Classification of Seismic Events Using Unsupervised Machine Learning

By Jeff Church (churchjm@umich.edu), Dongdong Yao (dongdony@umich.edu), and Yihe Huang (yiheh@umich.edu)

Note to instructors: Dr. Yao and Dr. Huang are faculty from the U of M Department of Earth and Environmental Sciences. They are not students enrolled in SIADS 697. Please also note that the format of my paper is not a blog post, but a draft for a scientific journal article.

Introduction

Isolating and classifying seismic events recorded in continuous seismograms is a fundamental step of many research problems in the field of seismology. Because the volume of data recorded by modern seismograph networks is too large for human experts to manually process, automated tools are needed to accomplish this task. A technique called template matching, which scans seismograms for regions of high correlation with known waveforms, is a popular choice among seismologists. While a powerful tool, template matching is limited by its reliance on events that have already been discovered.

We propose applying unsupervised machine learning techniques to build an event detection pipeline that does not rely on any a priori knowledge of the events contained in the seismogram of interest. We isolate events using the PhaseNet automated phase picking tool [1], and classify the events using both traditional machine learning and deep learning approaches. We test our pipeline on seismograms recorded in eastern Ohio. This region is interesting because it has seen a marked increase in seismicity since 2010, which is associated with increased hydraulic fracking (HF) activity in the same period of time [2]. Although not required for our pipeline to work, choosing a well-understood region allows us to validate our results. We expect to find two main clusters in this area: HF-induced earthquakes and mining blasts.

We found several similar works by other researchers regarding the classification of seismic events using unsupervised machine learning. Local and teleseismic earthquakes are classified in [3], five classes of seismic noise are clustered and labeled in [4], and precursory seismicity leading up to a landslide in Nuugaatsiaq, Greenland is discovered in [5]. However, to our knowledge we are the first to attempt unsupervised classification of HF-induced earthquakes and mining blasts in continuous seismograms.

Data

Our project analyzes vertical component waveforms recorded by the TA:O53A seismograph station located in New Philadelphia, OH. This location was chosen because of its proximity to

frequent HF and mining activity. Seismograms recorded by TA:O53A should contain a rich collection of waveforms from various sources, including fracking, mining, and tectonic earthquakes. Because tectonic earthquakes are rare in the area surrounding TA:O53A, all earthquakes recorded by this station are assumed to be HF-induced for our purposes.

We analyze two datasets in this project. The first is a labeled dataset consisting of 1,251 HF-induced earthquakes, and 5,152 mining blasts recorded between 2013 and 2018. The second dataset consists of 7,398 unlabeled events automatically picked by the PhaseNet tool with a minimum probability of 0.8 [1]. The PhaseNet dataset spans 2013-2016. See figure 1 for examples of an HF-induced earthquake and a mining blast.

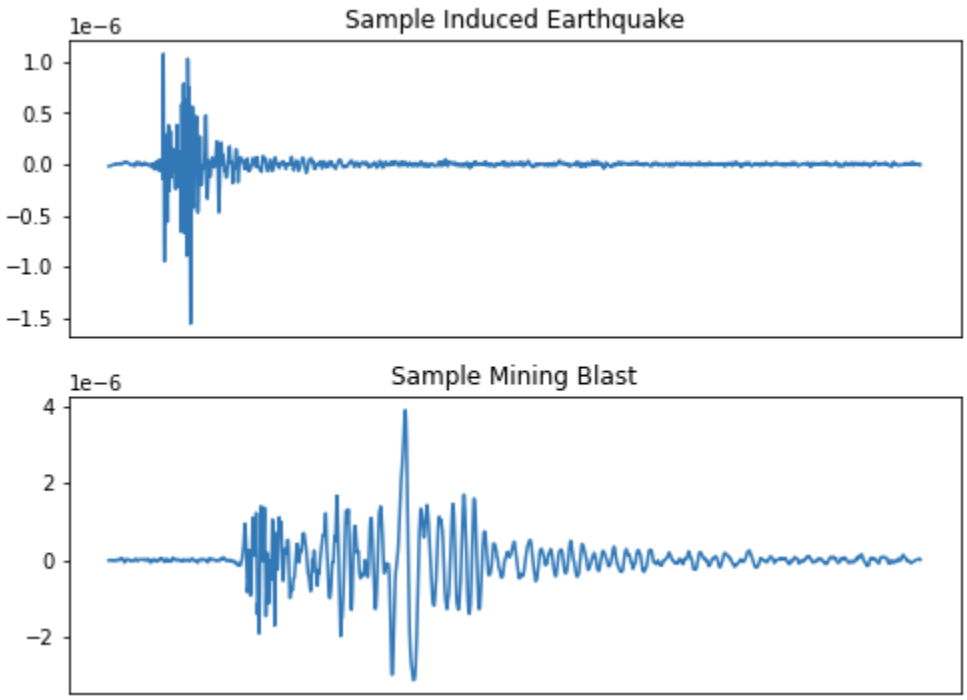

Figure 1. A sample HF-induced earthquake and mining blast from the labeled dataset.

Data Preparation

All data is downloaded and prepared using the obspy Python package. When a waveform is downloaded using obspy the seismometer's instrument response must be removed to convert the units from digital counts to velocity. To prevent response removal from introducing signal distortion, a bandpass pre-filter is applied. Following response removal a second 1-20 Hz bandpass filter is applied, as all seismic activity should be captured in this frequency range.

Figure 2 demonstrates the effects of this waveform preparation. The first row of the figure shows the raw, unprocessed data. The second row illustrates both low-frequency distortion (left) and high-frequency distortion (right) introduced by response removal absent a pre-filter.

The 1-20 Hz bandpass filter addresses this distortion to a degree, but high-frequency noise remains. Only when both filters are applied do we get the clean waveforms in the last row.

Following response removal and filtering, each waveform is normalized by its maximum value. This normalization is critical as waveform amplitudes in our data vary by several orders of magnitude. See data preparation notebook in supplemental materials for complete details.

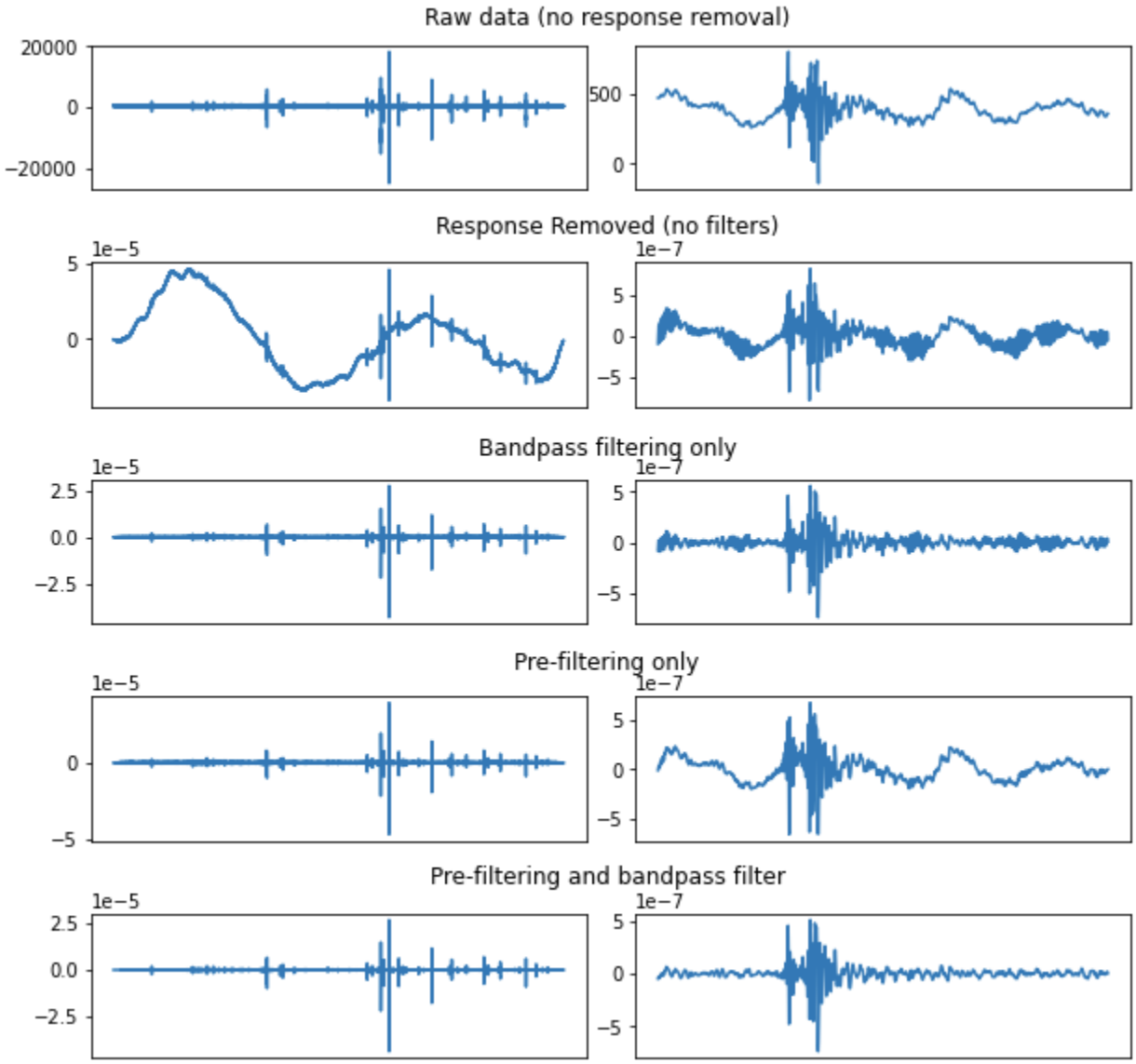

Figure 2. The effects of pre-filtering and bandpass filtering waveforms retrieved with obspy. The left column displays a full day of data to illustrate low-frequency distortions, and the right column displays 15 seconds of data to illustrate high-frequency distortions.

Machine Learning Analyses

The goal of our machine learning analyses is to separate (e.g. classify) HF-induced earthquakes and mining blasts in the PhaseNet dataset using unsupervised techniques. Modeling choices are validated on the labeled dataset. We investigate both a traditional machine learning approach utilizing k-means clustering and manually engineered features, and a deep learning approach utilizing an architecture called Deep Embedded Clustering (DEC) [6]. The advantages and disadvantages of each approach are explored in the Discussion section.

Traditional Machine Learning

We perform a traditional machine learning analysis using k-means clustering and manually engineered features by domain experts. Selected features are listed in table 1 and are taken from [4, 8]. These features are computed using the normalized waveform time-series downloaded with obspy.

Feature	**Description**
Integral of squared waveform	$\int x(t)^2 dt$
Top-three max spectral amplitudes	Computed using Fast Fourier Transform
Frequencies of top-three max spectral amplitudes	Computed using Fast Fourier Transform
Center frequency	$(\sum f_i * F_{xi}) / \sum F_{xi}$ where f_i = frequency, F_{xi} = spectral amplitude
Bandwidth	$\sqrt{(\sum f_i - f_{center})^2 / \sum F_{xi}}$ where f_{center} = center frequency
Zero crossing rate	Number of zero crossings divided by waveform length
Skewness	Skewness
Kurtosis	Kurtosis
Quarter comparison	Median absolute amplitude in last quarter of waveform divided by median absolute amplitude in first quarter
Maximum step	Largest difference between two consecutive samples
Pre-signal noise level	Standard deviation of amplitude distribution in first 4.5s of waveform (e.g. prior to event onset)
Extreme value	Max absolute deviation from mean, divided by variance
Peak absolute magnitude	Max absolute magnitude in waveform

Table 1. Feature set used in k-means clustering analysis.

Traditional Machine Learning - Labeled Dataset

Our choices of clustering algorithm and feature set are first validated on the labeled dataset, and then applied to the PhaseNet dataset where we lack labels to easily assess performance. To mitigate the curse of dimensionality, the feature space is reduced from 17 to three dimensions using Principal Component Analysis (PCA). These three components explain 80.4% of the original variance. See figure A1 in the appendix for PCA loadings.

No analysis is necessary to find the optimal number of clusters in the labeled dataset, as we already know the data contains two types of events; earthquakes and blasts. Training a two-cluster k-means model on the labeled dataset results in exceptional accuracy, assigning only eight waveforms to the incorrect cluster. Clustering results are shown in figure 3.

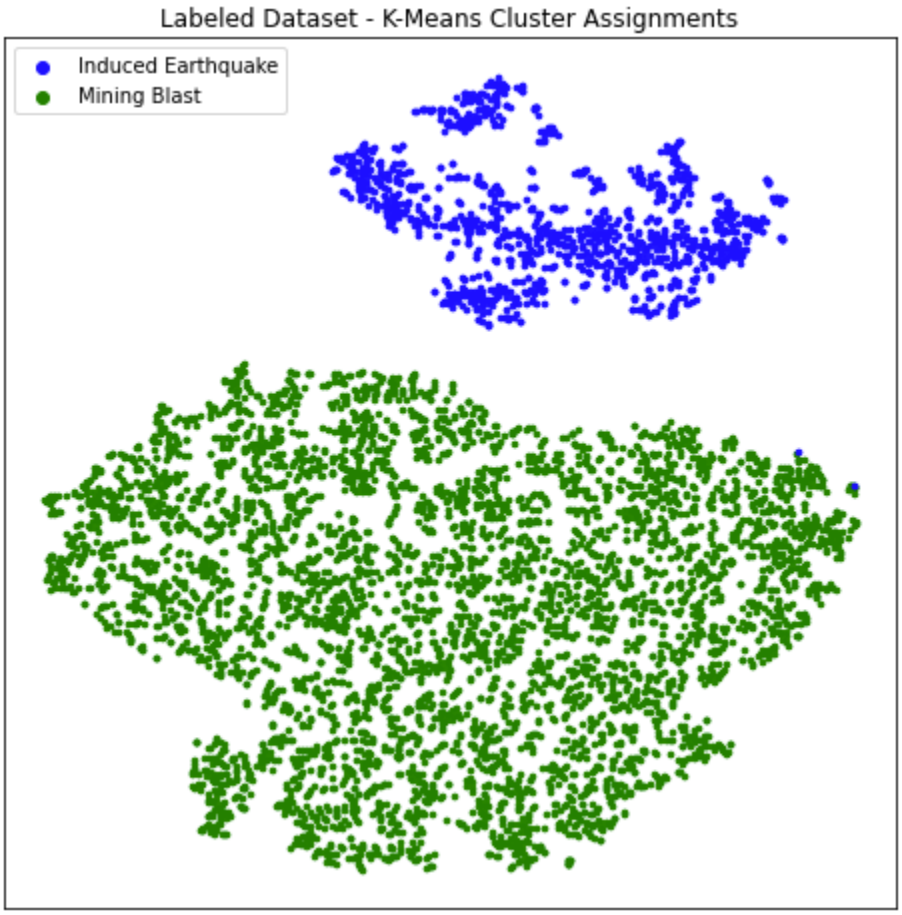

Figure 3. TSNE embedding of k-means clustering results on labeled dataset.

Misclassified waveforms are plotted in figure 4. These waveforms are somewhat ambiguous; they contain strong impulsive content early on, which is characteristic of earthquakes, but also have energy spread throughout their duration, which is characteristic of blasts. It makes sense that our model struggles with these particular events.

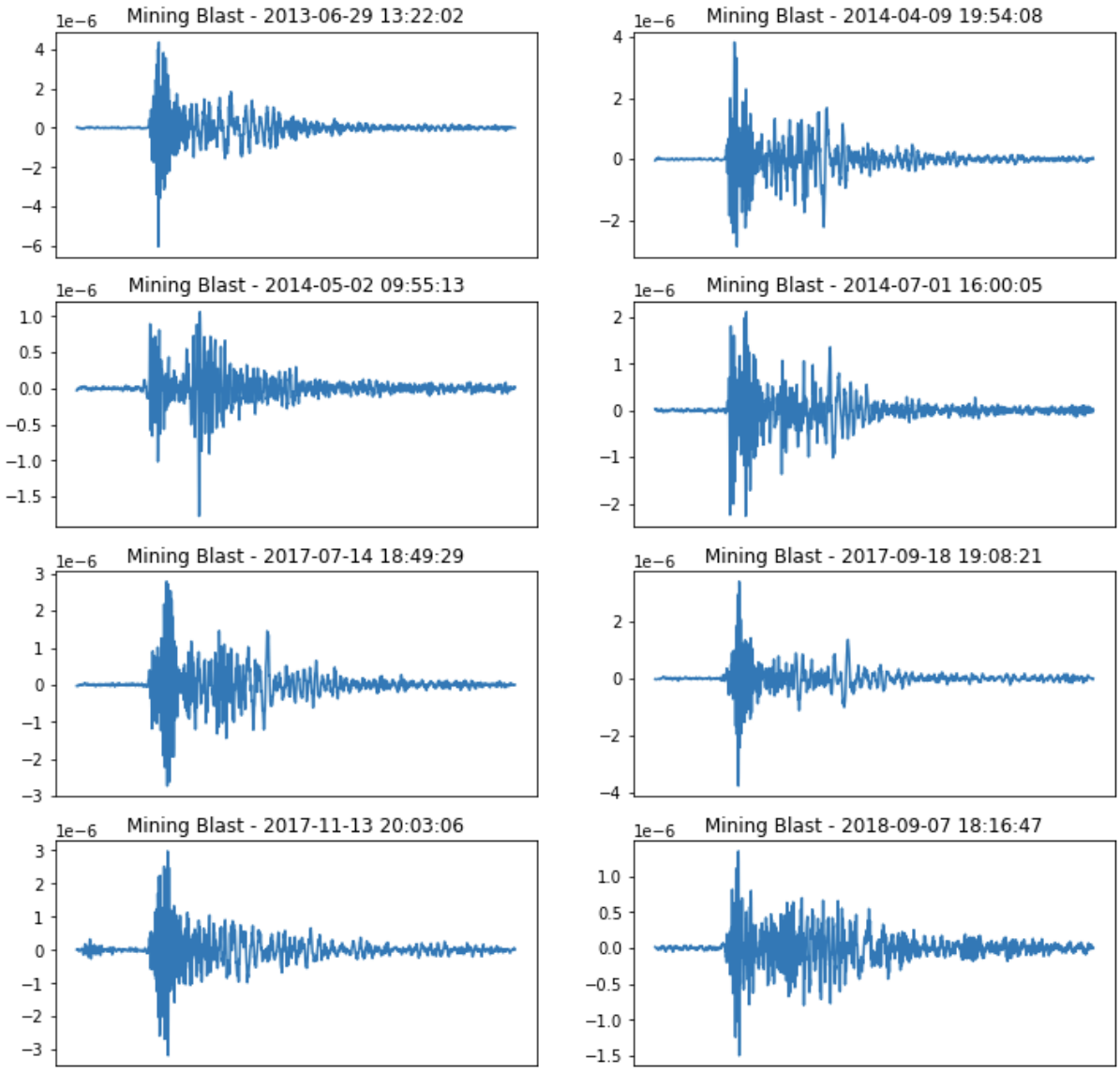

Figure 4. The eight mining blasts our k-means model incorrectly assigns to the earthquake cluster. These waveforms have characteristics of both earthquakes and blasts.

Traditional Machine Learning - PhaseNet Dataset

Based on the results achieved on the labeled dataset, we conclude that our modeling choices are suitable for clustering seismic events. We apply the same feature set to a k-means analysis of the unlabeled PhaseNet waveforms. An elbow plot (figure 5) is used to confirm that the optimal number of clusters is still two.

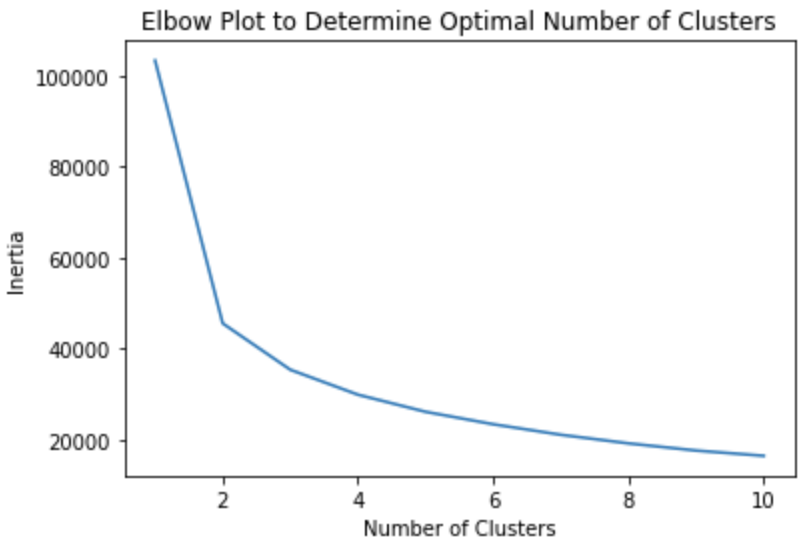

Figure 5. Elbow plot confirming that the optimal number of clusters in the PhaseNet dataset is two.

As with the labeled dataset, PCA is used to reduce the dimensionality of the feature space. In this case, four components are necessary to explain 82.2% of the original variance. See appendix figure A2 for the PhaseNet PCA loadings.

Clustering results from a two-cluster k-means model trained on the PhaseNet dataset are shown in figure 6. Based on the TSNE plots, results for both datasets appear to be consistent. However, since we have no way of calculating clustering accuracy without PhaseNet labels, alternate assessment methods are needed. We use the following list of criteria.

1. Manual expert review of clusters using a visualization tool developed for this purpose. The tool can be found at https://storage.googleapis.com/d3_phasenet_vis/trad.html, and is used to label the clusters as containing earthquakes or blasts, which is necessary for (2) and (3).
2. Examination of an hour-of-onset histogram. There should be very few mining blasts outside of working hours, while earthquakes may occur at any time throughout the day or night.
3. Examination of the waveforms predicted to be earthquakes, binned by month. There are four known earthquake clusters (e.g. earthquake swarms) in the years spanned by the PhaseNet dataset: Sep-Oct 2013, May 2014, Sep 2015, and Nov 2016 [2]. All of these clusters should be easily visible in a histogram.

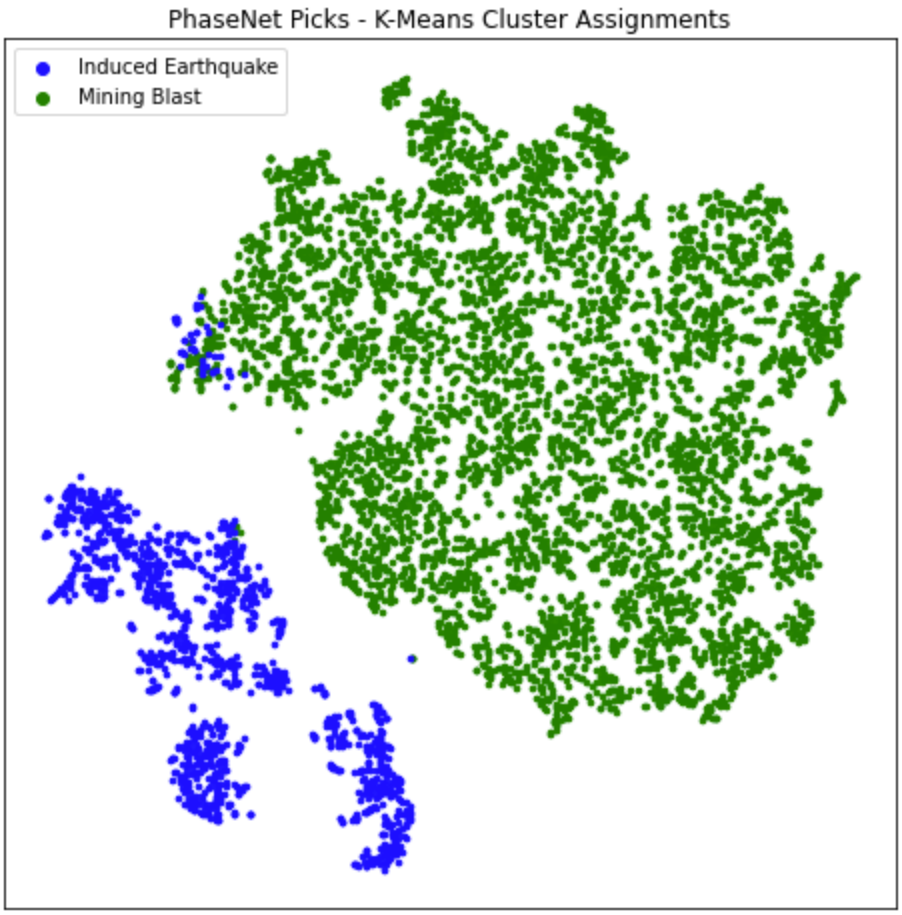

Figure 6. TSNE embedding of k-means clustering results on PhaseNet dataset.

Using the visualization tool listed in (1) above, we determine that the smaller blue cluster does indeed contain most of the induced earthquakes, while the larger green cluster contains most of the mining blasts. Based on this, the histograms described in (2) and (3) are created. See figures 7 and 8.

The histograms suggest accurate clustering results. Events in the mining blast cluster almost exclusively occur during business hours, while events in the earthquake cluster occur more uniformly throughout the day. In addition, all of the major earthquake clusters can be found in figure 8.

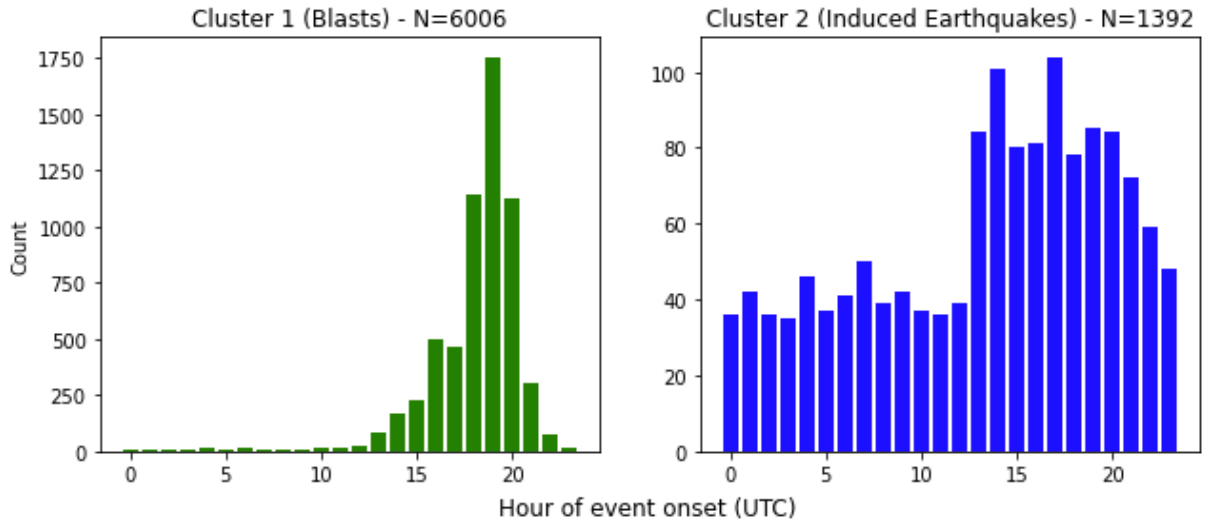

Figure 7. Hour-of-onset histogram created from the PhaseNet clusters. Mining blasts are focused in the afternoon hours, while earthquakes occur more uniformly throughout the day.

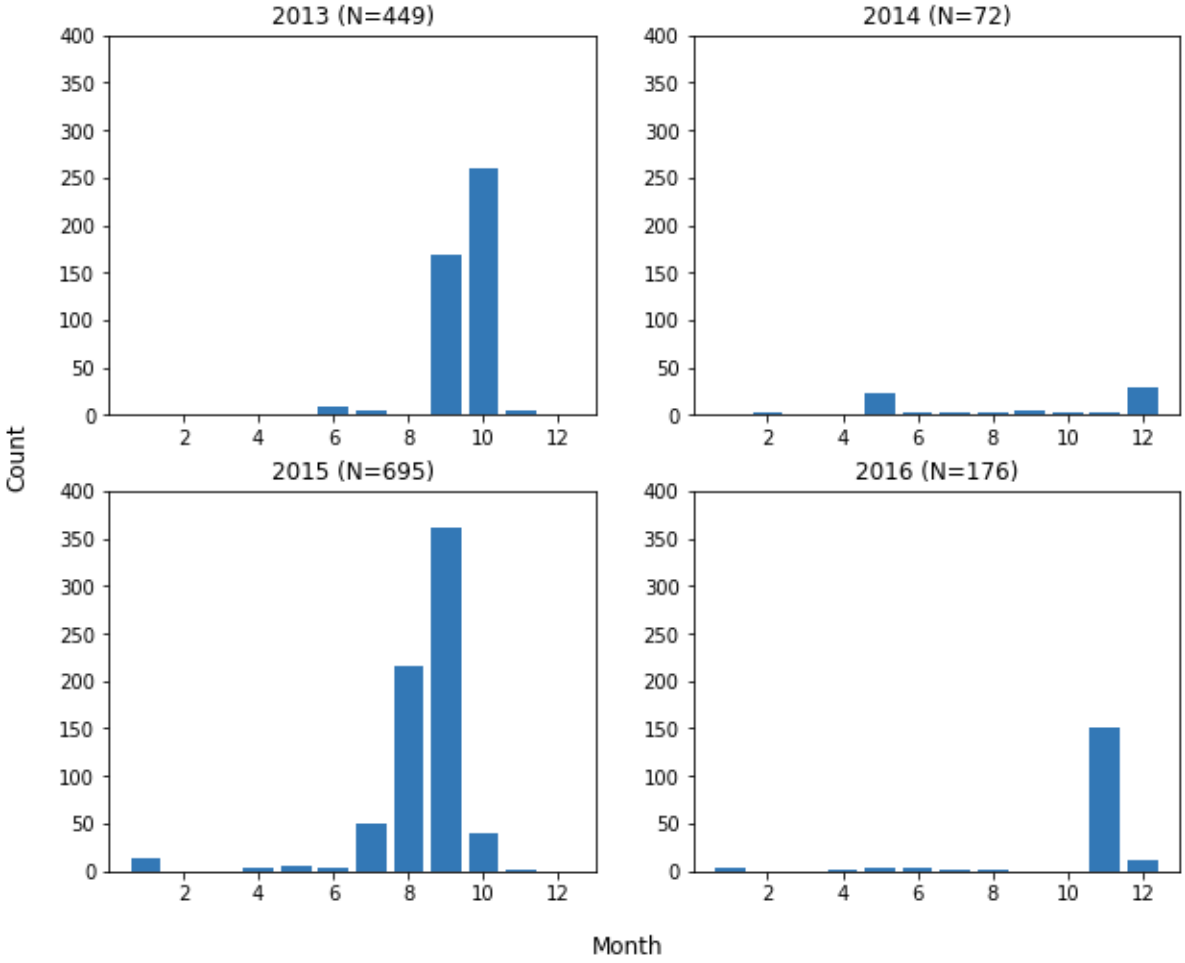

Figure 8. Induced earthquake counts, binned by month. The Sep-Oct 2013, May 2014, Sep 2015, and Nov 2016 earthquake clusters are all visible in the histogram.

Deep Learning

Our deep learning analyses are carried out using DEC, an autoencoder-based deep clustering architecture. Training a DEC network is a two step process. First, a stacked autoencoder (SAE) is pre-trained on the data to provide initial feature vectors (e.g. the bottleneck layer), which are then clustered using k-means to provide initial cluster centroids. We use a 10-dimensional bottleneck layer for all analyses [6].

The second step of DEC training is to simultaneously tune the encoder portion of the SAE, to provide feature vectors more appropriate for k-means clustering, and the cluster centroids themselves. This is done using a KL divergence loss function, where the distributions in question are the soft clustering assignments to the current cluster centroids, and an auxiliary distribution where the soft assignments are "tightened up." That is, in the auxiliary distribution high probability assignments are made even higher, and low probability assignments are made even lower. In this way, the DEC network learns from the high confidence soft clustering assignments [6]. See figure 9 for a diagram of the DEC architecture.

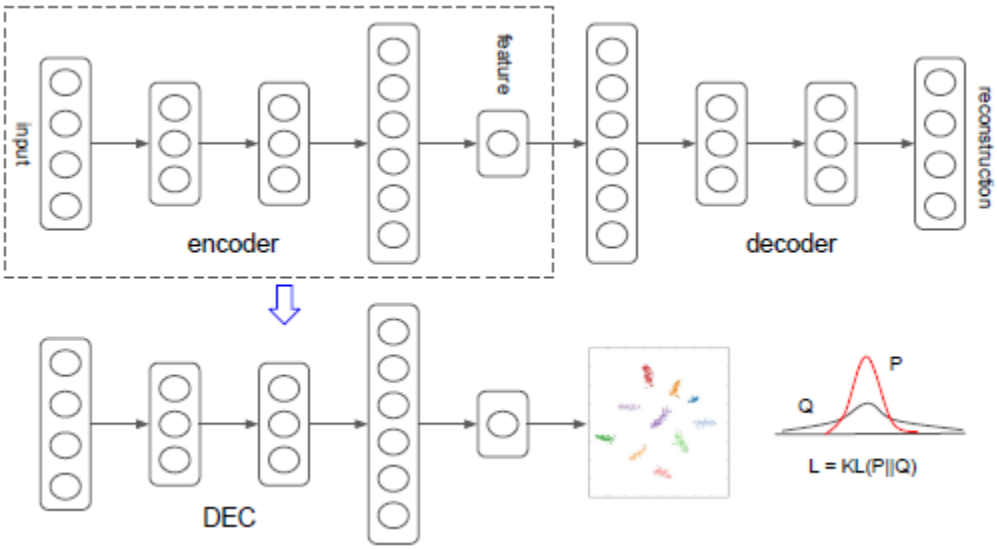

Figure 9. DEC network structure [6].

We use the DEC authors' code with some modification for our analyses [7]. To validate our DEC implementation we reproduce the DEC authors' MNIST experiment. There is some variance run-to-run, but we achieve a pretraining accuracy of about 82% with a 6-8% increase during the DEC tuning phase. These results are roughly in line with those of the DEC authors. Accuracy curves are plotted in figure 11, and other results are shown in appendix figures A3 and A4.

With our DEC implementation validated, we next apply it to the TA:O53A datasets. We compute spectrograms for all waveforms using the Short-time Fourier transform (STFT). These

spectrograms serve as our DEC training data, rather than the original waveform time-series. See data preparation notebook in supplemental materials for details.

Deep Embedded Clustering - Labeled Dataset

The labeled dataset (N=6,403) is much smaller than the MNIST dataset (N=70,000), so a simpler autoencoder is needed. In [9], the authors successfully apply a deep clustering architecture similar to DEC to the 20Newsgroup dataset (N=18,846) so we borrow their choice of autoencoder structure to analyze both the labeled and PhaseNet datasets. See DEC notebook in supplemental materials for details.

We achieve excellent clustering accuracy on the labeled dataset with this modification. On average, there are only about three incorrect cluster assignments after autoencoder pretraining. This slightly improves on the traditional analysis where eight events were misclassified. However, results following the tuning phase indicate that the labeled dataset is too small and/or simple to take advantage of the full DEC architecture end-to-end. Because results are already excellent following pretraining, the tuning phase has almost no room for improvement and often causes accuracy to slightly drop instead. See figure 11 for accuracy curves.

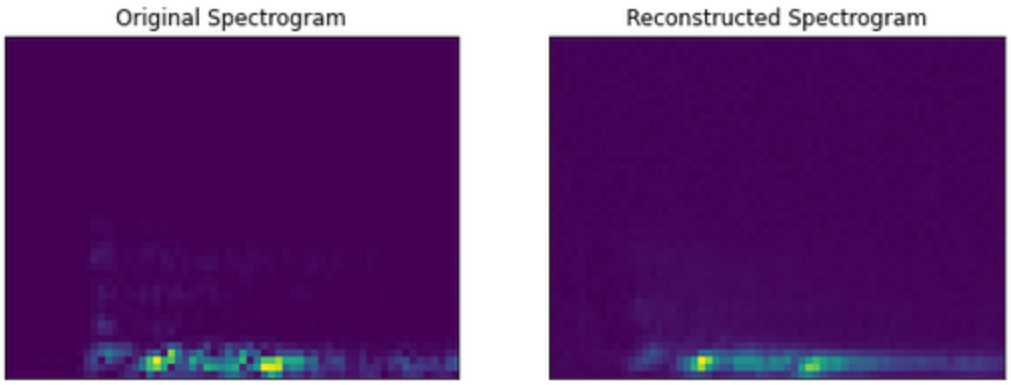

Figure 10. An STFT spectrogram and its SAE reconstruction following compression to 10-dimensions. Some detail is lost, but the overall character of the event remains, indicating an effective autoencoder architecture for clustering.

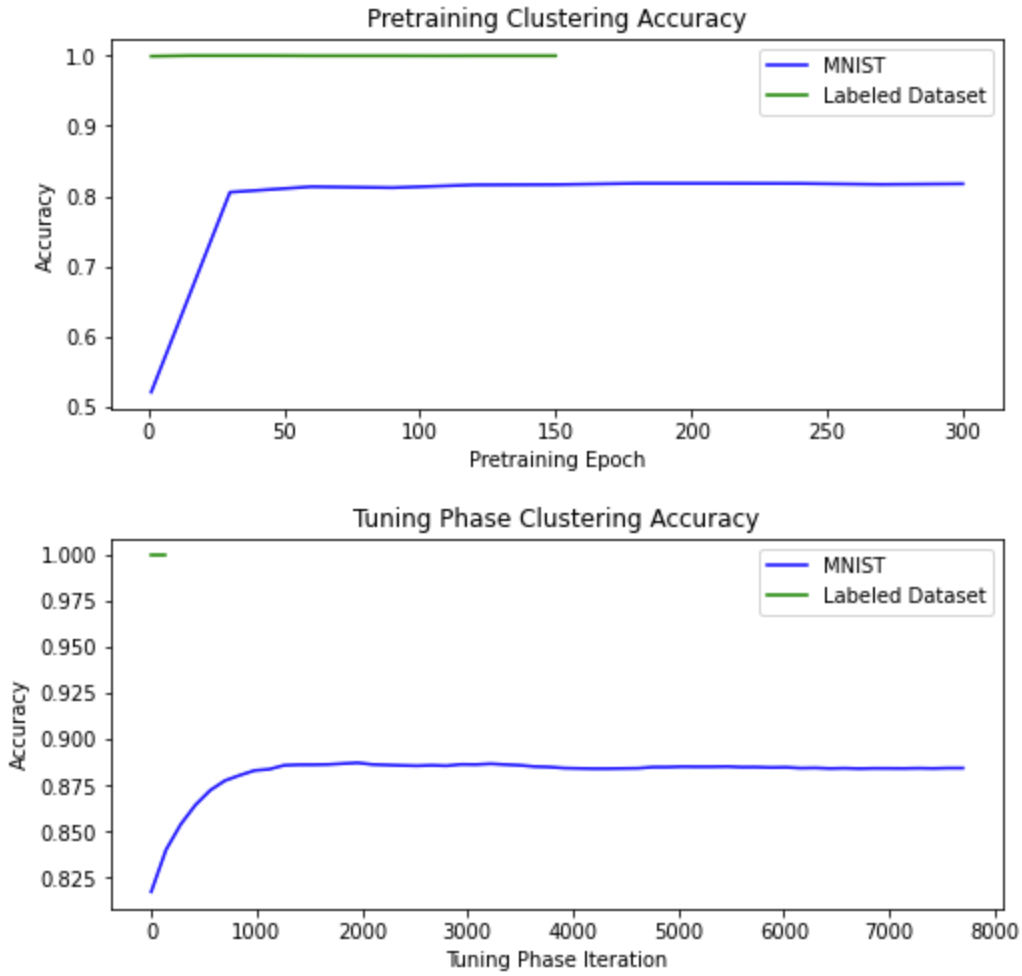

Figure 11. Pretraining and tuning phase accuracy curves for the MNIST and labeled waveform datasets. These results suggest the labeled dataset is too small and/or simple to benefit from the full DEC architecture. While the MNIST dataset requires nearly 8,000 tuning phase iterations to converge, the labeled dataset converges almost immediately.

Deep Embedded Clustering - PhaseNet Dataset

Despite the poor tuning phase results, the pretraining accuracy on the labeled dataset proves that the autoencoder architecture from [9] is suitable for encoding our seismic datasets. We apply the same network to the PhaseNet spectrograms. Unlike the labeled dataset, the PhaseNet dataset does benefit from the tuning phase. This is likely because the PhaseNet dataset is not curated and contains more noisey, ambiguous events making it a “harder” dataset. See figure 12 for PhaseNet DEC clustering results, before and after the tuning phase.

PhaseNet DEC results are evaluated using the same three criteria listed in the traditional machine learning section; manual inspection using a visualization tool, examination of an hour-of-onset histogram, and examination of earthquake counts, binned by month. The DEC

visualization tool is hosted at https://storage.googleapis.com/d3_phasenet_vis/2clusters.html. The histograms are plotted in figures 13 and 14.

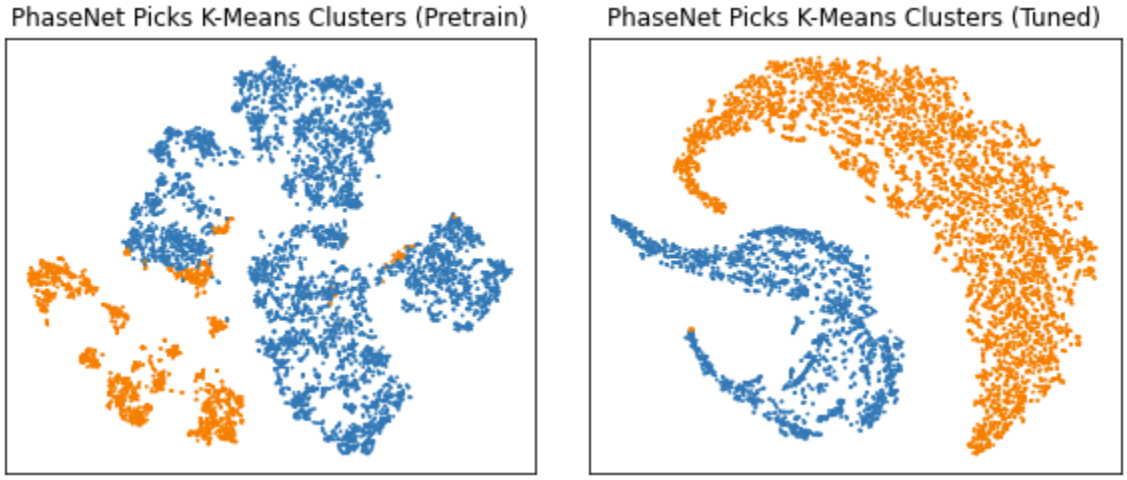

Figure 12. Pre and post-tuning phase DEC clustering results on the PhaseNet dataset. The PhaseNet dataset greatly benefits from the tuning phase, as the pre-tuning cluster structure is very weak.

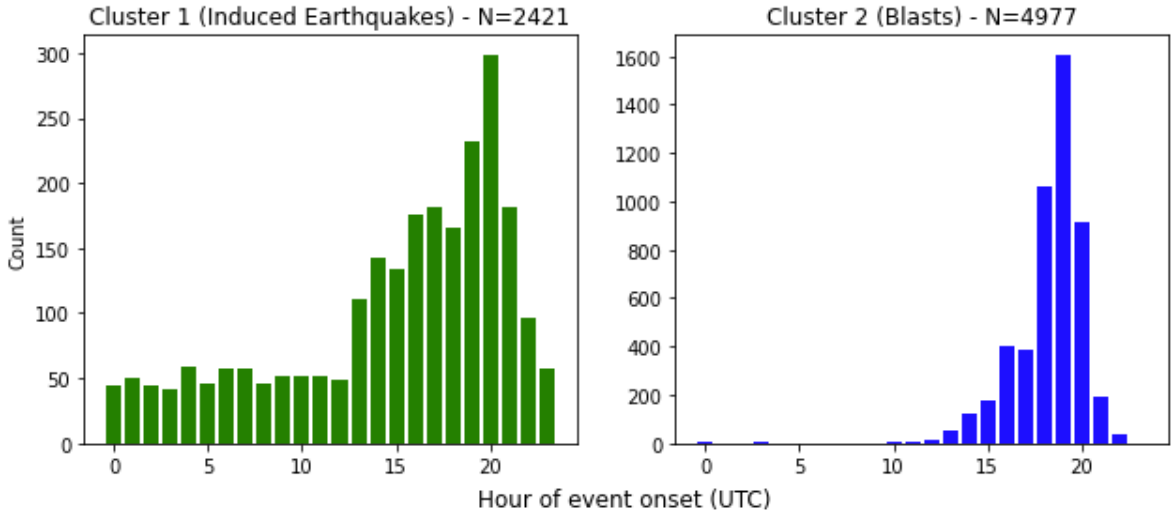

Figure 13. Hour-of-onset histogram created from the DEC PhaseNet clusters. Results are similar to the traditional analysis, but more events occurring during the workday are classified as earthquakes.

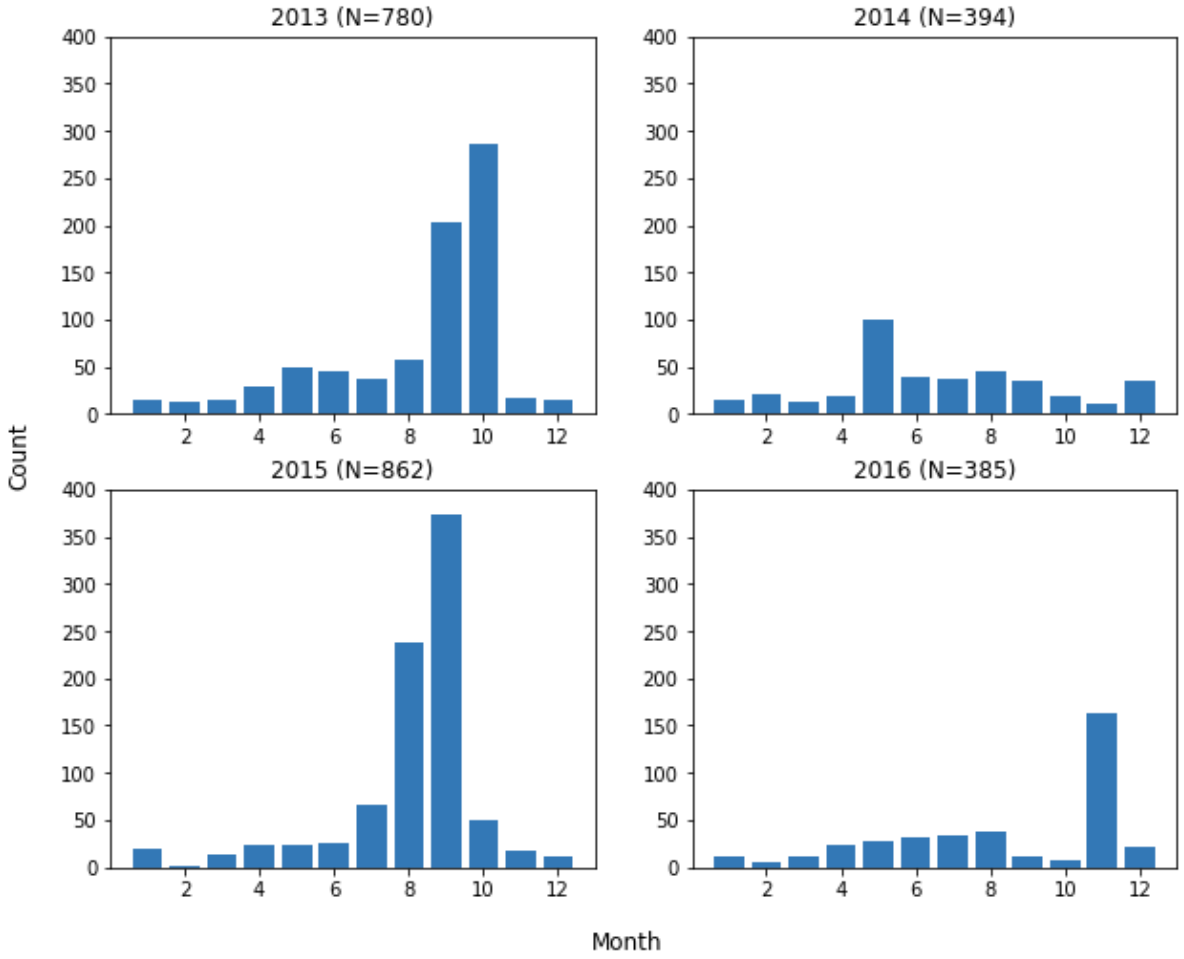

Figure 14. Induced earthquake counts from DEC PhaseNet results, binned by month. The results are similar to the traditional analysis, but there are more events classified as earthquakes in all months.

While these results "pass" all three criteria, they seem slightly worse than the traditional analysis results. There are more events in the earthquake cluster occurring during business hours, and although all major earthquake clusters are visible in the second histogram there appears to be more events classified as earthquakes in all months. This suggests that the DEC earthquake cluster includes some events that should be assigned to the mining blast cluster.

Discussion

Our results show that both the traditional machine learning and deep learning approaches can be used to separate HF-induced earthquakes and mining blasts in the PhaseNet dataset. Each approach has its own advantages and disadvantages.

Traditional Machine Learning

The most obvious advantage of the traditional k-means clustering analysis is that it seems to make fewer mistakes on the PhaseNet dataset compared to the deep learning approach, as the latter appears to include some mining blasts in the earthquake cluster. The traditional k-means analysis is also more interpretable due to using a manually engineered feature set. However, this interpretability is somewhat obfuscated by the use of PCA which necessitates consulting the PCA loadings to determine what each principal component represents.

The major disadvantage of the traditional k-means analysis is that it requires seismology domain expertise to engineer discriminative features. Traditional machine learning approaches might also benefit less from increased data volume compared to deep learning approaches.

Deep Learning

The primary advantage of the DEC architecture is that it doesn't require manually engineered features. This would theoretically allow us to apply a DEC-based solution to *any* seismograph station and immediately begin analyzing the various event types recorded by it without any prior knowledge of the area. The DEC architecture would also likely benefit more from increased data volume than the traditional k-means analysis.

The main disadvantage of DEC is that it seems to achieve slightly lower clustering accuracy on the PhaseNet dataset compared to the traditional k-means analysis. During our work on this project, we noticed that the DEC tuning phase tends to drive cluster sizes to be more equal. That is, the smaller cluster seems to always receive more events during tuning, and never lose any. This behavior suggests that although deep learning approaches clearly show potential for this task, there may be other deep clustering architectures better suited to analyzing imbalanced datasets.

Predicting New Data

Since both approaches utilize centroid-based clustering, soft cluster assignments can be computed for newly recorded events and used as a basis for classification. The computation is relatively simple (see equation 1). An appropriate classification threshold can be selected by consulting ROC or Precision-Recall curves [10].

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_j\|}{\|\mathbf{x}_i - \mathbf{c}_k\|} \right)^{\frac{2}{m-1}}}$$

Equation 1. Formula for computing the probability that data point *i* belongs to cluster *j*. This gives rise to a valid probability distribution, as cluster assignment probabilities sum to one [11].

Conclusion

Based on our results we conclude that the proposed pipeline is an effective tool for isolating and classifying seismic events recorded in continuous seismograms. As indicated by the earthquake monthly count histograms, the Phasenet tool appears to locate all interesting events in the timeframe of interest. Likewise, both of our clustering approaches successfully separate HF-induced earthquakes and mining blasts, the two primary event types encountered in eastern Ohio. The pipeline should work well in other regions as well, although the optimal number of clusters will likely vary. Further experimentation is needed to confirm that the pipeline is portable.

Future Work

The most critically needed enhancement of this work is the ability to quantify the PhaseNet clustering results. One potential solution is to use the 2017-2018 events in the labeled dataset as a held-out test set, since the PhaseNet dataset only includes events through 2016. However, the labeled dataset contains only four HF-induced earthquakes in these years so more would need to be collected.

We believe it might also be worthwhile to investigate replacing the STFT with other signal transformations, such as the continuous wavelet transform. The authors in [9] reported favorable results when using the scattering wavelet transform to preprocess the MNIST dataset before clustering, so we find it to be another intriguing candidate.

Finally, selecting a different deep clustering architecture might address DEC's shortcomings on imbalanced datasets. One possibility is JULE, which is based on hierarchical clustering [12].

Statement of Work

As trained seismologists, Dr. Huang and Dr. Yao served as domain expert advisors. Dr. Yao implemented the PhaseNet tool and provided the picks with >= 0.8 probability. He also devised our signal filtering scheme when downloading data from obspy. Jeff completed all machine learning analyses and wrote the paper.

Acknowledgements

We would like to thank Stanford's Dr. Mostafa Mousavi for his assistance in troubleshooting our STFT spectrograms and autoencoder implementation. We would also like to thank the University of Michigan's Dr. Elle O'Brien for participating in frequent progress reviews with us, ensuring our project stayed on the right track.

References

1. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. https://academic.oup.com/gji/article/216/1/261/5129142.
2. Maturity of nearby faults influences seismic hazard from hydraulic fracturing. https://www.pnas.org/content/115/8/E1720.
3. Unsupervised Clustering of Seismic Signals Using Deep Convolutional Autoencoders. https://ieeexplore.ieee.org/document/8704258.
4. Identifying Different Classes of Seismic Noise Signals Using Unsupervised Learning. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL088353.
5. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. https://www.nature.com/articles/s41467-020-17841-x.
6. Unsupervised Deep Embedding for Clustering Analysis. https://arxiv.org/abs/1511.06335.
7. Deep Embedded Clustering GitHub Repo. https://github.com/XifengGuo/DEC-keras.
8. Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018JB016661.
9. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. https://arxiv.org/abs/1610.04794.
10. How to Use ROC Curves and Precision-Recall Curves for Classification in Python. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/.
11. Confidence in k-means. https://towardsdatascience.com/confidence-in-k-means-d7d3a13ca856.
12. Joint Unsupervised Learning of Deep Representations and Image Clusters. https://arxiv.org/abs/1604.03628.

Appendix

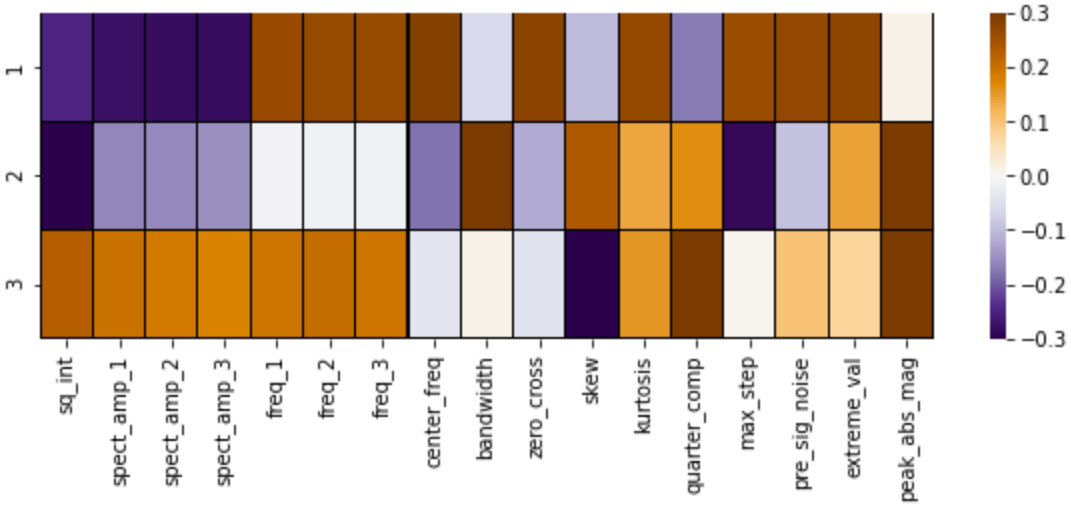

Figure A1. PCA loadings for the labeled dataset.

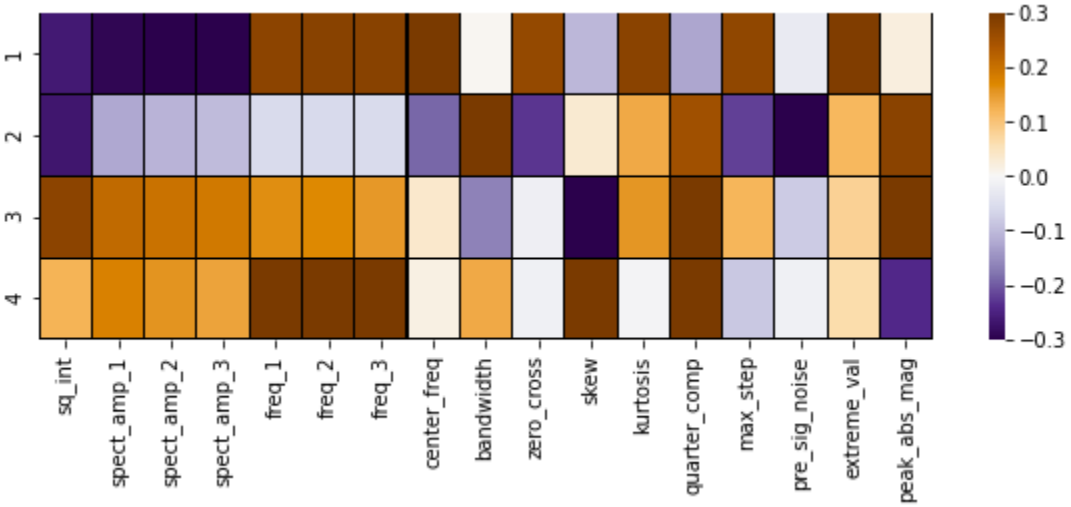

Figure A2. PCA loadings for PhaseNet dataset.

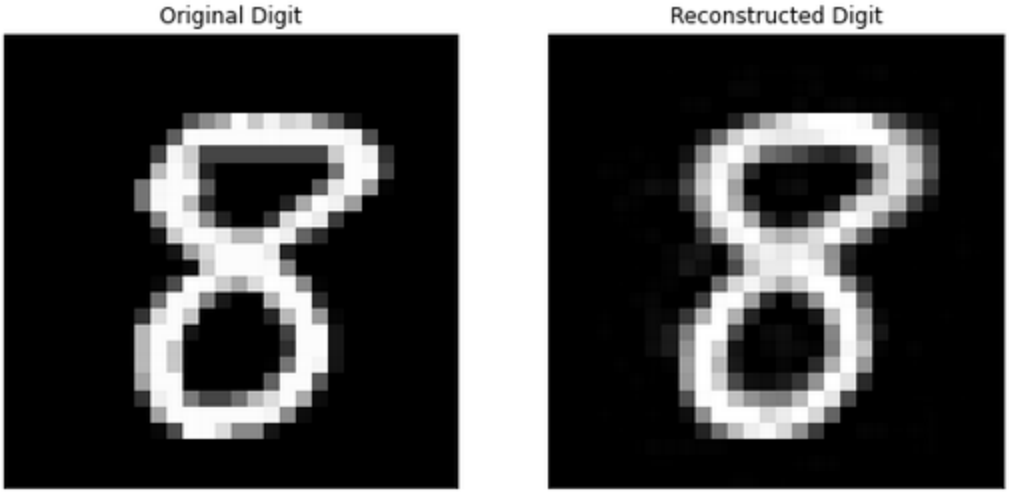

Figure A3. An MNIST digit and its SAE reconstruction following compression to 10-dimensions.

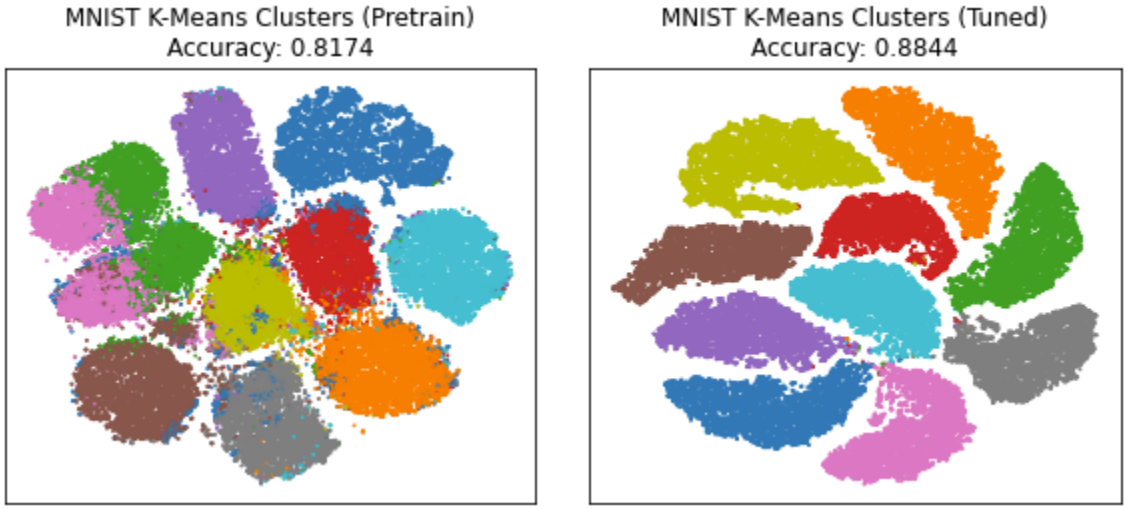

Figure A4. MNIST clustering results following pretraining (left) and DEC tuning phase (right).