Earthquake Signal Prediction Method Based on CNN and Clustering

Introduction

In recent years, seismic activity has become a matter of widespread concern in earthquake-prone regions. Japan, situated along the Circum-Pacific seismic belt, is one of the most seismically active regions in the world. Due to the convergence of four major tectonic plates—the Pacific Plate, the Philippine Sea Plate, the Eurasian Plate, and the North American Plate—earthquake activity in the country is frequent and intense. This tectonic setting makes Japan highly susceptible to the impacts of powerful earthquakes and tsunamis. These seismic events not only pose a threat to human lives but also result in significant economic impacts. In this context, earthquake preparedness and prediction are of paramount importance for Japan.

With the rise of deep learning technologies, Convolutional Neural Networks (CNNs) have transformed various research fields, from image recognition to natural language processing. In earthquake research, the inherent complex patterns and signals in seismic data make CNNs an ideal choice for analysis. In summary, employing advanced CNN techniques to process Japan's seismic data not only offers a new method for earthquake detection but also brings hope for future earthquake prediction and preventive research.

On the other hand, the volume of earthquake data collected globally is continuously growing, exceeding the existing analytical capabilities. Up until now, such datasets have been analyzed through human expert-intensive, supervised methods. To address these two challenges, we have developed a new unsupervised machine learning framework for detecting and clustering earthquake signals in continuous seismic records.

In this study, we utilize CNNs to analyze Japan's seismic data and to predict earthquake information. The analysis results indicate that, prior to the occurrence of certain earthquake events, there is a significant increase in low-frequency signals. This change in frequency could be a precursor, signaling variations in tectonic movements that may lead to earthquakes. These insights are invaluable and pave the way for more proactive measures to be taken in earthquake-prone regions.

Seismic noise is the result of seismic waves generated by seismic activity propagating through the Earth's interior and surface, exhibiting characteristics and patterns related to seismic activity. Seismic noise is typically classified as natural noise and has specific frequency ranges and spatiotemporal features. Anthropogenic noise refers to interference signals in seismic monitoring systems caused by human activities. In seismic monitoring, anthropogenic noise can disrupt the accurate detection and analysis of seismic events, making its study and control crucial. In addition to anthropogenic noise, there are many other noises from nature that can cause interference with seismic monitoring.

It is an important task to identify and classify seismic signals and background noise in seismology. Traditionally, this task has relied heavily on the expertise of seismologists and manual analysis. However, with the rapid growth of seismic data, manual methods of analysis are becoming increasingly difficult. In addition, the results of manual analysis can be influenced by human bias. Seismograms offer a valuable window into the Earth's internal activities. However, deciphering them can be challenging due to the intertwined signals they contain. Urban settings further complicate this with multiple overlapping disturbances.Traditional methods, which lean heavily on manual interpretations, often fall short in terms of efficiency and accuracy. It is necessary to have a more automated and objective methods.

Unsupervised deep learning, a machine learning method free of labeling, has the power to

learn and discern the innate patterns and structures within data. By using this approach, we

can automatically detect and differentiate seismic signals from background noise, clustering

them together. It has emerged as a transformative tool across various domains, showcasing its potential in analyzing intricate datasets. By categorizing and observing seismic noise, we are able to discern statistical features and variations that can serve as early warning signals and provide predictive information for seismic activity. This capability holds significant potential for the proactive prevention and mitigation of earthquake disasters.

Data

Source and Introduction

The data source is NIED (National Research Institute for Earth Science and Disaster Resilience), which is a leading institution in Japan dedicated to earthquake research and monitoring to enhance the nation's preparedness and response to seismic disasters.

The dataset was collected based on the Hi-Net network and saved as an h5 file.

The following is the structure of one of the h5 files.

```
H5 File
|
|-- attrs
|   |-- 'ele' (Elevation: 0.330)
|   |-- 'lat' (Latitude: 37.050000)
|   |-- 'lon' (Longitude: 142.320833)
|   |-- 'mag' (Magnitude: 7.0)
|   |-- 'time' (Event Time: 2014-07-11T19:22:00.440000Z)
|
|-- 'IWEH' (Dataset for a measurement station)
|   |-- attrs
|   |   |-- 'dist_m' (Distance from the event: 0120189)
|   |   |-- 'ele' (Elevation: -145)
|   |   |-- 'lat' (Latitude: 37.0264)
|   |   |-- 'lon' (Longitude: 140.9702)
|   |-- Data shape: (3, 270000000)
|
|-- 'KI2H' (Dataset for another measurement station)
|
|-- 'MKJH' (Dataset for another measurement station)
|
|-- 'NMEH' (Dataset for another measurement station)
```

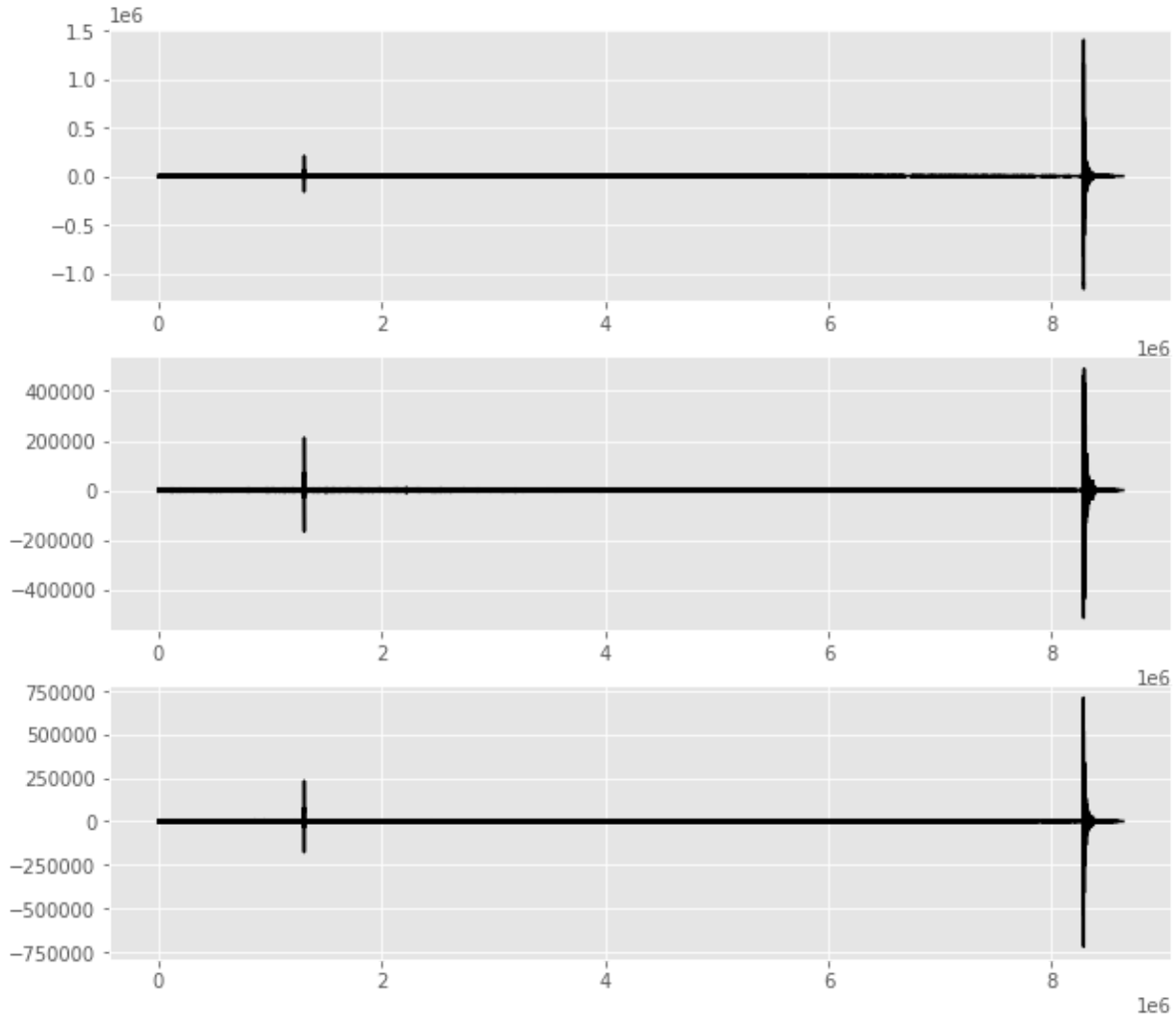
Each h5 file is an event (an earthquake) data, first attrs records some basic information about this earthquake: Elevation, Latitude Longitude, Magnitude, time. This earthquake event occurred at a latitude of 37.05 degrees, a longitude of 142.320833 degrees, and an elevation of 0.330 meters above sea level. 37.05 degrees, a longitude of 142.320833 degrees, and an elevation of 0.330 meters above sea level. It had a magnitude of 7.0, making it a significant and potentially damaging seismic event. The earthquake took place on July 11, 2014, at 19:22:00.440 UTC.

We then have seismic signal data from multiple stations, 'IWEH', 'KI2H', 'MKJH', 'NMEH'. Each site has 3 channels of data (E,N,Z). The signal on each channel is a time series of length 270000000. The site receivers are sampled at 100hz, so the signal data is recorded for a total of 31.25 days before and after the earthquake.

Now we have multiple h5 files, i.e., data from multiple earthquakes, and observations from multiple sites for each record. We hope to make use of these multiple earthquakes and multiple sites to analyse the data, and make some explorations in earthquake prediction.

Visualisation

Since we are trying to explore the features near the time node of earthquake occurrence, we first take one of the nodes as an example, and draw the signals of the 3 channels 24h before the earthquake occurs. We can find that the signals on different channels (i.e., different directions) differ in intensity and peak location.



Models and methods

Traditional matching strategy

In seismology, researchers need to rapidly and effectively identify and locate earthquake events within massive volumes of seismic data, which presents a significant challenge. One strategy to address this issue is the template matching approach.

The fundamental principle of this strategy is that earthquake events exhibit certain similarities in their waveforms, especially those occurring within the same seismic zone. Researchers can capitalize on these similarities by selecting a representative earthquake waveform as a template. This template is then compared with continuous seismic data to identify earthquake events that resemble the template.

Method:

- First, one or multiple templates are chosen. These templates are typically waveforms of known earthquake events.
- Then, these templates are cross-correlated with the continuous seismic data.
- If a segment of the continuous data has a similarity with the template that exceeds a predefined threshold, it is considered to be an earthquake event similar to the template.

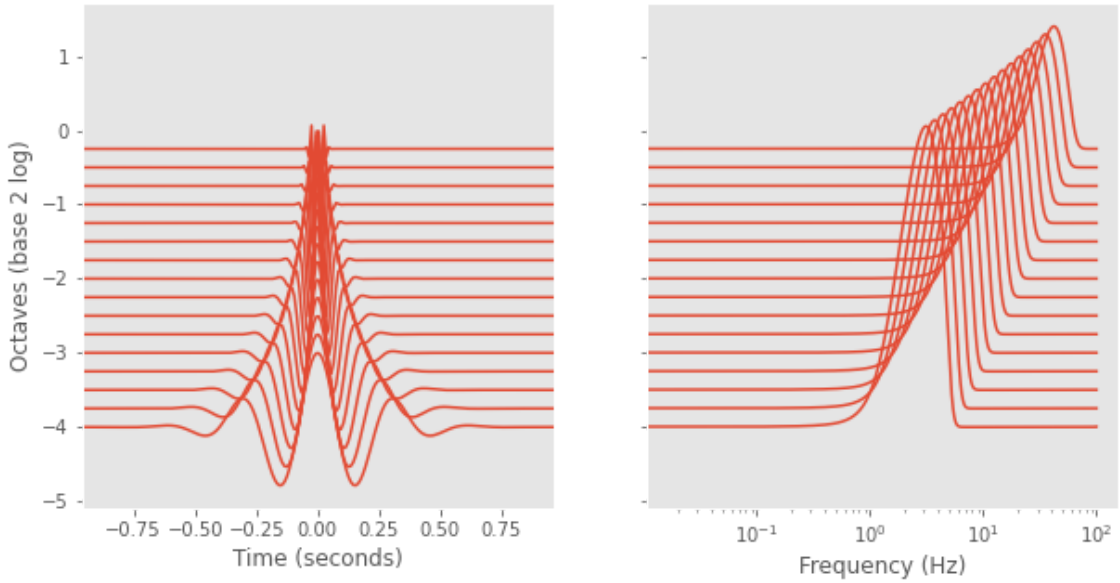
Drawbacks:

1. Dependence on Template Quality: The template matching strategy is highly reliant on the quality of the templates, which includes aspects such as duration and frequency band of the template. Inappropriate selection of templates may lead to inaccurate detection results.
2. High Computational Load: Template matching requires cross-correlation calculations for every potential earthquake event. If the data volume is large, the computational load can be extremely high, which may pose difficulties in practical applications.

CNN-Based Clustering

The process can be broadly summarized as follows:

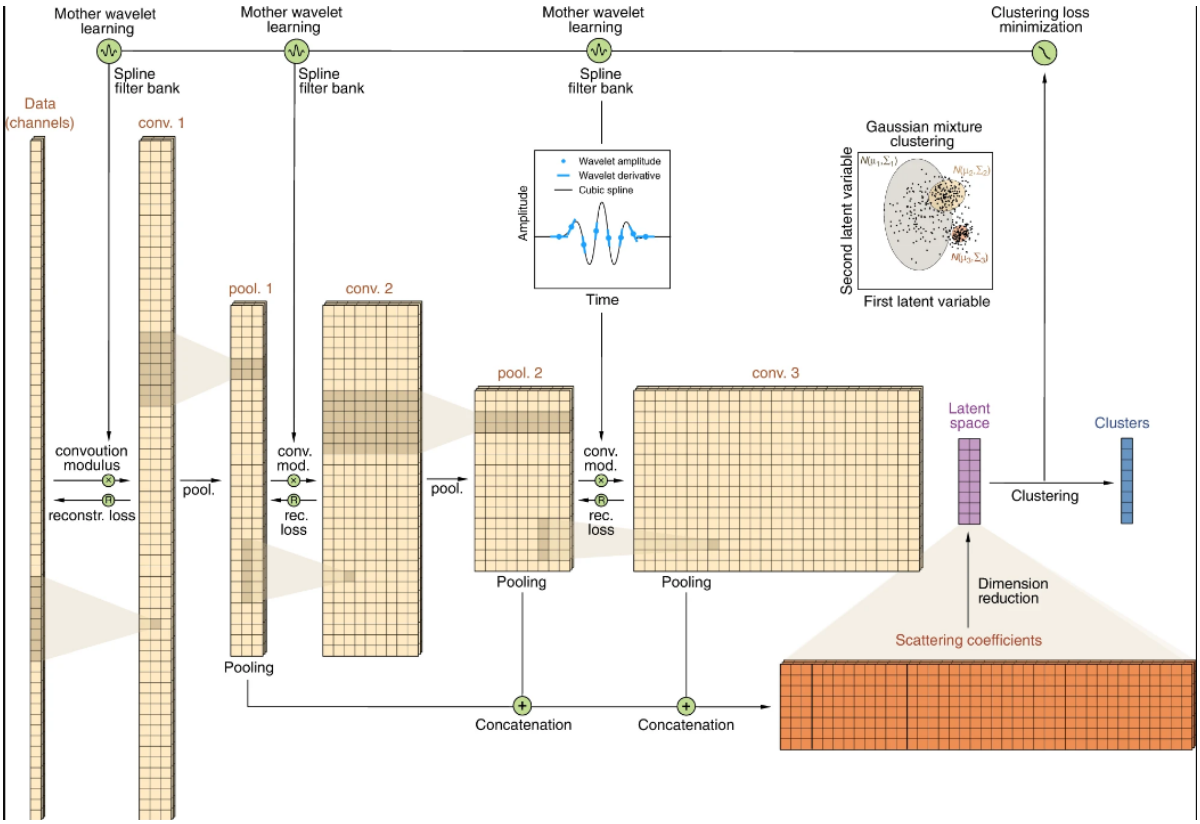
- First, we compute the deep scattering spectrum of continuous three-component seismic graphs using a deep scattering network. The deep scattering network is a deep convolutional neural network, in which the convolutional filters are constrained to be wavelets and the activation is modulus operation. A network permits the establishment of outputs at each layer.



- We apply a sliding window on the single-component seismic graphs and compute the first-order spectrogram using a wavelet transform. A second wavelet transform is applied to the first-order spectrogram, forming a second-order spectrogram. A pooling operation collapses the time axis of the spectrogram, recovering the first and second order scattering coefficients. For each component of the ground motion recording, we compute the scattering coefficients and concatenate them. We repeat this process for each window and retrieve the deep scattering spectrum.
- The deep scattering spectrum is a redundant high-dimensional representation and, due to the curse of dimensionality, is not directly suitable for clustering. Therefore, we extract the most relevant features or characteristics and reduce dimensionality through Independent Component Analysis (ICA). The number of most relevant features (or independent components) is often unknown and should be inferred; initially, we have chosen 10 features, with adjustments made subsequently based on actual circumstances.
- Finally, we perform hierarchical clustering in the low-dimensional feature space constructed by the independent components. The aim of clustering is to group objects—defined here as data points in a given feature space—based on measures of similarity or dissimilarity. K-means clustering can be employed. However, it requires manual setting of the number of clusters (we initially chose 10 clusters).

This process outlines a rigorous, systematic approach to analyzing seismic data through the application of deep convolutional neural networks, specifically tailored wavelet transforms, and advanced dimensionality reduction techniques, ultimately aiming for precise and interpretable clustering of seismic events based on their inherent characteristics in the data.

The figure below is a visual illustration of the above processes



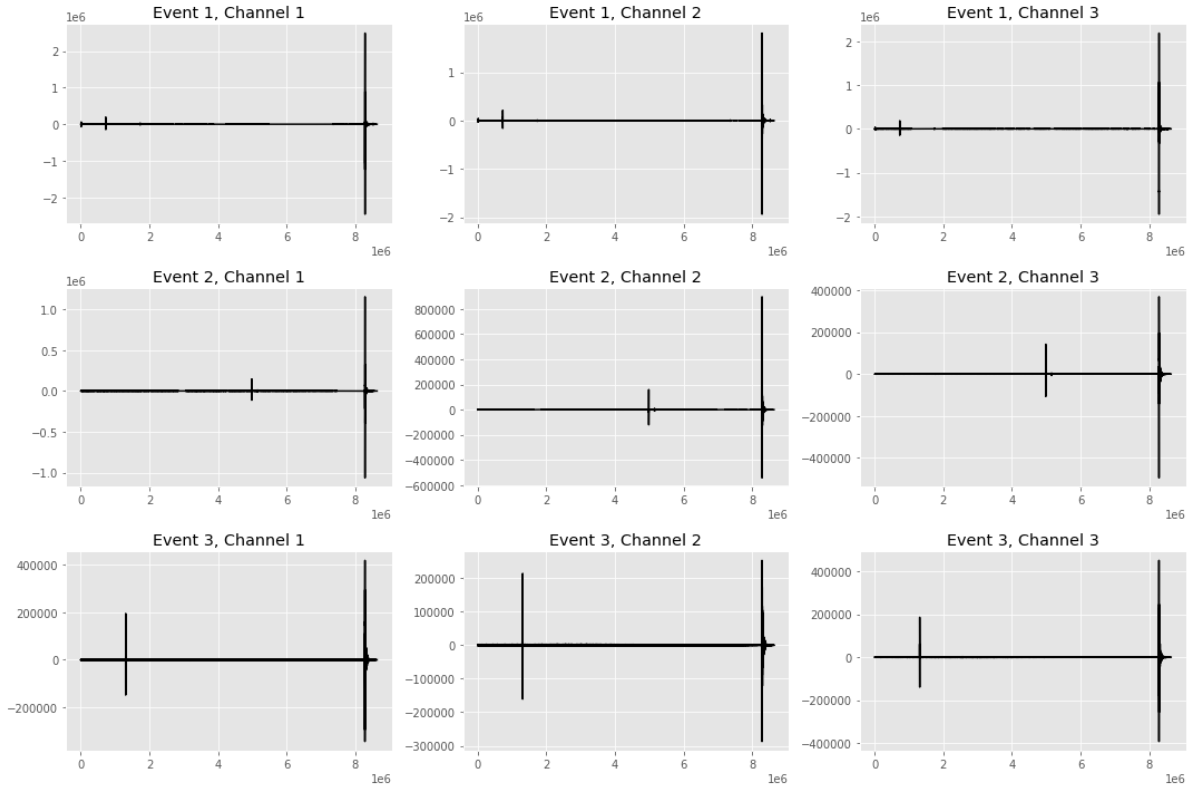
Results

We selected the data of the site 'NMEH' for analysis. There are 3 earthquake records, which occurred in:

- 2016-11-21T20:59:46.890000Z
- 2021-03-20T09:09:44.830000Z
- 2014-07-11T19:22:00.440000Z.

Basic Visualization

First, we visualize the seismic signals from the three channels for the three earthquakes (capturing the 24 hours preceding each earthquake event).



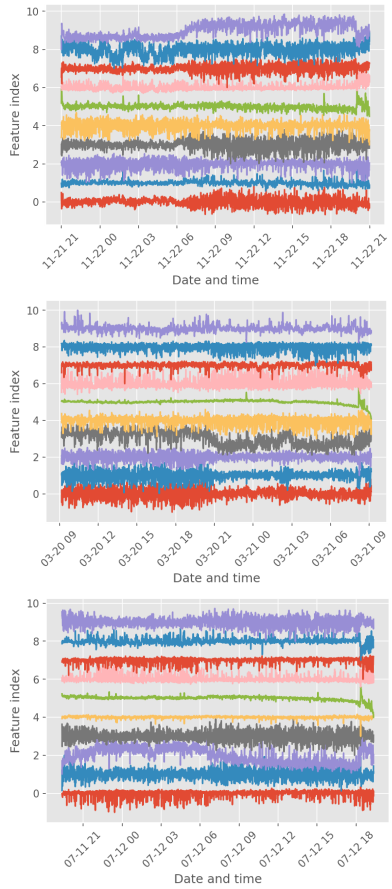
We find that:

- The overall properties of the signals from each channel are quite similar: they reach a peak when the earthquake occurs, and there is a smaller peak before the earthquake. However, these smaller peaks occur at different times, and their relative strengths vary between different earthquakes.
- The signal strength varies among different earthquakes. Even for the same earthquake, the signal strength can differ across various channels (directions). This adds complexity to subsequent analyses that combine data from different earthquakes.

Single Earthquake Analysis

Feature Extraction

Based on the signal visualization above, there is a significant difference in signal strength among different earthquakes. We will first conduct the aforementioned analysis process for each earthquake signal separately. Below are three subplots, each representing the data analysis results for one earthquake. We cluster the earthquake signals into 10 categories, and each curve illustrates how signals from different categories change over time.



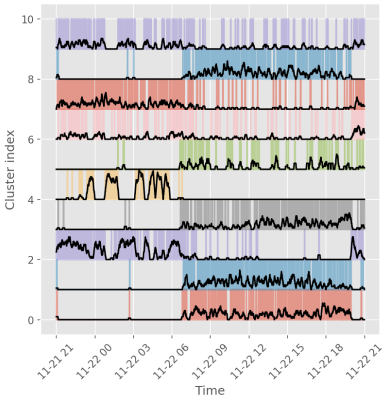
We can observe the following:

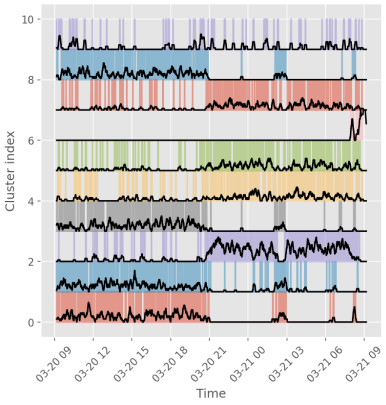
- In all three subplots, many feature signals exhibit abrupt changes, either increasing or decreasing dramatically, at the moments when earthquakes occur. This is good news for us, as our goal is to extract features from the seismic signals that can be used for earthquake prediction.
- Specifically, we observe that the fifth feature signal (corresponding to the light green curve) demonstrates a similar trend during each earthquake: it remains smooth and stable before the arrival of the earthquake, surges instantaneously at the moment the earthquake strikes, then gradually recedes and eventually falls below the intensity level prior to the earthquake. The nature of the light green curves in all three subplots is also quite similar. Therefore, we plan to further explore the signal characteristics corresponding to the light green curve to uncover more insightful information.

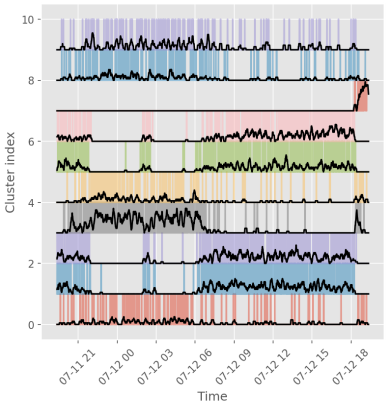
Clustering

Of course, the aforementioned analysis only extracted 10 features and examined their variations over time. Next, we aim to integrate multiple features examined earlier, with the goal of extracting features with stronger correlations, that is, clustering the previously extracted features (which is essentially similar to a principal component decomposition transformation).

In the figure below, we apply K-means clustering to the initial 10 feature categories, resulting in 10 clusters. Following the same analytical approach as before, we plot the temporal trends of the signals within each cluster.



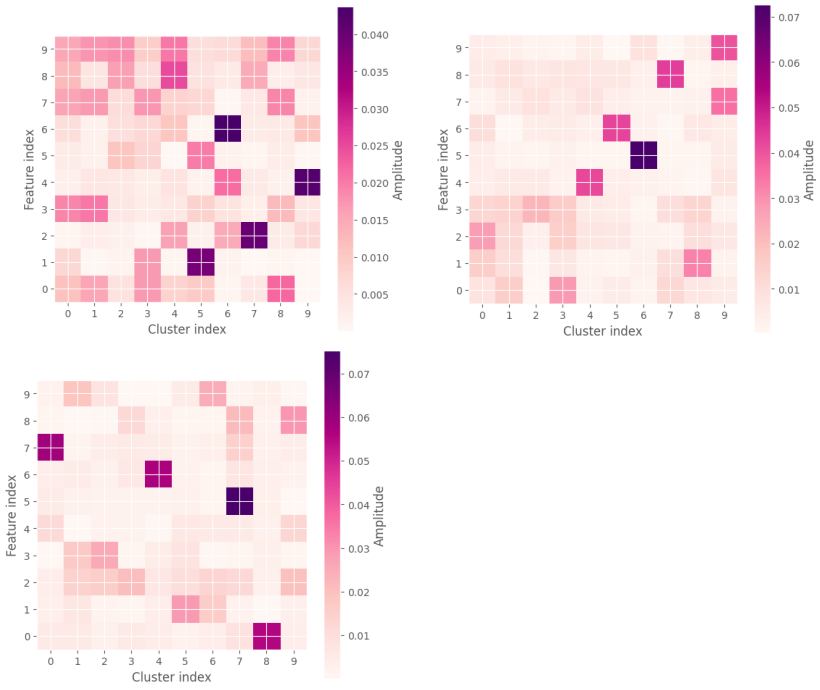




We can observe that:

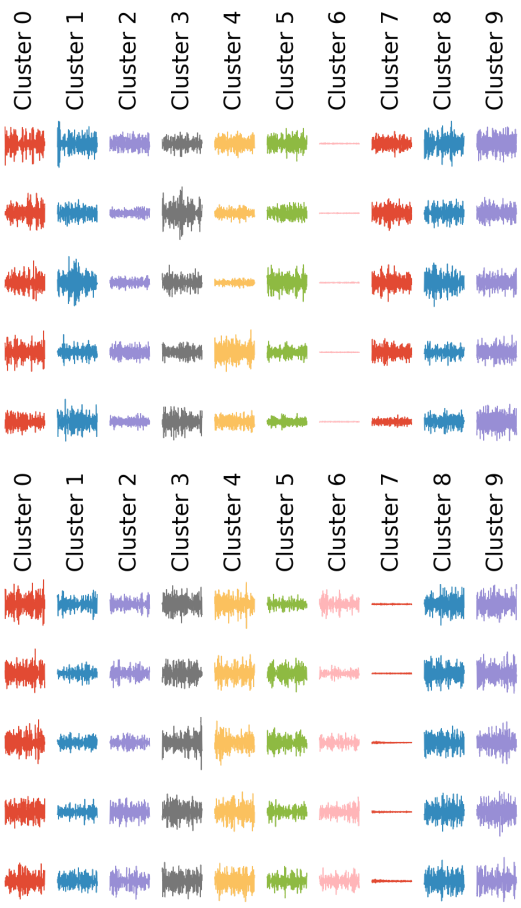
- Signals of many categories exhibit dramatic changes at the moment when earthquakes occur. Particularly noticeable are the signals of category 6 in event 2 and category 7 in event 3. Prior to the occurrence of the earthquakes, these two categories of signals are virtually non-existent, but they spike dramatically at the moment the earthquakes strike. Therefore, we have reason to suspect that these types of signals have a strong relationship with earthquakes, and this relationship is not coincidental, as similar observations have been made across multiple earthquakes.

To gain a clearer understanding of the relationship between the aforementioned features and categories, we have produced correlation graphs for each feature with each category (the three subplots represent the results for the three earthquakes, with darker colors indicating stronger correlations).

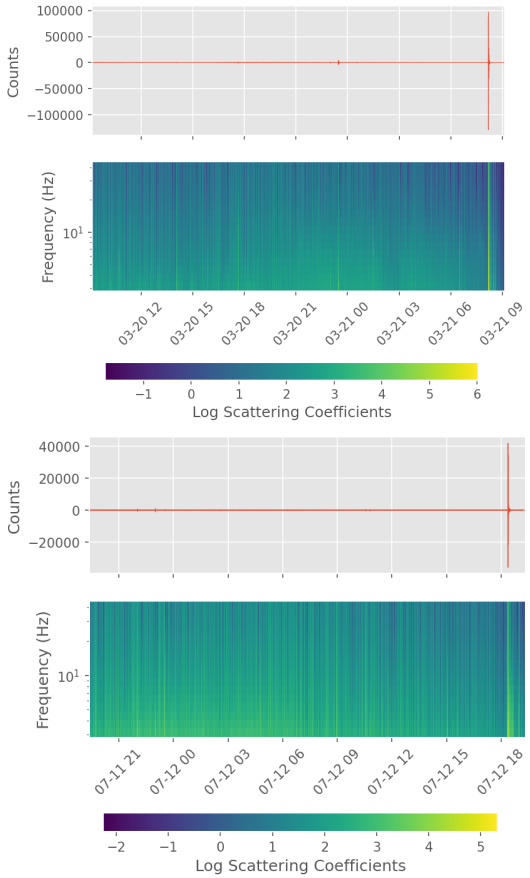


As we have already observed, the fifth feature may have a significant correlation with earthquake signals. In event 2, cluster 6 has the strongest correlation with feature 5, and the signals corresponding to cluster 6 also begin to surge dramatically at the moment the earthquake strikes. The same analysis also applies to cluster 7 in event 3.

Next, we wish to explore what the signals corresponding to cluster 6 in event 2 and cluster 7 in event 3 actually look like. From the K-means clustering, we selected the five most representative "points" (representing the scattering coefficients obtained after five 1-minute-long signals have passed through the scattering network). We converted these "points" back into seismic signal waveforms, yielding the following results (the two subplots represent events 2 and 3).



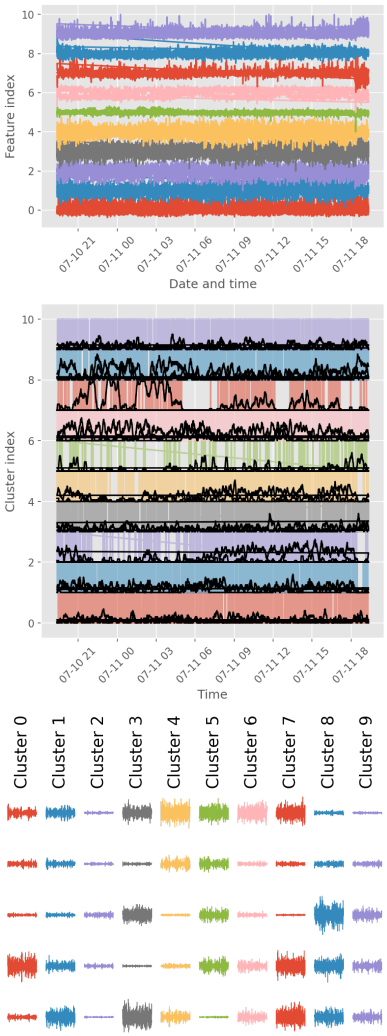
Here, we observe that the waveforms of the signals corresponding to cluster 6 in event 2 and cluster 7 in event 3 are both very stable, small signals. Thus, we speculate that when an earthquake is imminent, the frequency of these small signals will dramatically increase. This can be reflected through the spectral change diagram. We need to focus on the lower part of the two subplots below, which reflects the frequency distribution of the signals at each time point. The darker the color, the stronger the signal at that frequency. We can see that the spectrogram becomes noticeably lighter when an earthquake is about to occur, indicating an increase in small signals, which is consistent with our previous analysis.



From the analysis above, we find that the frequency of small signals increases dramatically when an earthquake is imminent. Therefore, by monitoring the changing trend of these small signals, we may be able to make predictions about impending earthquakes.

Integrated Analysis of Multiple Earthquakes:

Previously, we analyzed the signals for each earthquake separately. In this section, we analyze the signals of multiple earthquakes together. As mentioned earlier, the signal strength of different earthquakes varies, and even the signal strength of the same earthquake can differ in different channels (directions). Analyzing all earthquake data together might not be entirely rational, but we consider this analysis as a complement to the previous ones.



We find that when we attempt to train and test with all the event data from the same station combined, the patterns are not as obvious as before, and the results can be quite messy. The information in the above figure is, in fact, quite limited; each simulated signal is relatively uniform, and the changes around the time of the earthquake are not very clear.

We speculate that different events have varying earthquake intensities and signal strengths. For example, a signal of the same strength could be a strong signal in event 1 and a weak signal in event 2. If we use the same time window and combine all event data for training and testing, the differences in signal strength within the same event can be easily obscured.

Conclusion

In the multiple earthquake signals from the same station, we discovered that certain types of signals increase dramatically before an earthquake, are converted into waveforms, and are found to be small and smooth waveforms. Importantly, we observed similar phenomena in the analysis of numerous earthquake signals. That is, when we analyze events 1, 2, and 3 from the same station, we observe that these small and smooth signals increase dramatically before an earthquake. We believe that these small signals may provide some guidance for earthquake prediction, which will be convenient for subsequent analyses.

Reference

1. Seydoux, L., Balestriero, R., Poli, P. et al. *Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning.* Nat Commun 11, 3972 (2020). **https://doi.org/10.1038/s41467-020-17841-x**
2. Barkaoui, S., Lognonné, P., Kawamura, T., Stutzmann, É., Seydoux, L., de Hoop, M. V., … & Banerdt, W. B. (2021). *Anatomy of continuous Mars SEIS and pressure data from unsupervised learning.* Bulletin of the Seismological Society of America, 111(6), 2964-2981. **https://doi.org/10.1785/0120210095**
3. Steinmann, R., Seydoux, L., Beaucé, E., & Campillo, M. (2022). *Hierarchical exploration of continuous seismograms with unsupervised learning.* Journal of Geophysical Research: Solid Earth, 127(1), e2021JB022455. **https://doi.org/10.1029/2021JB022455**
4. Steinmann, R., Seydoux, L., & Campillo, M. (2022). *AI-Based Unmixing of Medium and Source Signatures From Seismograms: Ground Freezing Patterns* Geophysical Research Letters, 49(15), e2022GL098854. **https://doi.org/10.1029/2022GL098854**