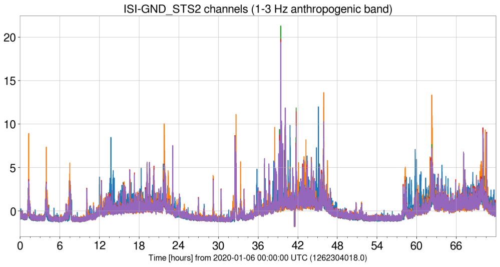

Figure 2: Periodic elevation in noise floor in the anthropogenic band during daytime hours, followed by reduction during nighttime hours.

These complementary methods will be used to implement a pipeline which can in real time determine the seismic state of the detector.

2 Objectives

Given environmental data from PEM sensors, we plan to determine the seismic state of the LIGO detector using clustering algorithms. The specific objectives are summarized below.

- **Objective 1: Dataset creation.**

  From the time-series PEM data, fixed-length segments of time will be extracted, and a feature extraction process will transform these segments into a scalar dataset. Alternatively, a raw time-series dataset may be created.

- **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 will be identified, classifying time periods according to statistical similarities.

- **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

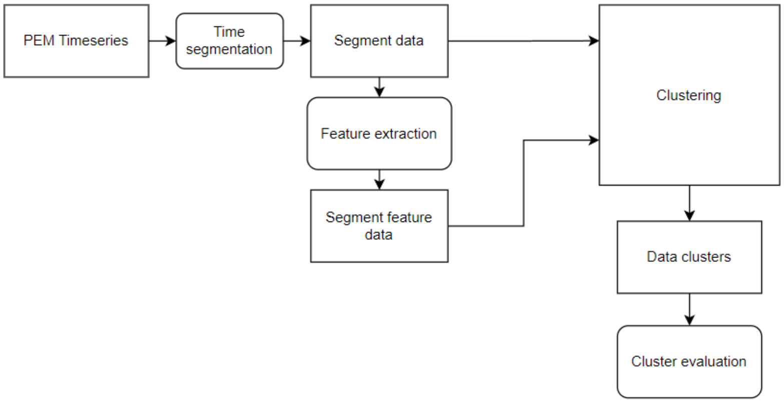

Figure 3: Planned pipeline structure

An initial detection pipeline has been completed. This system implements the central features of the machine learning system, from data acquisition to cluster evaluation. A feature extraction methodology is currently in use, and the DBSCAN and k-means algorithms have been put in place. An initial machine learning pipeline has been completed. It acquires accelerometer data from sensors located in the Internal Seismic Isolation system, filtered with a BL-RMS filter and bandpassed according to apparent project standards. This data is split into segments and run through **tsfresh** feature extraction. The resulting features are clustered using either a k-means [2] or DBSCAN [3] algorithm, as implemented in the **scikit** library. [4] The clusters are then visualized with the **matplotlib** plotting function, originally with a simple line plot and soon with color coding.

A rudimentary time-segment label set is generated, grouping time segments which display similar characteristics (from the point of view of the algorithm). This classification is visually represented as a graph or set of graphs with lines colored according to their cluster identifier. Notable events in the sensor data can then be compared with the clustering result, allowing insight into the characteristics deemed important by the algorithm.

So far, the k-means algorithm shows the most promise in feature clustering. It identifies times of interest across a broad range of segment lengths; so far, 30 second and 300 second segments have both been shown to be effective in visual analysis, with 240 second segments being the most carefully explored. However, some inconsistency has been noted. Because k-means is not deterministic, running the same process on the same data multiple times can yield drastically different results. In current testing this seems to occur most often when higher values of k are used. The algorithm is initialized with the default settings, using the k-means++ initialization; this chooses initial centroids based on a probability distribution in order to speed up convergence. This specific implementation selects from multiple trials for extra precision.

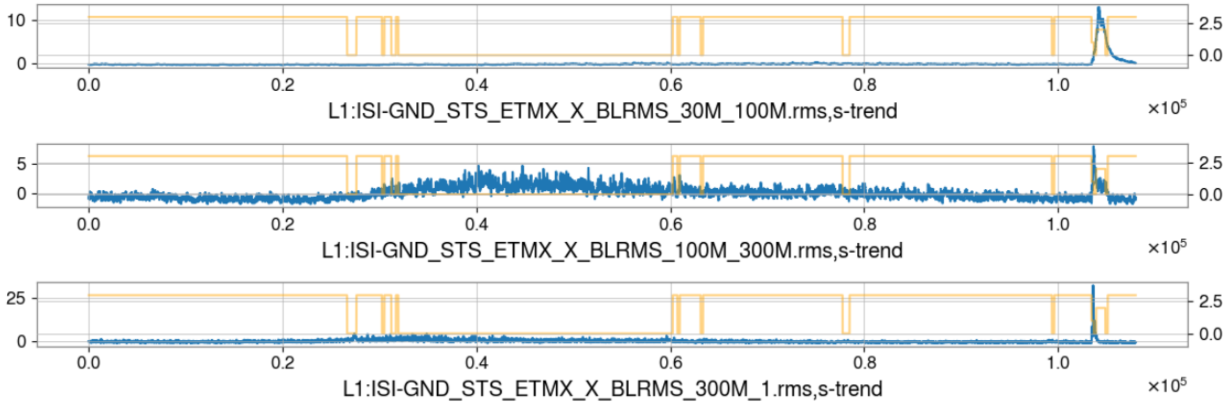

Figure 4: A promising clustering result; yellow lines follow the ID of the segment label

The DBSCAN algorithm has also been explored. Initial results are not promising. A grid search across the available settings (EPS and minimum samples [4]) did not produce useful clusters. The EPS parameter determines the classification of each point based on the number of points within range, while the value of minimum samples determines the points required to form a cluster. EPS is most important for cluster behavior, so the grid search focused on its value. Setting EPS too low causes all points in the dataset to be treated as noise, while setting it too high merges all points into a single cluster. For the 30-hour dataset in use, with 240-second segments, EPS values from approximately 12000 to 13500 were tested; this covered the range from all noise to single cluster. Steps of 100 were tested. The minimum sample parameter was varied from 2 to 5 for these tests. None of the resulting cluster sets correlated with visually obvious points of interest in the sensor data.

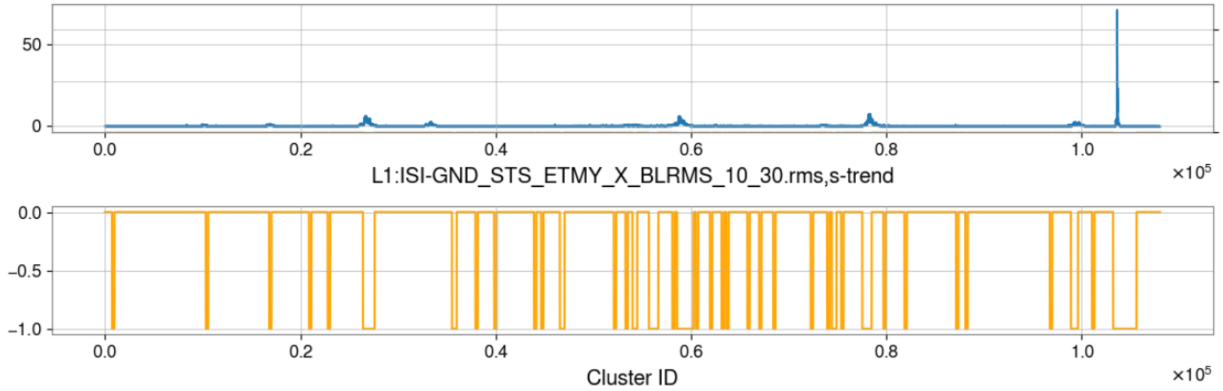

Figure 5: Note the lack of values besides 0 and -1 – we have a single cluster and much noise

Cluster evaluation metrics have been applied to various k-means cluster 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 the tested metrics as applied to the data and algorithms displayed in Figure 4; the silhouette coefficient and D-B index both point towards a cluster number of 4, which does indeed work well for this situation.

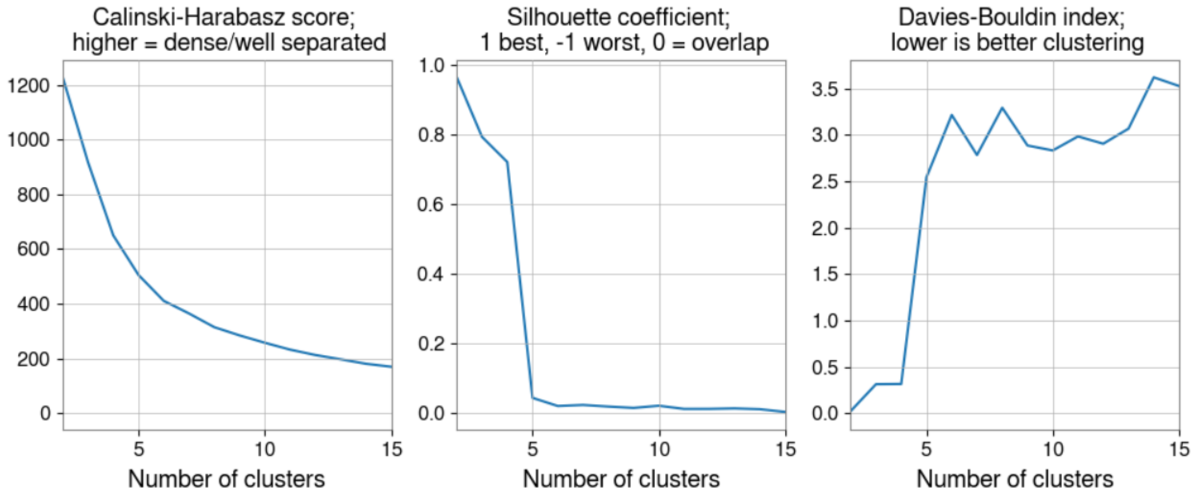

Figure 6: Cluster evaluation metrics for 30 hour dataset split into 240-second chunks

4 Challenges and Future Work

Cluster analysis is difficult because of the aforementioned lack of ground truth. The Davies-Bouldin, silhouette, and Calinski-Harabasz metrics are all usable, as well as the Hopkins statistic. However, these are less helpful when the challenge is creating clusters that correlate with expected behavior. The general trend is towards a worse score with a higher number of clusters (see figure 6), which is expected behavior for a clustering algorithm; however, this only corroborates the results of visual evaluation.

A continued search through clustering parameters and algorithms will include analysis of raw data, different feature extraction parameters, and a larger dataset. Analyzing raw data may require different algorithms and different distance metrics such as dynamic time warping (most helpful when similar events occur over different timescales). **tsfresh** provides a list of default features which can be easily modified; evaluating the importance of each feature to the clustering results will allow the choice of a more focused parameter set, which will improve clustering performance and reduce computation requirements. And more data can be explored, such as LIGO glitch rate information and detector range statistics.

Points of interest in the datastreams include known noise events, false GW events, and other unexpected behaviors. Analysis of PEM and internal seismic data from those times should provide some insight into the factors that negatively affect the detector. And as this will act as a ground truth state, clustering can become more automated; this allows different metrics and different algorithms to be used for supervised learning.

References

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

[2] 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.

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

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