Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL088353

Identifying Different Classes of Seismic Noise Signals Using Unsupervised Learning

Christopher W. Johnson[1,2], Yehuda Ben-Zion[3], Haoran Meng[3], and Frank Vernon[1]

[1]Scripps Institution of Oceanography, University of California, San Diego, San Diego, CA, USA, [2]Now at Los Alamos National Laboratory, Los Alamos, NM, USA, [3]Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA

Key Points:
- Emergent and impulsive signals in continuous seismic waveforms are identified using cluster analysis on a dense array data
- An unsupervised learning model is trained to identify multiple classes of noise using temporal and spectral data features
- A more complete understanding of seismic noise signals will improve the ability to detect genuine microseismic events

Supporting Information:
- Supporting Information S1
- Movie S1

Correspondence to:
C. W. Johnson,
cwj004@ucsd.edu

Citation:
Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes of seismic noise signals using unsupervised learning. *Geophysical Research Letters*, *47*, e2020GL088353. https://doi.org/10.1029/2020GL088353

Received 9 APR 2020
Accepted 24 JUN 2020
Accepted article online 8 JUL 2020

Abstract Proper classification of nontectonic seismic signals is critical for detecting microearthquakes and developing an improved understanding of ongoing weak ground motions. We use unsupervised machine learning to label five classes of nonstationary seismic noise common in continuous waveforms. Temporal and spectral features describing the data are clustered to identify separable types of emergent and impulsive waveforms. The trained clustering model is used to classify every 1 s of continuous seismic records from a dense seismic array with 10–30 m station spacing. We show that dominate noise signals can be highly localized and vary on length scales of hundreds of meters. The methodology demonstrates the complexity of weak ground motions and improves the standard of analyzing seismic waveforms with a low signal-to-noise ratio. Application of this technique will improve the ability to detect genuine microseismic events in noisy environments where seismic sensors record earthquake-like signals originating from nontectonic sources.

Plain Language Summary Improvements in microseismic detection will advance observations of failure processes on faults subjected to slowly accumulating tectonic stress. Continuous seismic waveforms contain copious variations of nontectonic signals that inhibit the detection of genuine microearthquakes. Developing a framework to identify emergent and impulsive signals originating from natural and anthropogenic activity will advance seismic network monitoring capabilities. We apply unsupervised machine learning techniques to classify multiple types of weak ground motions that occur ubiquitously in continuous seismic records. A trained model is used to label every 1 s of a dense geophone array to provide a high-resolution description of the anatomy of continuous seismic records. Further methodology developments and application in various environments have high potential for improving the ability to monitor earthquakes and other sources of ground motion.

1. Introduction

Seismic waveforms contain abundant information that traditional signal processing techniques do not utilize fully, allowing for new discoveries to characterize weak ground motions. Advancements in machine learning and large seismic data sets provide opportunities to explore continuous waveforms and identify new signals (Bergen et al., 2019; Kong et al., 2018). Machine learning is a tool to map relationships through a functional model defined using information extracted from data. Various studies concluded that supervised machine learning can identify subtle seismic waveform variations to classify earthquakes, tremors, landslides, avalanches, and mining explosions (Aguiar & Beroza, 2014; Hammer et al., 2013; Linville et al., 2019; Mousavi et al., 2016; Perol et al., 2018; Ross, Meier, & Hauksson, 2018; Ross, Meier, Hauksson, & Heaton, 2018; Rouet-Leduc et al., 2019). Correctly labeled data with a diverse mixture of seismic signals and a large enough quantity with respect to the complexity of the classification task and type of employed model are necessary for supervised model training (Bishop, 2006). Unsupervised machine learning infers a functional model from the underlying structure of features describing the data without any a priori classification or categorization (Chen, 2017; Holtzman et al., 2018; Mousavi et al., 2019; Yoon et al., 2015). Applying unsupervised learning to data features extracted from seismic waveforms may be used to label nontectonic weak ground motions that regularly occur in continuous seismic recordings and originate from multiple sources.

Labeling new classes of seismic signals requires a clear understanding of weak sources of ground motion that recur as emergent and impulsive nontectonic signals in the continuous wavefield (Gradon et al., 2019;

Johnson et al., 2019; Meng & Ben-Zion, 2018a). Seismic waveforms contain a superposition of naturally occurring nontectonic processes that produce ongoing weak ground motions such as atmospheric pressure variations (Qin et al., 2019; Sorrells & Goforth, 1973; Tanimoto & Wang, 2018), temperature changes (Hillers & Ben-Zion, 2011; Johnson et al., 2019), interactions of ocean waves with rocks (Gerstoft & Tanimoto, 2007; Hillers et al., 2012), and wind (De Angelis & Bodin, 2012; Johnson, Vernon, et al., 2019). Similarly, anthropogenic activity such as automobile, train, and air traffic and wind power generation produce observable nontectonic ground motions that contribute to the seismic noise wavefield (Inbal et al., 2018; Marcillo & Carmichael, 2018; Meng & Ben-Zion, 2018a) and may be utilized for monitoring fault zones (e.g., Brenguier et al., 2019). Nontectonic ground motions occupy a significant fraction of the day and can obscure low-amplitude tectonic events such as microseismicity and tremors (Inbal et al., 2018; Johnson, Vernon, et al., 2019; Meng et al., 2019; Meng & Ben-Zion, 2018a). Robust earthquake detections using data-driven techniques require correctly identifying emergent and impulsive waveforms as ground motions produced by sources below the surface (Ross, Meier, Hauksson, & Heaton, 2018). Spatially dense seismic arrays can identify earthquakes and tremor that have a signal-to-noise ratio close to 1 (Ben-Zion et al., 2015; Inbal et al., 2016; Li et al., 2018; Meng & Ben-Zion, 2018b) and are embedded in waveforms that are dominated by multiple weak nontectonic sources originating at and above the surface (Gradon et al., 2019; Johnson, Vernon, et al., 2019; Meng & Ben-Zion, 2018a). Correctly identifying the dominant sources of ground motions will contribute to understanding the composition of continuous waveforms and improve the ability to track tectonic faulting and image fault zone structures.

This study implements an unsupervised data-driven machine learning methodology to describe the anatomy of waveforms. The focus is on nontectonic emergent and impulsive noise signals. Supervised machine learning models can be trained to accurately identify earthquake phase arrivals, even when applied to unfiltered data that contain low signal-to-noise earthquake signals (Zhu & Beroza, 2018). The majority of seismic waveforms contain signals from unknown sources (Meng et al., 2019), and unsupervised machine learning can exploit information coded in data to develop novel labels for new signals to more completely describe the recorded weak ground motions. Using data from a spatially dense seismic array, feature vectors are calculated for >170 million wavelets to subdivide the emergent and impulsive noise signals using cluster analysis. Known earthquake waveforms are excluded from the training process, and the noise signals are grouped into five classes with varying spectral and temporal waveform properties. The trained models allow for rapid classification of large volumes of continuous seismic waveforms recorded by the dense seismic network. Coherent signals from multiple classes of weak ground motion are observed across the array. Additional progress can lead to new discoveries about the full range of weak ground motion sources and significant improvements in the ability to detect small genuine microearthquakes and tremor.

2. Seismic Waveform Features for Training Data

2.1. Seismic Data

The data are continuous waveforms from 1,108 vertical component ZLand geophones (10 Hz) recording from 8 May 2014 to 7 June 2014 (days of year 128 to 158) at 500 Hz in a densely spaced square configuration, extending 600 × 600 m with sensors aligned in rows orthogonal to the San Jacinto fault trace (Ben-Zion et al., 2015). The configuration of the dense deployment consists of 20 rows spaced about 30 m apart with 50 geophones spaced approximately 10 m; the remaining 108 sensors extend multiple rows in the study area (supporting information Figure S1). The geophone locations were measured using a real-time-kinematic global positioning system survey with submeter accuracy for the relative position and elevation. Located within the dense deployment is the Plate Boundary Observatory borehole seismometer B946 at a depth of 147 m. The borehole station is a short-period three-component sensor with a 1 Hz corner frequency recording at 100 samples per second and provides the timing of *P*-wave arrivals at the site. A 30 day catalog containing 805 local and regional handpicked *P*-wave arrival times for all $M \geq 0$ earthquakes within about 100 km (Ben-Zion et al., 2015) and all $M \geq 2$ earthquakes within 200 km is used to remove earthquake waveforms from the training data set. This ensures that known tectonic signals are excluded from the analysis.

2.2. Waveform Features for Unsupervised Model Training

Seismic waveforms on 25–26 May 2014 (days of year 145 and 146) for every sensor are used to develop the training data set. A 1 Hz high-pass filter is applied to the daily waveforms, and the data units are in

counts. The waveforms are cut into 1 s intervals (500 samples), and data within ±60 s of all *P*-wave arrivals are removed for a total of 178,683,245 unlabeled seismic noise examples.

For each waveform, feature vectors are calculated to use in clustering analysis. The data features are time and frequency domain scalar values that include the integral of the squared waveform, maximum spectral amplitude, frequency at the maximum spectral amplitude, center frequency, signal bandwidth, zero upcrossing rate, and the rate of spectral peaks (Table S1). The use of engineered features provides domain expertise in the data characteristics of interest to isolate variations in weak ground motions with changing spectral properties over distances of tens of meters (e.g., Johnson, Meng, et al., 2019). Here the focus is primarily on high-frequency (5–200 Hz) signals that are generated by different unknown sources (Meng et al., 2019). The features are selected to include energy based (e.g., Hulbert et al., 2019; Rouet-Leduc et al., 2019) and spectral characteristics (e.g., Hammer et al., 2012; Mousavi et al., 2016) that are utilized in seismic waveforms regression and classification problems. The total number of features is limited to 7 to avoid clustering a high-dimensional feature space.

The data set is shuffled and split 90/10% (160,814,916/17,868,329) for training and evaluating the cluster model. The training data features are standardized by removing the mean and scaling to unit variance. The standardization coefficients are saved and applied to all other data in the analysis. Principal component analysis (PCA) is applied to the standardized features using seven components and indicates that 99.5% of the variance is explained by six linear independent vectors. The PCA covariance matrix shows that correlation exists between the center frequency and zero upcrossing rate; whitening is applied to decorrelate the components. The whitened PCA components are saved, and all other data in the analysis are projected to this coordinate space when performing the cluster analysis.

3. Unsupervised Cluster Analysis

3.1. *k*-Means Clustering

A *k*-means cluster analysis is performed using the seven standardized feature vectors projected onto the principal component orientations. The *k*-means method is an efficient clustering algorithm but can produce poor results by assuming a flat geometry; however, all unsupervised clustering techniques suffer from limitations about the underlying structure of the data. The *k*-means algorithm randomly selects centroid vectors to partition the data into a predefined number of clusters. Model optimization is obtained by iteratively updating the centroid vectors by minimizing the within-cluster variance of each.

We are interested in evaluating the optimum number of clusters to partition the waveforms. The gap statistic is evaluated to determine the appropriate number of separable clusters by comparing the data dispersion from the centroids to the dispersion of a synthetic feature space generated using an equivalent random normal distribution (Tibshirani et al., 2001). The gap statistic is the difference in the dispersion of the clusters from the data features and null features. Clusters are formed using 2–20 centroids, and the method is reinitialized for 100 iterations using different randomly chosen centroid seeds, with the final model having the lowest dispersion from the cluster centers. For each of the 2–20 number of centroids, 19 total, the process is repeated 500 times using a population of 15,000 randomly selected data features to assess the total inertia mean and deviation for each number of clusters. The procedure uses 142.5 million data features, with about 18 million not selected. The process is repeated three times to ensure the inclusion of all data, and the results are similar for each iteration. The gap statistic shows a change point between five and nine cluster centroids (Figure S2). The rate of change decreases sharply at five centroids, which is chosen as the optimal number of separable clusters for the analysis.

The final model is calculated for five cluster centroids using the 17,868,329 evaluation data. The initial centroid values are obtained using batches of 50,000 randomly selected data, with 500 initializations each, and repeated until no improvement is obtained after 250 iterations. The initial centroids are used to seed the *k*-means model for optimization using all evaluation data and apply a cluster label to each. The clustering partitions the data into Labels 1 (19.1%), 2 (22.4%), 3 (14.6%), 4 (23.9%), and 5 (19.9%). The percentage totals are between 14.6% and 23.9% and are expected to be homogenous since the *k*-means algorithm is minimizing the centroid variance, which requires similar totals for each group.

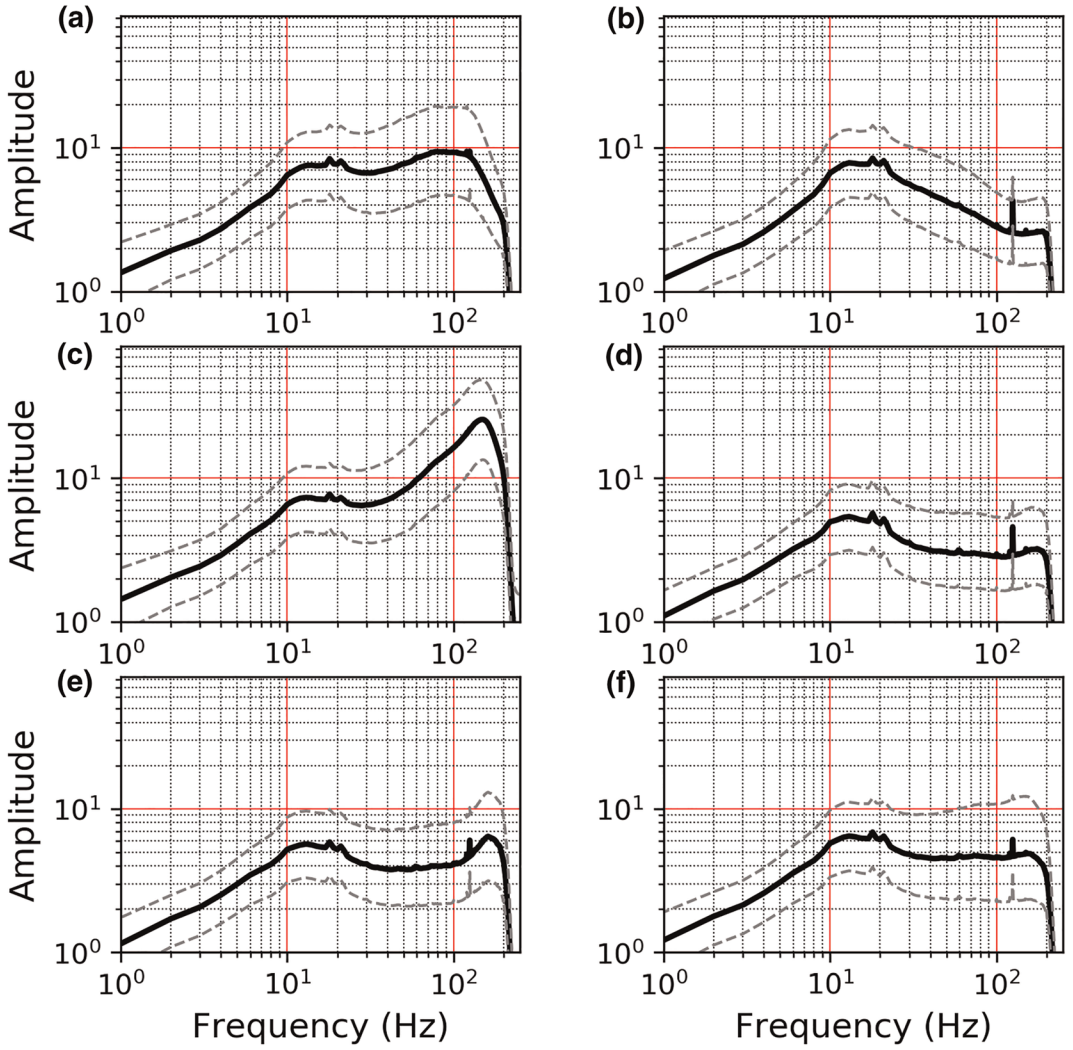

Figure 1. Stacked spectra of evaluation data waveforms for cluster labels (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, and (f) all. The median is shown with the solid black line and the 25th and 75th percentiles with the gray dashed line.

3.2. Spectral Characteristics of Waveform Clusters

The feature vectors contain spectral properties to separate the waveforms into different classes. The spectra for the five different noise clusters in the evaluation data are stacked to identify similarities and differences in the signals (Figure 1). The Label 1 spectra show spectral peaks around 15 and 100 Hz, with the higher frequencies being greater. Label 2 results show a peak around 15–20 Hz with lower amplitude across all other bands. Label 3 spectra contain high amplitude above 100 Hz and are the greatest amplitude for all the spectral groups. Label 4 and Label 5 show relatively lower amplitudes compared to the other groups, with Label 5 indicating higher amplitude above 100 Hz and Label 4 most consistent with the median of all spectra (Figure 1f). We note the similarity of the spectra for Noise Labels 4 and 5 but retain them as separate classes because the higher-frequency signals in Label 5 may reflect a different process generating ground motions.

4. Labeling the Noise in the Entire Dense Array

The daily waveforms of each geophone in the array are partitioned into 1 s intervals, and the temporal and spectral features are calculated for each. The feature vectors are standardized and projected onto the principal components estimated during the training procedure. Each feature vector is assigned the label for the

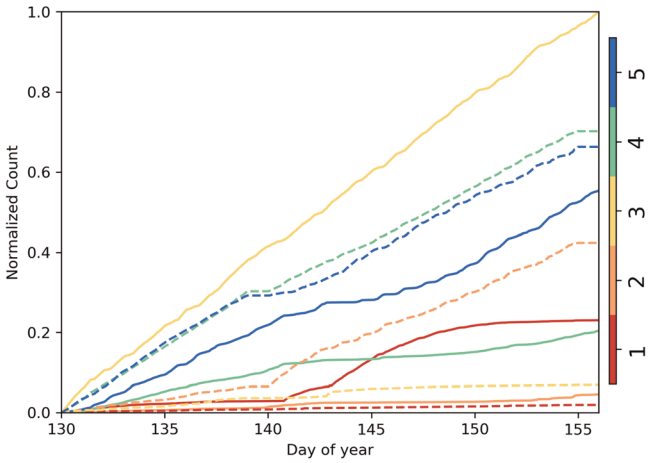

Figure 2. Normalized cumulative number of detection for the five noise labels using a geophone located on the hillslope east of the fault within the trees (solid lines) and another sensor located in basin fill north of built structures in the study area (dashed lines).

nearest cluster centroid using the trained *k*-means cluster model. The procedure produces a label representing specific spectral and temporal properties for every 1 s of seismic waveforms in the dense array, which allows analysis of the spatiotemporal variations of the weak ground motions in the study area.

Waveform noise characteristics vary over distances of tens to hundreds of meters (Johnson, Meng, et al., 2019). Examining the cumulative occurrence of each label at two geophones located in different parts of the array shows variations in the dominant source of noise (Figure 2). The locations shown were selected to highlight the quantifiable difference in weak sources of ground motion propagating in the shallow subsurface of the study area (Johnson, Meng, et al., 2019; Meng et al., 2019). The geophone located near the trees (3,514; Figure S1) has primarily Noise Label 3, with Noise Label 5, the next most recurring signal, occurring 50% less often. Noise Label 3 (Figure 1c) contains increased amplitudes >100 Hz and is consistent with wind gusts that shake vegetation and produce high-frequency weak ground motions (Meng et al., 2019). The geophone located in the basin fill west of the fault, away from dense vegetation, and north of the built structures (Figure S1) shows primarily Noise Labels 4 and 5 (Figures 1d and 1e). The spectra of these noise signals are generally lower in amplitude, with some containing higher amplitudes above 100 Hz, which is consistent with wind-generated ground motions produced by in situ built structures (Johnson, Meng, et al., 2019). The lower amplitude and broader spectrum suggest less impulsive and emergent signals at this particular location in the array. Interestingly, Noise Label 3 is the second lowest detected signal (<10%) at the geophone near the built structure and is in the part of the study area with sparse vegetation. This suggests that Noise Label 3 is separable from other types of noise and is possibly associated with ground motions produced by vegetation coupling atmospheric energy into the subsurface.

Noise labels for 15 May 2014 (day of year 135) highlight the spatial and temporal variability within the 600 × 600 m array. Figure 3 shows the maximum label occurrence during 1 hr intervals for local hour 0 (nighttime) and 12 (daytime) at each geophone. During local hour 0 when conditions are expected to be relatively quiet (Meng et al., 2019), many sensors record noise with Labels 4 and 5, which have low-amplitude broadband spectra. In areas containing built structures and in situ unused machinery, Label 2 is observed, suggesting that the detection might be a localized source (Figure S1; Johnson, Meng, et al., 2019). During midday at local hour 12, almost the entire array shows the five types of noise waveforms. Noise Labels 1, 2, and 3 are dominant across the array, possibly due to elevated daytime wind conditions that interact with surface objects and act as local sources (Johnson, Meng, et al., 2019). The findings are consistent throughout the deployment with Noise Labels 4 and 5 observed during the nighttime hours, Labels 1 and 2 throughout the daytime hours, and Label 3 observed both day and night in groups of neighboring stations (Movie S1).

The labeled 1 hr record sections during the afternoon of 15 May 2014 (day of year 135) along a transect of 54 geophones further illustrate the localization of noise sources and the ability of the cluster model to separate them spatially and temporally (Figure 4a). The percentage for each class is 16.6, 24.3, 7.6, 29.4, and 22.1 for Labels 1, 2, 3, 4, and 5, respectively. Specific locations have dominant noise labels that concentrate in areas with 10–100 m spacing. The first transect at 20 m is a mixture of all labels with consecutive observations of Label 3 showing increased amplitudes for many minutes. Moving east the amplitudes increase and correspond to the location of ground shaking produced by structures on the property (Johnson, Meng, et al., 2019). The largest amplitude signals between 120 and 200 m transition from a mix of Labels 2 and 4 to high-amplitude waveforms with many as Label 5, but all classes of noise waveforms are present. Beyond 200 m the amplitudes decrease and the waveforms show all noise signals with increased amplitudes between 300 and 400 m in the fault damage zone (Ben-Zion et al., 2015).

Focusing on a single record section highlights the diversity of nontectonic signals throughout the hour (Figure 4b). The lowest amplitudes signals are generally Label 5, while the strongest impulsive signals are

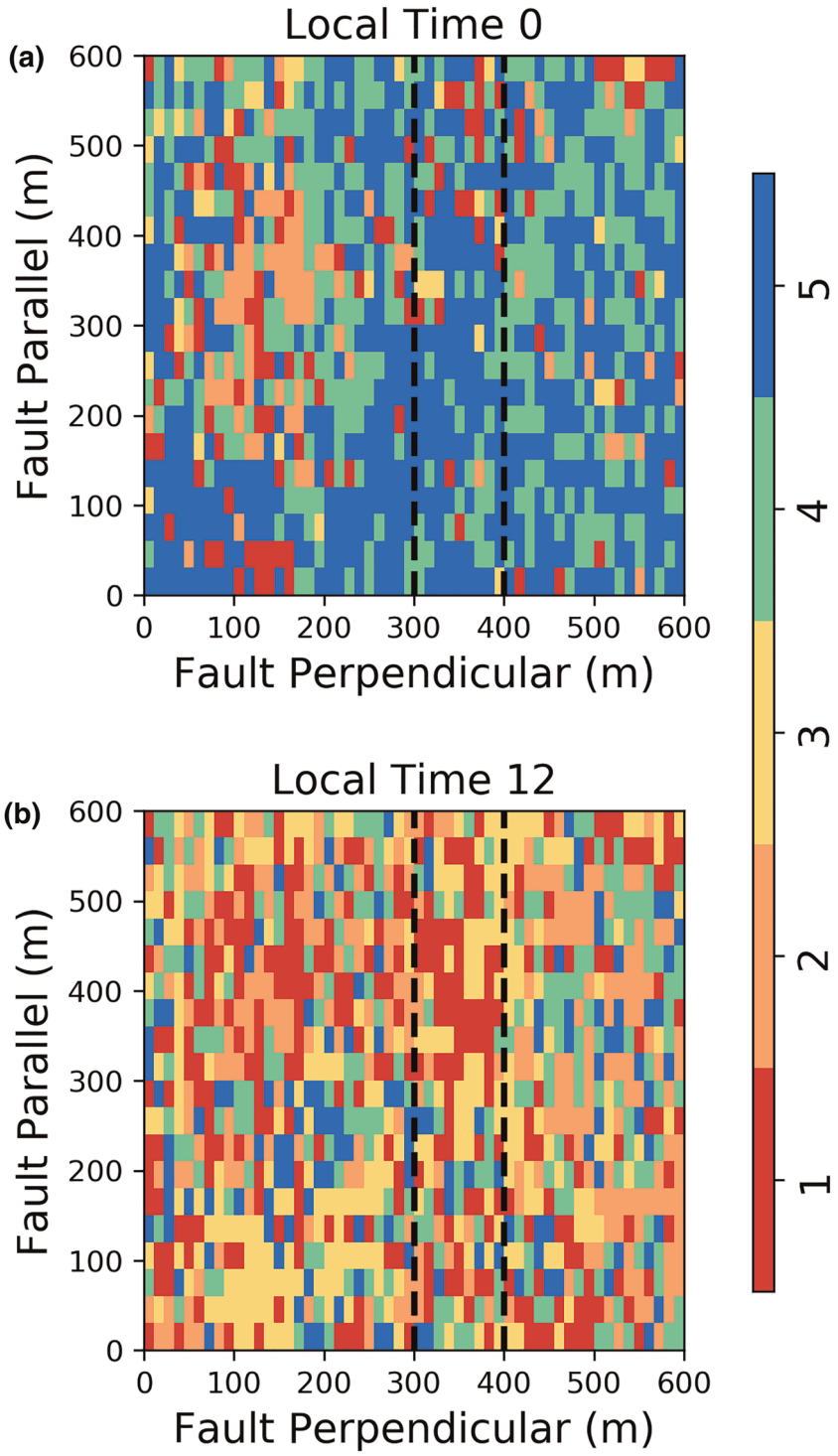

Figure 3. Cluster labels for each geophone in the array as the maximum noise label occurring during the 1 hr period. (a) Snapshot at midnight local time shows the majority of stations as Labels 4 and 5 during the quiet nighttime hours. (b) During daytime hours, the waveforms exhibit a mixture of Labels 1, 2, and 3 with clustering of noise types in specific locations. The vertical black dashed lines indicate the location of the fault damage zone.

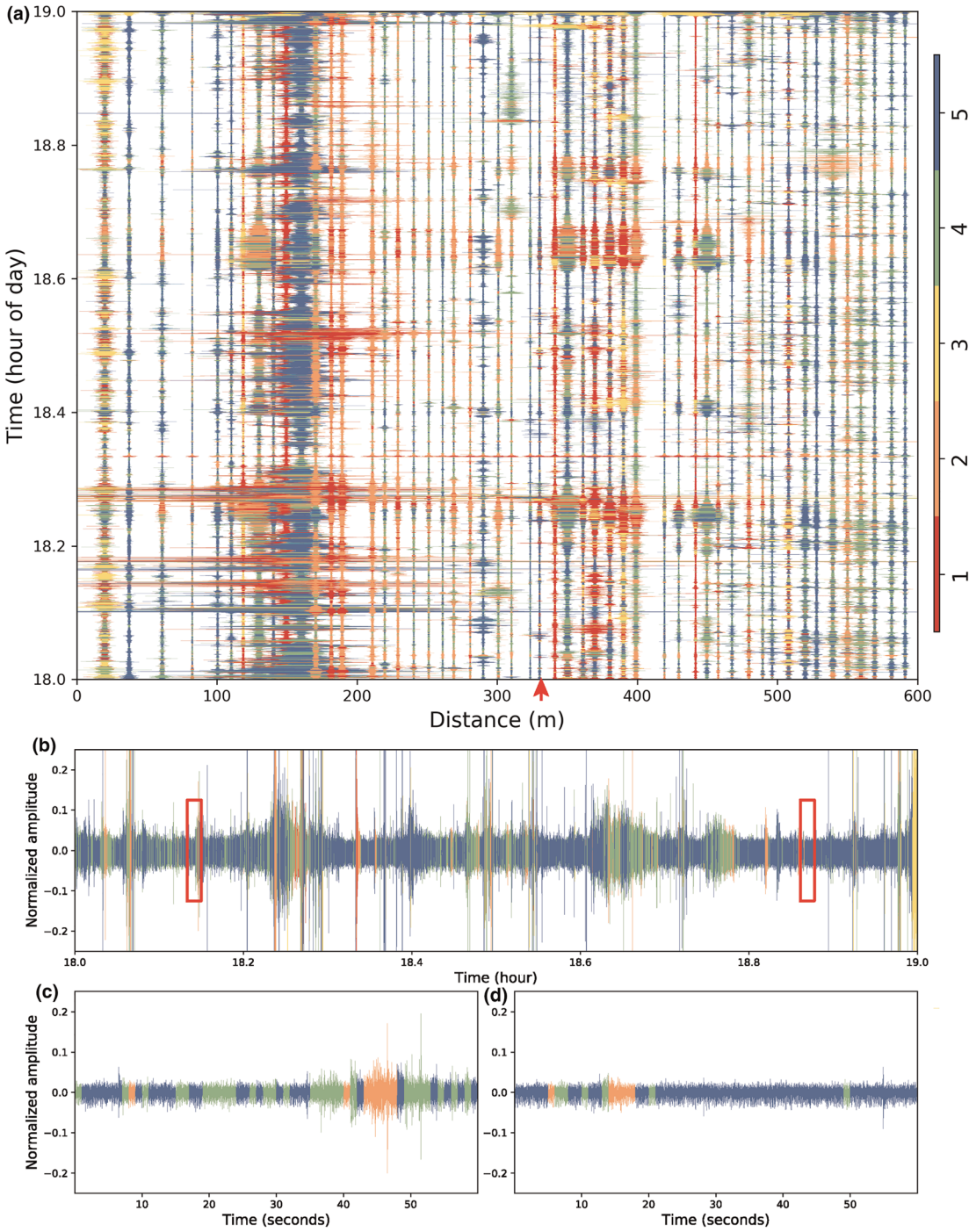

Figure 4. (a) Hour-long waveforms from 18–19 local time on 15 May 2014 for 54 geophones along a transect of the array (Figure S1) with a label assigned every 1 s. The distance is relative to the westernmost geophone and extends about 600 m. (b) Hour-long waveform for the geophone at 325 m (green dot in Figure S1) shown with red arrow in (a). The red box at 18:09 corresponds to panel (c) with a 1 min detailed view of the waveform labeling that contains a small earthquake observed at 42 s. The red box at 18:51 corresponds to panels (d) to show identification of subtle waveform variations.

labeled 2. The signals with Label 4 occur frequently as packets of emergent noise. The 1 min interval starting at 18:09 shows all labels with a microseismic event occurring at 42 s with the *P* wave as Label 4 and the *S* wave as Label 2 (Figure 4c). Emergent waveforms are identified and assigned different labels (Figure 4d), implying that the feature vectors to describe each wavelet to data are adequately describing the key spectral properties of the waveforms.

5. Discussion and Conclusions

The results represent significant progress toward fully classifying continuous seismic waveforms, which consist primarily of different types of noise (Meng et al., 2019). To our knowledge, this is the first attempt to label every 1 s of seismic waveforms, specifically in a dense array, and evaluate the spatiotemporal evolution of transient weak ground motions. Improved detection of microearthquakes requires better characterization of common types of nontectonic sources of ground shaking to properly separate these signals from subsurface processes. We focus on five classes of noise signals that are present at the study area; these classes are likely to exist elsewhere, possibly in a somewhat modified form and with additional classes of noise. The five classes of noise waveforms contain differing spectral properties that provide a separable feature space for an unsupervised clustering model. The study area used in this work is isolated in a rural environment with minimum human activity. Performing a similar analysis near an urban area would potentially produce entirely new classes of nontectonic waveforms. We do not expect the clustering model to perform exactly the same when applied to different regions, but the framework presented to identify different classes of noise is particularly useful for dense seismic arrays designed for microseismicity detection and can help develop metrics for nontectonic signals. The unsupervised approach is ideal to design and build a data set for a supervised classification model. Further refinements in extracting information from the data, for example, with convolutional autoencoders (Mousavi et al., 2019), could provide new information about subtle changes in the noise waveforms to develop a labeled training data set designed for more sophisticated machine learning models and augment libraries of labeled data. Determining the physical source for each cluster is also needed and requires specific experiments to collect known noise signals to incorporate into supervised learning models.

The implemented cluster analysis does not constrain physical labels, but insight is gained from the variety of noise signals in the data. Assigning a physical label to signals originating from wind shaking obstacles above the surface, air traffic, automobiles, trains, and other nontectonic signals will provide important information for designing new training data sets. The different classes of noise often show spatial coherency with neighboring geophones and sometimes form groups throughout the array that provide evidence of weak nontectonic ground motions propagating in the shallow crust (Figure 3 and Movie S1). Signals originating within the array from surface objects producing weak ground motions (Johnson, Meng, et al., 2019) are identifiable in groups of neighboring sensors having the same label for multiple seconds (Figure 2). A potential improvement to isolate the different classes of signals is applying the cluster labels to array processing techniques that consider waveform similarity metrics to isolate groups of sensors identifying the weak ground motion (Cheng et al., 2020). Stacking these signals would reduce uncorrelated noise and potentially identify new processes that could be associated to local sources at the surface or deformation at the subsurface material.

Detecting and locating microseismicity is of great importance to understanding fault mechanical processes. Machine learning techniques are useful to identify earthquakes (e.g., Ross, Meier, & Hauksson, 2018; Zhu & Beroza, 2018), but assembling training data without domain expertise on the various signals contained in seismic waveforms can produce erroneous results such as false detection of microearthquakes and tremor. The dense array design contains information useful to investigate earthquake detections by developing training data that specifically includes nontectonic waveforms with earthquake-like features. Efforts to detect and locate microseismic events in the study area suggest that many impulsive signals are associated with ground motion generated by anthropogenic and atmospheric sources (Gradon et al., 2019; Johnson, Meng, et al., 2019; Meng & Ben-Zion, 2018a, 2018b) that shake surface objects and produce earthquake-like signals. Figure 4b shows multiple impulsive earthquake-like signals that are assigned Label 2 or 4, similar to the known earthquake arrival shown in Figure 4c. No earthquake waveform information is provided in the training data, but the stacked spectra of Noise Label 2 show a peak in amplitude around 15 Hz with a falloff at higher frequencies, similar to the spectra of microseismicity. A local M0.6 earthquake located 44 km from the array is clearly observed and assigned Noise Label 2 for the majority of the array sensors (Figure S3). These observations support the notion that nontectonic emergent and impulsive signals commonly detected contain earthquake-like characteristics. Additional analysis using more than five clusters to partition the data might isolate tectonic signals into their own class, but earthquakes occupy such a small fraction of the waveforms that it would be difficult to independently identify. Instead, we suggest incorporating new classes of noise waveforms into data sets designed for supervised learning techniques, so the model can learn to differentiate nontectonic earthquake- and tremor-like

signals in low signal-to-noise (<1) data from low-magnitude (M < 0) microseismicity. Developing a training data set using labeled noise data and low signal-to-noise earthquakes will provide metrics to quantify the types of noise that are more prone to produce false-positive or false-negative tectonic detections.

Conflict of Interest

The authors declare no competing interests.

Data Availability Statement

Original data used in this study are publicly available through Incorporated Research Institutions for Seismology. The training data are available through the International Federation of Digital Seismograph Networks (https://doi.org/10.7914/SN/9A_2014) (Johnson, 2019).

Acknowledgments
We thank the two anonymous reviewers and editor Gavin Hayes for very constructive comments that helped to improve the manuscript. C. W. J is funded by the National Science Foundation EAR Postdoctoral Fellowship Award 1725344. The study was supported by the U.S. Department of Energy (Awards DE-SC0016520 and DE-SC0016527).

References

Aguiar, A. C., & Beroza, G. C. (2014). PageRank for earthquakes. *Seismological Research Letters*, *85*(2), 344–350. https://doi.org/10.1785/0220130162

Ben-Zion, Y., Vernon, F. L., Ozakin, Y., Zigone, D., Ross, Z. E., Meng, H., et al. (2015). Basic data features and results from a spatially dense seismic array on the San Jacinto fault zone. *Geophysical Journal International*, *202*(1), 370–380. https://doi.org/10.1093/gji/ggv142

Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433), eaau0323. https://doi.org/10.1126/science.aau0323

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Brenguier, F., Boué, P., Ben-Zion, Y., Vernon, F., Johnson, C. W., Mordret, A., et al. (2019). Train traffic as a powerful noise source for monitoring active faults with seismic interferometry. *Geophysical Research Letters*, *46*, 9529–9536. https://doi.org/10.1029/2019GL083438

Chen, Y. (2017). Automatic microseismic event picking via unsupervised machine learning. *Geophysical Journal International*, *212*(1), 88–102. https://doi.org/10.1093/gji/ggx420

Cheng, Y., Ben-Zion, Y., Brenguier, F., Johnson, C. W., Li, Z., Share, P., et al. (2020). An Automated Method for Developing a Catalog of Small Earthquakes Using Data of a Dense Seismic Array and Nearby Stations. *Seismological Research Letters*. https://doi.org/10.1785/0220200134

De Angelis, S., & Bodin, P. (2012). Watching the wind: Seismic data contamination at long periods due to atmospheric pressure-field-induced tilting. *Bulletin of the Seismological Society of America*, *102*(3), 1255–1265. https://doi.org/10.1785/0120110186

Gerstoft, P., & Tanimoto, T. (2007). A year of microseisms in southern California. *Geophysical Research Letters*, *34*, L20304. https://doi.org/10.1029/2007GL031091

Gradon, C., Moreau, L., Roux, P., & Ben-Zion, Y. (2019). Analysis of surface and seismic sources in dense array data with match field processing and Markov chain Monte Carlo sampling. *Geophysical Journal International*, *218*(2), 1044–1056. https://doi.org/10.1093/gji/ggz224

Hammer, C., Beyreuther, M., & Ohrnberger, M. (2012). A seismic-event spotting system for volcano fast-response systems. *Bulletin of the Seismological Society of America*, *102*(3), 948–960. https://doi.org/10.1785/0120110167

Hammer, C., Ohrnberger, M., & Fäh, D. (2013). Classifying seismic waveforms from scratch: A case study in the alpine environment. *Geophysical Journal International*, *192*(1), 425–439. https://doi.org/10.1093/gji/ggs036

Hillers, G., & Ben-Zion, Y. (2011). Seasonal variations of observed noise amplitudes at 2–18 Hz in southern California. *Geophysical Journal International*, *184*(2), 860–868. https://doi.org/10.1111/j.1365-246X.2010.04886.x

Hillers, G., Graham, N., Campillo, M., Kedar, S., Landès, M., & Shapiro, N. (2012). Global oceanic microseism sources as seen by seismic arrays and predicted by wave action models. *Geochemistry, Geophysics, Geosystems*, *13*, Q01021. https://doi.org/10.1029/2011gc003875

Holtzman, B. K., Paté, A., Paisley, J., Waldhauser, F., & Repetto, D. (2018). Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field. *Science Advances*, *4*(5), eaao2929. https://doi.org/10.1126/sciadv.aao2929

Hulbert, C., Rouet-Leduc, B., Johnson, P. A., Ren, C. X., Rivière, J., Bolton, D. C., & Marone, C. (2019). Similarity of fast and slow earthquakes illuminated by machine learning. *Nature Geoscience*, *12*(1), 69–74. https://doi.org/10.1038/s41561-018-0272-8

Inbal, A., Ampuero, J. P., & Clayton, R. W. (2016). Localized seismic deformation in the upper mantle revealed by dense seismic arrays. *Science*, *354*(6308), 88–92. https://doi.org/10.1126/science.aaf1370

Inbal, A., Cristea-Platon, T., Ampuero, J. P., Hillers, G., Agnew, D., & Hough, S. E. (2018). Sources of long-range anthropogenic noise in southern California and implications for tectonic tremor detection. *Bulletin of the Seismological Society of America*. https://doi.org/10.1785/0120180130

Johnson, C. W. (2019). Sage Brush Flats Labled noise data. International Federation of Digital Seismograph Networks, https://doi.org/10.7914/SN/9A_2014

Johnson, C. W., Meng, H., Vernon, F., & Ben-Zion, Y. (2019). Characteristics of ground motion generated by wind interaction with trees, structures, and other surface obstacles. *Journal of Geophysical Research: Solid Earth*, *124*, 8519–8539. https://doi.org/10.1029/2018JB017151

Johnson, C. W., Vernon, F., Nakata, N., & Ben-Zion, Y. (2019). Atmospheric processes modulating noise in Fairfield Nodal 5 Hz geophones. *Seismological Research Letters*. https://doi.org/10.1785/0220180383

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2018). Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, *90*(1), 3–14. https://doi.org/10.1785/0220180259

Li, Z., Peng, Z., Hollis, D., Zhu, L., & McClellan, J. (2018). High-resolution seismic event detection using local similarity for large-N arrays. *Scientific Reports*, *8*(1), 1646. https://doi.org/10.1038/s41598-018-19728-w

Linville, L., Pankow, K., & Draelos, T. (2019). Deep learning models augment analyst decisions for event discrimination. *Geophysical Research Letters*, *46*, 3643–3651. https://doi.org/10.1029/2018gl081119

Marcillo, O. E., & Carmichael, J. (2018). The detection of wind-turbine noise in seismic records. *Seismological Research Letters*, *89*(5), 1826–1837. https://doi.org/10.1785/0220170271

Meng, H., & Ben-Zion, Y. (2018a). Characteristics of airplanes and helicopters recorded by a dense seismic array near Anza California. *Journal of Geophysical Research: Solid Earth*, *123*, 4783–4797. https://doi.org/10.1029/2017JB015240

Meng, H., & Ben-Zion, Y. (2018b). Detection of small earthquakes with dense array data: Example from the San Jacinto fault zone, southern California. *Geophysical Journal International*, *212*(1), 442–457. https://doi.org/10.1093/gji/ggx404

Meng, H., Ben-Zion, Y., & Johnson, C. W. (2019). Detection of random noise and anatomy of continuous seismic waveforms in dense array data near Anza California. *Geophysical Journal International*, *219*(3), 1463–1473. https://doi.org/10.1093/gji/ggz349

Mousavi, S. M., Horton, S. P., Langston, C. A., & Samei, B. (2016). Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression. *Geophysical Journal International*, *207*(1), 29–46. https://doi.org/10.1093/gji/ggw258

Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. (2019). Unsupervised clustering of seismic signals using deep convolutional auto-encoders. *IEEE Geoscience and Remote Sensing Letters*, *16*(11), 1693–1697. https://doi.org/10.1109/LGRS.2019.2909218

Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake detection and location. *Science Advances*, *4*(2), e1700578. https://doi.org/10.1126/sciadv.1700578

Qin, L., Vernon, F. L., Johnson, C. W., & Ben-Zion, Y. (2019). Spectral characteristics of daily to seasonal ground motion at the Piñon Flats Observatory from coherence of seismic data. *Bulletin of the Seismological Society of America*, *109*(5), 1948–1967. https://doi.org/10.1785/0120190070

Ross, Z. E., Meier, M.-A., & Hauksson, E. (2018). *P* wave arrival picking and first-motion polarity determination with deep learning. *Journal of Geophysical Research: Solid Earth*, *123*, 5120–5129. https://doi.org/10.1029/2017JB015251

Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5A), 2894–2901. https://doi.org/10.1785/0120180080

Rouet-Leduc, B., Hulbert, C., & Johnson, P. A. (2019). Continuous chatter of the Cascadia subduction zone revealed by machine learning. *Nature Geoscience*, *12*(1), 75–79. https://doi.org/10.1038/s41561-018-0274-6

Sorrells, G. G., & Goforth, T. T. (1973). Low-frequency earth motion generated by slowly propagating partially organized pressure fields. *Bulletin of the Seismological Society of America*, *63*(5), 1583–1601.

Tanimoto, T., & Wang, J. (2018). Low-frequency seismic noise characteristics from the analysis of co-located seismic and pressure data. *Journal of Geophysical Research: Solid Earth*, *123*, 5853–5885. https://doi.org/10.1029/2018JB015519

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Yoon, C. E., O'Reilly, O., Bergen, K. J., & Beroza, G. C. (2015). Earthquake detection through computationally efficient similarity search. *Science Advances*, *1*(11), e1501057. https://doi.org/10.1126/sciadv.1501057

Zhu, W., & Beroza, G. C. (2018). PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273. https://doi.org/10.1093/gji/ggy423