Unsupervised Clustering of Continuous Ambient Noise Data to Get Higher Signal Quality in Seismic Surveys

Joseph Soloman Thangraj (josephsthangraj8923@gmail.com)
Baylor University
Jay Pulliam
Baylor University
Mrinal K. Sen
The University of Texas at Austin

Research Article

Keywords: Seismic interferometry, selective stacking, EGF convergence, spatio-temporal features, unsupervised learning, K-means

Posted Date: December 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1136687/v1

Abstract

Seismic interferometry has been shown to extract body wave arrivals from ambient noise seismic data. However, surface waves dominate ambient noise data, so cross-correlating and stacking all available data may not succeed in extracting body wave arrivals. A better strategy is to find portions of the data in which body wave energy dominates and to process only those portions. One challenge is that passive seismic recordings comprise huge volumes of data, so identifying portions with strong body-wave energy could be difficult or time-consuming. We use spatio-temporal features, calculated with data recorded by all receivers together, to perform unsupervised clustering. Using data recorded by a dense seismic array in Sweetwater, TX we were able to identify five clusters, representing a subsets of the complete dataset that contain similar features, and extract a 7 km/s body wave arrival from one cluster. This arrival did not emerge when we performed the same cross-correlation and stacking regimen on the entire dataset.

Highlights

- We calculated spatio-temporal features that characterize data recorded by nodal seismic arrays.
- Unsupervised clustering successfully identified noise windows with strong body wave energy.

Non-technical Summary

Passive seismic interferometry methods make use of naturally occurring seismic sources to conduct minimally invasive and inexpensive seismic surveys. The most common way to perform seismic interferometry is to record data continuously in the field and process all acquired data. This approach implicitly assumes that stacking more data will produce better signal-to-noise characteristics in the result. However, one can show that this assumption is often not true; it is easy to produce examples in which processing a subset of the data produces a superior result to processing the entire dataset. However, identifying subsets of the data that have desirable characteristics is challenging. We use unsupervised clustering, a machine learning application, to identify useful portions of the data and find that processing only these portions both enhances the quality of the seismic survey and allows us to reduce the duration of data acquisition. The methods used here can be used for additional applications in which little information is available prior to a survey, including monitoring volcanoes, prospecting for geothermal or hydrocarbon resources, and exploring extraterrestrial terrains.

1 Introduction

Ambient noise seismic interferometry (ANSI) extracts waves travelling between two stations by cross-correlating the time series data recorded by each station and thus estimates an Empirical Green's Function (EGF), the impulse response of the Earth's subsurface between the stations. One station acts as a "virtual source" and the other as a "receiver" (Dantas et al., 2018; Draganov et al., 2013; Nakata et al., 2015; Olivier et al., 2015; Panea et al., 2014; Poli et al., 2012; Quiros et al., 2016; Shapiro & Campillo, 2004; Thangraj et al., 2020a; Thangraj & Pulliam, 2021; Wang et al., 2014; Wapenaar et al., 2010, 2014).

However, for virtual source gathers to be an accurate estimate of the EGF there must be 1) a dense and homogeneous distribution of uncorrelated noise sources and 2) an equipartitioned noise wavefield (Wapenaar et al., 2005). Unfortunately, these conditions are rarely met in real applications. To compensate for a non-uniform distribution of sources, correlations computed for short time windows (e.g., several minutes) are stacked over hours or days to produce a robust estimate of an EGF. To meet the second condition, sources are required to be located on the axis that connects the two stations. For the case of reflected waves, sources are required to be located in stationary phase zones (Draganov et al., 2013) or in a region in which a reflection between two receivers can be retrieved. In this study, cross-correlation results for a single time window is referred to as a single “cross-correlation panel”. Final EGF estimates result from stacking a set of cross-correlation panels.

Stacking correlations over a long period of time was proposed previously as a way to attain the broadest possible distribution of sources (Schimmel et al., 2011; Snieder, 2004) and increase the signal-to-noise ratio of arrivals by suppressing random noise. However, it has been challenging to establish how long correlations need to be stacked, or for passive seismic surveys to acquire data, to ensure that EGFs converge. The convergence rate of noise correlations has been studied by Sabra et al. (2005), who established theoretical relations between variance, recording length, frequency bandwidth and recorded energy. EGF convergence also depends on the choice of window length, processing parameters, and quality of data (Chamarczuk et al., 2021a; Medeiros et al., 2015; Weemstra et al., 2014).

Surface wave studies are more common than studies of body waves in ANSI (Agrawal et al., 2015; Bensen et al., 2007; Chmiel et al., 2019; Nishida et al., 2009; D. A. Quiros et al., 2018; Y. Yang et al., 2007) because surface waves dominate ambient noise sources (and recordings). Consequently, extracting body waves from ANSI is difficult and relatively few successful studies are found in the literature (Chamarczuk et al., 2021b; Cheraghi et al., 2017; Draganov et al., 2009; Panea et al., 2014; Roux et al., 2016; Ruigrok et al., 2011; Thangraj & Pulliam, 2021; Zhan et al., 2010). A particular challenge in retrieving body wave arrivals is the indeterminate nature of seismic sources, and insufficient knowledge of source locations, frequency characteristics and their variation with time. It can be shown that naively stacking all data in ANSI does not always result in the retrieval of body wave arrivals (Girard & Shragge, 2019; Olivier et al., 2015; Thangraj & Pulliam, 2021). Selectively stacking a subset of correlation panels in which body wave energy dominates can extract high signal-to-noise ratio body wave arrivals; selective stacking based on post-correlation metrics has been shown to be successful at extracting body waves (Cheraghi et al., 2017; Olivier et al., 2015; Panea et al., 2014; Safarkhani & Shirzad, 2019; Thangraj & Pulliam, 2021). With the exception of Thangraj & Pulliam (2021), the above-mentioned studies require an estimate of velocity or slowness to discriminate between the cross-correlation panels. There are few pre-correlation automated selective stacking methods (Draganov et al., 2013). Pre-correlation estimates can save significant computation time in computing cross-correlations and also would be less biased by pre-processing routines used in ANSI. In this study we show that it is possible to identify noise windows in which body wave energy dominates with the aid of machine learning methods, specifically unsupervised clustering.

Machine learning methods have been used extensively in seismology. Examples include earthquake phase picking (Soto & Schurr, 2021, Suarez & Given, 2021, Ming et al., 2019), seismic phase association (Ross et al., 2019), seismic wave discrimination (Li, Meier, et al., 2018), earthquake detection (Thomas et al., 2021; S. Yang et al., 2020; Walter et al., 2020), earthquake localization (Shen & Shen, 2021) and event discrimination (Renouard et al., 2021; Linville et al., 2019). However, most of the above-mentioned studies can be classified as “supervised learning” approaches because they require a training dataset and information about the data beforehand. Supervised techniques may therefore work well in the lab, when applied to previously-acquired data, but applications in the field, in real time, would be challenging or impossible. Ambient noise studies are often carried out on datasets for which little information is available beforehand, so “unsupervised clustering” or “unsupervised machine learning”, if feasible, would be better suited to exploring ambient noise datasets. Unsupervised clustering has been used previously to discriminate between local and teleseismic arrivals (Mousavi et al., 2019), to automate microseismic event picking (Chen, 2020), and to detect seismic events (Li, Peng, et al., 2018a). However, few studies have applied and assessed the capabilities of unsupervised learning methods to identify signals recorded by a passive seismic array. Johnson et al. (2020) and Chamarczuk et al. (2020) report on applications of unsupervised learning methods that succeed in identifying classes of signals in seismic data. Identifying noise windows with dominant body wave energy in ANSI studies remains an important and often time-consuming problem. This study assesses the prospects for unsupervised learning methods to save the time and effort required to separate these panels (Draganov et al., 2013; Vidal et al., 2014). We engineer features and tailor unsupervised learning to identify clusters in which body waves arrivals dominate cross-correlation panels, and ultimately selectively stack panels within clusters to reveal body wave arrivals with higher signal-to-noise ratios than is found by stacking all available panels. We also show that machine learning algorithms can help extract information required to label data for subsequent supervised learning studies.

2 Data

The data used in this study were recorded by a mixed array comprised of twenty Nanometrics Trillium Compact Posthole seismometers, five Nanometrics Trillium 120 Posthole seismometers, and 2639 Fairfield Zland nodes from Nodal Seismic deployed at Sweetwater, TX (Barklage et al., 2014; Robert Woodward et al., 2014; Sumy et al., 2015). The Zland nodes were divided into three subarrays, namely, the active source array consisting of 2322 nodes, the backbone array consisting of 292 nodes that spanned the study area for passive seismic data analysis and the outlier array that consisted of 25 nodes spaced ~5 km around the study area. The Nanometrics Trillium stations recorded data with a sampling rate of 200 Hz; Fairfield nodes recorded at 500 Hz. The Sweetwater array collected data from 7th March to 30th April.

We use a 132-station subset to satisfy array processing assumptions, namely that departures from homogeneous structure are negligible. These 132 stations are confined to the 9x9 km area shown in Fig. 1a. For unsupervised clustering we use data from 29-30 March 2014, a total of 48 hours. The array recorded a variety of sources, including airplanes, oil pump jacks, vehicles, Vibroseis trucks, and wind

turbines (Fig. 1b), among others. Possible sources of distant urban noise sources are Snyder airport (38 km west of the Sweetwater array), Anson airport (60 km to the east), and Abilene (74 km to the southeast: 116^0 azimuth). Seismicity near Snyder, TX has been associated with supercritical CO_2 injection (Gan & Frohlich, 2013; Lund Snee & Zoback, 2016) and has been used previously in seismic modeling of the Gulf Coastal Plain (Thangraj et al., 2020).

Figure 2 (a) shows two car signals (Schippkus et al., 2020) from spectrograms of data recorded by nodal station 4X506, which is close to the road. On the other hand, Figure 2 (b) shows the spectrogram of an oil pumpjack, where signals with a periodicity of 6.6 s are observed at ~50 Hz. A plane passing over the array can be seen in Figure 2 (c), where the frequencies shift with time between 100 Hz and 80 Hz. Multiple Vibroseis shots can be observed in Figure 2(d), which is a spectrogram of broadband station 6X542. The number and variety of anthropogenic noise sources, coupled with a lack of seismic activity, makes the identification of seismic arrivals at long offsets challenging. In addition to the noise sources we were able to identify, we also observed a constant signal at ~3 Hz in the spectrogram of uncertain origin. Other studies have identified wind turbines as sources of similar signals (Neuffer et al., 2021). There are certainly turbines in the region, many of which were installed in 2014 or shortly beforehand, but efforts to determine the direction and distance to these sources, described below, point to locations of no known wind turbines, in addition to locations of dense groups of turbines.

3 Methods

Workflow

We apply unsupervised machine learning methods to continuous seismic data to find clusters with high signal-to-noise body wave arrivals. Specifically, we compute the seismic features beamforming, interferometric location (Chamarczuk et al., 2020) and frequency domain amplitudes in an effort to capture characteristics of ambient noise signals that can be interpreted in terms of direction, velocity, location and frequency content of seismic sources. Our preprocessing routine involves removing linear trends from each noise window and downsampling to 200 Hz using Lanczos resampling. To calculate features and perform subsequent clustering, we first divide 48 hours of continuous seismic data into 20 s windows to compute cross-correlation panels. Window length is an important parameter in seismic interferometry studies; shorter noise windows have been shown to extract body wave arrivals more effectively from ambient noise seismic data (Chamarczuk et al., 2021b; Cheraghi et al., 2017; Draganov et al., 2007). Following this step, our workflow consists of three steps (1) feature calculation (2) unsupervised clustering using K-means, and (3) ambient noise processing and stacking of each cluster.

3.1 Feature Generation

We compute features for various durations of time windows (e.g., 20 s, 40 s, 60 s, etc.). *Beamforming* is a "delay and sum" technique for tuning an array of sensors to receive signals from a particular azimuth,

while rejecting signals from other azimuths. Data from array sensors are combined in such a way that waves arriving from user-specified directions sum constructively, while waves arriving from other directions experience destructive interference. The “beamforming” feature is therefore sensitive to azimuth and apparent velocity of sources located around the array. We use Obspy’s beamforming function (Beyreuther et al., 2010), which is a correlation approach (Ruigrok et al., 2016) where cross-correlations of waveforms from different receivers are used as inputs for beamforming. Based on the node spacing of 800 m, the minimum resolvable wavelength would be 800 m if we consider a minimum frequency of 0.5 Hz (Löer et al., 2018). We calculate beamforming features between 0.5 to 10 and limit slownesses to -1.0 s/km to 1.0 s/km to avoid spatial aliasing. We calculate a total of 400 beamforming features and subsequently smooth and downsample to 40 features. The number is further reduced to 19 by imposing a 95% variance threshold on the feature correlation matrix. An example of unscaled beamforming features can be seen in Fig. 3.

Interferometric location is a spatiotemporal tool that scans within the confines of an array to find source locations (Dales, et al., 2017). Interferometric location has been applied successfully to regional earthquake monitoring (Poiata et al., 2016), to studies of earthquake sequences (Baker et al., 2021), and to identifying persistent noise sources for subsurface monitoring (Dales et al, 2017). The inputs are cross-correlations and a constant velocity, which are used to calculate the intensity of sources at each location. The feature for location (x, y), $S(x, y)$ is given by

$$S(x,y) = \sum_{m=1}^{N} \sum_{n=m+1}^{N} C_{mn}(\tau_m - \tau_n), \quad (1)$$

where τ_m and τ_n are time differences between location (x, y) and station *m*, and location (x, y) and station *n*, respectively.

To calculate the interferometric location feature we first divide the area covered by the array into 20x20 grids. After applying a bandpass filter to the data with corner frequencies of 1 Hz and 80 Hz and spectral whitening, we calculate all inter-receiver cross-correlations using CuPy. We then use the cross-correlations to calculate a total of 400 interferometric location features. Subsequently, we smooth and downsample the features to 100 features. An example of unscaled interferometric location features is shown in Fig. 3a.

Frequency domain amplitudes helps us capture variations in frequency content across the array. To calculate this feature, we calculate the amplitude spectrum for each station using a Fast Fourier Transform. We then grid the array area with 4x4 cells and sum the amplitudes of signals of receivers that fall within the same cell. Then we sum the signals within each cell between 1 – 20 Hz, 20 – 40 Hz, 40 – 60 Hz, 60 – 80 Hz, 80 – 100 Hz, to produce a total of 80 features. We choose 4x4 grids because of the density of stations in the Sweetwater array. The frequency bands were chosen, after examining sample sources, to allow different types of sources to be localized in (i.e., separated into) different intervals. An example of frequency domain amplitudes can be seen in Fig. 4.

3.2 Unsupervised clustering using K-means

Clustering is an exploratory data analysis technique often used to identify subgroups in a dataset that share common characteristics. Unsupervised clustering is a way of identifying subgroups without any prior labels in the dataset. K-means is an unsupervised distance-based algorithm that iteratively tries to partition the data into distinct, non-overlapping subgroups (Jin & Han, 2010).

One of the hyperparameters that is required to be initialized for K-means is the number of clusters. To determine the optimal value for number of clusters we use silhouette analysis, which measures the similarity of points within clusters and their dissimilarity from points in other clusters. Supplementary S1 includes more details concerning silhouette analysis.

To evaluate the utility of each feature type we evaluate the K-means clustering performance for each type by performing silhouette analysis on a range of cluster numbers. We find that only frequency domain amplitudes have good clustering performance. From silhouette analysis, we find that five clusters have the maximum average silhouette value, so the optimal number of clusters is five (Fig. 5a). To understand the distribution of noise windows in each cluster, we plot the distribution of silhouette coefficient values (Fig. 5b). From Fig. 5b, we can see that Group Two has the highest silhouette coefficient values, while all groups other than Group Four have silhouette coefficient values greater than 0.35.

3.3 Ambient noise seismic processing and selective stacking

We selectively stack data within each cluster after applying ambient seismic noise processing. We detrend, demean, and downsample the data to 200 Hz for each of the 134 stations. We then bandpass-filter the data with corner frequencies of 1 and 80 Hz, split the results into 20-s noise windows, and apply spectral whitening between 1 Hz and 80 Hz to balance the contribution between different frequencies. We cross-correlate the whitened traces for a maximum lag of 10 s, because limiting lag times saves computational time. Finally, the cross-correlation panels are stacked according to their cluster relationship.

4 Results And Discussion

4.1 K-means clustering results

We retrieve a 7 km/s arrival in the positive lag of Virtual Source Gather (VSG) 6X542 - Group 4 obtained by stacking correlations within Group Four (Fig. 6b), which is observed to ~9 km offsets. Additionally, in the positive lag, we retrieve arrivals with apparent velocities of 4 km/s and 1 km/s. In VSG 6X542 - Group 0 (Fig. 6a), obtained by stacking correlations within Group 0, we observe arrivals with velocities 1 km/s in both positive and negative lags, although they are more dominant in the negative lag. We interpret the 7

km/s arrival to be a body wave due to its lack of any dispersive character; we interpret the 1 km/s arrival to be a surface wave due to its clearly visible dispersive nature (Fig. 6b). It is important to note that the 7 km/s arrival (Fig. 6b) only becomes visible when a bandpass filter with corners at 1 and 5 Hz is applied. Its frequency content and a high velocity suggest a mid-crustal arrival, likely originating from the direction of Oklahoma city. Moreover, selectively stacking correlations in Group 3 (Supplementary Fig. S3c) retrieves no coherent arrivals. Among other clusters, we observe 1 km/s arrivals in the negative lag of Group 1, however, we observe a lot of scattered energy in the negative lag as well (Supplementary Fig. S3a). VSG 6X542 (Fig. 6c), obtained by stacking all data, retrieves only surface wave arrivals in the negative lag, thereby indicating that the dominant energy over the period of two days is contained in surface waves.

Upon selectively stacking within groups clustered by K-means, we find that all groups except for Group 4 have high silhouette values. High silhouette implies good clustering performance, i.e., that noise sources in each cluster have high similarity in location and frequency content. To gain further insight into the clustering performance, we analyze the noise window distribution in each group (Supplementary Fig. S1). Both Group 1 (Supplementary Fig. S1b) and Group 2 (Supplementary Fig. S1c) show that noise windows within this group fall mainly within night times, when noise levels are generally low. Group 1 (Supplementary Fig. S1c) noise windows fall largely within the days and daylight hours when most Vibroseis shots were shot within and around the array. Group 3 (Supplementary Fig. S1d) majorly includes day-time activity on 30th March. Group 4 has an average silhouette value of 0.2, which suggests high variance within the cluster. High variance implies that sources in this cluster were strong and unique, i.e., not repeating in time or location. From interferometric location features we infer that a number of Vibroseis shots were shot within the array on 29th March. Regular moves of the Vibroseis sources to new shot locations, as is routine during surveys, may explain the high variance observed.

4.2 *Identification of sources*

Since the 7 km/s body wave arrival is only seen in the positive lag and at negative offsets, the noise source is very likely located east of the array. Also, the arrival is visible in the Virtual Source Gather only when a 1-5 Hz bandpass filter is applied, which makes it unlikely that Vibroseis shots are the sources. However, the observed surface waves appear to come from all directions in the frequency band 0.5 - 10 Hz (Fig. 7). Consequently, we conclude that higher energy surface wave sources lie to the west of the array, because the surface wave amplitudes are higher in negative lags in VSG 6X542 – Group 0 (Fig. 6a). Also, in beamforming plots at frequencies of 0.5 Hz - 10 Hz (Fig. 7) we see high amplitudes at 290^0 azimuth with slowness values between 0.8 s/km and 0.9 s/km. This direction points toward the airport in Lubbock, TX and we note that airplane landings, at least, are efficient generators of surface waves.

4.3 Interferometric location and Beamforming feature clustering

Although beamforming and interferometric features do not help retrieve body wave arrivals in the case described here, and they suffer from poor clustering performance, useful information is nevertheless conveyed by these features. From the interferometric location features shown in Fig. 2d we can identify Vibroseis sources with varying strength, and note that interferometric location features change significantly as time windows move from low strength to high strength (Fig. 8). This suggests that at least two different Vibroseis trucks were used in the survey, although it should also be noted that the locations of sources are only approximate (Dales, et al., 2017). Given the interferometric location feature's sensitivity to energy sources within the array, it can be used to identify strong, persistent noise sources. Examples of such noise sources are mining, drilling, or construction activities. These, and other, noise sources can be used in ambient seismic monitoring or tomography studies.

The beamforming feature, despite its poor clustering performance in our study, can help identify distant sources of energy. Attenuation and scattering tend to lower the frequency content of seismic signals recorded by distant seismic stations, so we compute beamforming features with the more sensitive broadband instruments, rather than nodes. The beamforming result in Supplementary Fig. S2 shows energy with slowness 0.1 s/km arriving from an azimuth of 290^0 degrees, which could be generated at the Lubbock airport, and energy with slowness 0.2 s/km coming from an azimuth of 30^0 degrees in the frequency band 0.01-0.1 Hz (Supplementary Fig. S2). In the band 0.1-3 Hz (Supplementary Fig. S2), the dominant energy comes from the southeast with slowness 0.1 s/km, which is likely to have been generated by microseisms at the Texas/Gulf of Mexico coast.

Although only the frequency domain amplitude features helped identify panels in which body wave energy dominates, additional features are likely to be sensitive to body wave energy. For example, Snover et al. (2021) use spectrograms as features to perform deep clustering, Ben-Zion et al. (2015) use matched field processing algorithms to locate individual sources of energy, and Cros et al. (2011; Kuperman & Turek (1997) and Li, et al. (2018b) use local similarity to detect events. In addition to these features, we can also use rms energy, kurtosis and skewnesss as features to extract more information from the array. For real-time applications, computational demands of some of these features may render them impractical or infeasible.

5 Conclusions

We investigate the utility of a K-means algorithm to cluster data recorded by the Sweetwater seismic array deployed in north-central Texas in Spring 2014 and we extract body wave arrivals from one of the clusters. Extracting body wave arrivals is often a challenge in ambient noise seismic interferometry because surface waves typically dominate seismic recordings. We cluster the data into five groups based upon variations in the frequency domain amplitude feature, which measures energy distribution in user-specified frequency bands at locations within the array. We retrieve body waves with velocities of 7 km/s and 4 km/s by processing the data within just one high-amplitude cluster; neither of these waves was retrieved by processing all recorded data. We conclude that selectively stacking subsets of ambient noise records produces better results, in the sense that more waves can be identified and measured, than can

be achieved by stacking all available data. Further, it appears that the K-means algorithm may be able to efficiently identify subsets whose stacking leads to better results.

Convergence in ambient noise seismic interferometry has been approached previously as a “brute-force”, time-sequential problem. However, because surface wave energy associated with anthropogenic activities is so abundant, brute stacking may not succeed in extracting body wave arrivals. In such cases, identifying and refining measures that identify time windows of the data in which body wave arrivals dominate could be an effective approach. Machine learning algorithms offer increasingly accessible and simple-to-implement strategies for identifying data subsets by clustering portions of the data that share common characteristics with the help of spatiotemporal features computed for the whole array. In the case described here, one such shared feature led to constructive stacking that revealed previously unseen body waves. Perhaps most importantly, we retrieved body wave arrivals with no prior information about the subsurface.

Unsupervised clustering is easy to automate, to perform periodic processing of data recorded by nodal arrays. Additionally, unsupervised clustering can efficiently determine optimal ambient seismic noise processing parameters. Consequently, these techniques can also be used for monitoring applications in which recurring sources, if available and if identified, can be used to investigate changes in subsurface structure.

Declarations

Acknowledgments

We thank the IRIS Data Management Center for archiving and disseminating the data used in this study (Woodward et al., 2014). Data available from: The IRIS Data Management Center (IRISDMC): http://service.iris.edu/fdsnws/dataselect/1/.

Author contribution statement

J.S.T wrote the manuscript, analyzed the data, developed the method, wrote codes to implement it, made all the figures and interpreted the results.

J.P. helped with editing the manuscript, guiding and formulating the problem statement and interpretation of results..

M.K.S helped with editing the manuscript, machine learning validation, guiding and formulating the problem statement.

All authors reviewed the manuscript

References

1. Agrawal, M., Pulliam, J., Sen, M. K., & Gurrola, H. (2015). Lithospheric structure of the Texas-Gulf of Mexico passive margin from surface wave dispersion and migrated Ps receiver functions. *Geochemistry, Geophysics, Geosystems*, *16*(7), 2221–2239. https://doi.org/10.1002/2015GC005803
2. Baker, B., Holt, M. M., Pankow, K. L., Koper, K. D., & Farrell, J. (2021). Monitoring the 2020 Magna, Utah, Earthquake Sequence with Nodal Seismometers and Machine Learning. *Seismological Research Letters*, *92*(2A), 787–801. https://doi.org/10.1785/0220200316
3. Barklage, M., Hollis, D., Gridley, J. M., Woodward, R., & Spriggs, N. (2014). A large-N mixed sensor active+ passive seismic array near Sweetwater, TX. In *AGU Fall Meeting Abstracts* (Vol. 2014, pp. S11D-4368).
4. Bensen, G. D., Ritzwoller, M. H., Barmin, M. P., Levshin, A. L., Lin, F., Moschetti, M. P., et al. (2007). Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophysical Journal International*, *169*(3), 1239–1260. https://doi.org/10.1111/j.1365-246X.2007.03374.x
5. 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
6. Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010). ObsPy: A Python Toolbox for Seismology. *Seismological Research Letters*, *81*(3), 530–533. https://doi.org/10.1785/gssrl.81.3.530
7. Chamarczuk, M., Nishitsuji, Y., Malinowski, M., & Draganov, D. (2020). Unsupervised Learning Used in Automatic Detection and Classification of Ambient-Noise Recordings from a Large-N Array. *Seismological Research Letters*, *91*(1), 370–389. https://doi.org/10.1785/0220190063
8. Chamarczuk, M., Malinowski, M., & Draganov, D. (2021a). 2D body-wave seismic interferometry as a tool for reconnaissance studies and optimization of passive reflection seismic surveys in hardrock environments. *Journal of Applied Geophysics*, *187*, 104288.
9. Chamarczuk, M., Malinowski, M., & Draganov, D. (2021b). 2D body-wave seismic interferometry as a tool for reconnaissance studies and optimization of passive reflection seismic surveys in hardrock environments. *Journal of Applied Geophysics*, *187*, 104288. https://doi.org/10.1016/j.jappgeo.2021.104288
10. Chen, Y. (2020). Automatic microseismic event picking via unsupervised machine learning. *Geophysical Journal International*, *222*(3), 1750–1764. https://doi.org/10.1093/gji/ggaa186
11. Cheraghi, S., White, D. J., Draganov, D., Bellefleur, G., Craven, J. A., & Roberts, B. (2017). Passive seismic reflection interferometry: A case study from the Aquistore CO_2 storage site, Saskatchewan, Canada. *GEOPHYSICS*, *82*(3), B79–B93. https://doi.org/10.1190/geo2016-0370.1
12. Chmiel, M., Mordret, A., Boué, P., Brenguier, F., Lecocq, T., Courbis, R., et al. (2019). Ambient noise multimode Rayleigh and Love wave tomography to determine the shear velocity structure above the Groningen gas field. *Geophysical Journal International*. https://doi.org/10.1093/gji/ggz237

13. Cros, E., Roux, P., Vandemeulebrouck, J., & Kedar, S. (2011). Locating hydrothermal acoustic sources at Old Faithful Geyser using matched field processing. *Geophysical Journal International*, *187*(1), 385–393.
14. Dales, P., Audet, P., Olivier, G., & Mercier, J.-P. (2017). Interferometric methods for spatio temporal seismic monitoring in underground mines. *Geophysical Journal International*, *210*(2), 731–742. https://doi.org/10.1093/gji/ggx189
15. Dales, P., Audet, P., & Olivier, G. (2017). Seismic Interferometry Using Persistent Noise Sources for Temporal Subsurface Monitoring. *Geophysical Research Letters*, *44*(21), 10,863-10,870. https://doi.org/10.1002/2017GL075342
16. Dantas, O. A. B., do Nascimento, A. F., & Schimmel, M. (2018). Retrieval of Body-Wave Reflections Using Ambient Noise Interferometry Using a Small-Scale Experiment. *Pure and Applied Geophysics*, *175*(6), 2009–2022. https://doi.org/10.1007/s00024-018-1794-0
17. Draganov, D., Wapenaar, K., Mulder, W., Singer, J., & Verdel, A. (2007). Retrieval of reflections from seismic background-noise measurements. *Geophysical Research Letters*, *34*(4), L04305. https://doi.org/10.1029/2006GL028735
18. Draganov, D., Campman, X., Thorbecke, J., Verdel, A., & Wapenaar, K. (2009). Reflection images from ambient seismic noise. *GEOPHYSICS*, *74*(5), A63–A67. https://doi.org/10.1190/1.3193529
19. Draganov, D., Campman, X., Thorbecke, J., Verdel, A., & Wapenaar, K. (2013). Seismic exploration-scale velocities and structure from ambient seismic noise (>1 Hz): AMBIENT-NOISE EXPLORATION-SCALE IMAGING. *Journal of Geophysical Research: Solid Earth*, *118*(8), 4345–4360. https://doi.org/10.1002/jgrb.50339
20. Gan, W., & Frohlich, C. (2013). Gas injection may have triggered earthquakes in the Cogdell oil field, Texas. *Proceedings of the National Academy of Sciences*, *110*(47), 18786–18791. https://doi.org/10.1073/pnas.1311316110
21. Girard, A. J., & Shragge, J. (2019). Automated processing strategies for ambient seismic data. *Geophysical Prospecting*. https://doi.org/10.1111/1365-2478.12794
22. Jin, X., & Han, J. (2010). K-Means Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 563–564). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_425
23. Kuperman, W. A., & Turek, G. (1997). Matched field acoustics. *Mechanical Systems and Signal Processing*, *11*(1), 141–148.
24. Li, Z., Peng, Z., Hollis, D., Zhu, L., & McClellan, J. (2018a). High-resolution seismic event detection using local similarity for Large-N arrays. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-19728-w
25. Li, Z., Peng, Z., Hollis, D., Zhu, L., & McClellan, J. (2018b). 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

26. Li, Z., Meier, M.-A., Hauksson, E., Zhan, Z., & Andrews, J. (2018). Machine Learning Seismic Wave Discrimination: Application to Earthquake Early Warning. *Geophysical Research Letters*, *45*(10), 4773–4779. https://doi.org/10.1029/2018GL077870
27. Linville, L., Pankow, K., & Draelos, T. (2019). Deep Learning Models Augment Analyst Decisions for Event Discrimination. *Geophysical Research Letters*, *46*(7), 3643–3651. https://doi.org/10.1029/2018GL081119
28. Löer, K., Riahi, N., & Saenger, E. H. (2018). Three-component ambient noise beamforming in the Parkfield area. *Geophysical Journal International*, *213*(3), 1478–1491. https://doi.org/10.1093/gji/ggy058
29. Lund Snee, J.-E., & Zoback, M. D. (2016). State of stress in Texas: Implications for induced seismicity: STATE OF STRESS IN TEXAS. *Geophysical Research Letters*, *43*(19), 10,208-10,214. https://doi.org/10.1002/2016GL070974
30. Medeiros, W. E., Schimmel, M., & do Nascimento, A. F. (2015). How much averaging is necessary to cancel out cross-terms in noise correlation studies? *Geophysical Journal International*, *203*(2), 1096–1100. https://doi.org/10.1093/gji/ggv336
31. Ming Z., Shi C., LiHua F., & Yuen D. A. (2019). Earthquake phase arrival auto-picking based on U-shaped convolutional neural network. *Chinese Journal of Geophysics*, *62*(8), 3034–3042. https://doi.org/10.6038/cjg2019M0495
32. Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. (2019). Unsupervised Clustering of Seismic Signals Using Deep Convolutional Autoencoders. *IEEE Geoscience and Remote Sensing Letters*, *16*(11), 1693–1697. https://doi.org/10.1109/LGRS.2019.2909218
33. Nakata, N., Chang, J. P., Lawrence, J. F., & Boué, P. (2015). Body wave extraction and tomography at Long Beach, California, with ambient-noise interferometry. *Journal of Geophysical Research: Solid Earth*, *120*(2), 1159–1173. https://doi.org/10.1002/2015JB011870
34. Neuffer, T., Kremers, S., Meckbach, P., & Mistler, M. (2021). Characterization of the seismic wave field radiated by a wind turbine. *Journal of Seismology*, *25*(3), 825–844. https://doi.org/10.1007/s10950-021-10003-6
35. Nishida, K., Montagner, J.-P., & Kawakatsu, H. (2009). Global Surface Wave Tomography Using Seismic Hum. *Science*, *326*(5949), 112–112. https://doi.org/10.1126/science.1176389
36. Olivier, G., Brenguier, F., Campillo, M., Lynch, R., & Roux, P. (2015). Body-wave reconstruction from ambient seismic noise correlations in an underground mine. *GEOPHYSICS*, *80*(3), KS11–KS25. https://doi.org/10.1190/geo2014-0299.1
37. Panea, I., Draganov, D., Vidal, C. A., & Mocanu, V. (2014). Retrieval of reflections from ambient noise recorded in the Mizil area, Romania. *Geophysics*, *79*(3), Q31–Q42. https://doi.org/10.1190/geo2013-0292.1
38. Poiata, N., Satriano, C., Vilotte, J.-P., Bernard, P., & Obara, K. (2016). Multiband array detection and location of seismic sources recorded by dense seismic networks. *Geophysical Journal International*, *205*(3), 1548–1573.

39. Poli, P., Campillo, M., Pedersen, H., & LAPNET Working Group. (2012). Body-Wave Imaging of Earth's Mantle Discontinuities from Ambient Seismic Noise. *Science*, *338*(6110), 1063–1065. https://doi.org/10.1126/science.1228194
40. Quiros, D. A., Pulliam, J., Barman, D., Polanco Rivera, E., & Huerfano, V. (2018). Ambient Noise Tomography Images Accreted Terranes and Igneous Provinces in Hispaniola and Puerto Rico. *Geophysical Research Letters*, *45*(22), 12,293-12,301. https://doi.org/10.1029/2018GL080095
41. Quiros, Diego A., Brown, L. D., & Kim, D. (2016). Seismic interferometry of railroad induced ground motions: body and surface wave imaging. *Geophysical Journal International*, *205*(1), 301–313. https://doi.org/10.1093/gji/ggw033
42. Renouard, A., Maggi, A., Grunberg, M., Doubre, C., & Hibert, C. (2021). Toward False Event Detection and Quarry Blast versus Earthquake Discrimination in an Operational Setting Using Semiautomated Machine Learning. *Seismological Research Letters*. https://doi.org/10.1785/0220200305
43. Robert Woodward, Dan Hollis, & Neil Spriggs. (2014). Sweetwater Array. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/XB_2014
44. Ross, Z. E., Yue, Y., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2019). PhaseLink: A Deep Learning Approach to Seismic Phase Association. *Journal of Geophysical Research: Solid Earth*, *124*(1), 856–869. https://doi.org/10.1029/2018JB016674
45. Roux, P., Moreau, L., Lecointre, A., Hillers, G., Campillo, M., Ben-Zion, Y., et al. (2016). A methodological approach towards high-resolution surface wave imaging of the San Jacinto Fault Zone using ambient-noise recordings at a spatially dense array. *Geophysical Journal International*, *206*(2), 980–992. https://doi.org/10.1093/gji/ggw193
46. Ruigrok, E., Campman, X., & Wapenaar, K. (2011). Extraction of P-wave reflections from microseisms. *Comptes Rendus Geoscience*, *343*(8–9), 512–525. https://doi.org/10.1016/j.crte.2011.02.006
47. Ruigrok, E., Gibbons, S., & Wapenaar, K. (2016). Cross-correlation beamforming. *Journal of Seismology*. https://doi.org/10.1007/s10950-016-9612-6
48. Sabra, K. G., Roux, P., & Kuperman, W. A. (2005). Emergence rate of the time-domain Green's function from the ambient noise cross-correlation function. *The Journal of the Acoustical Society of America*, *118*(6), 3524–3531. https://doi.org/10.1121/1.2109059
49. Safarkhani, M., & Shirzad, T. (2019). Improving C1 and C3 empirical Green's functions from ambient seismic noise in NW Iran using RMS ratio stacking method. *Journal of Seismology*, *23*(4), 787–799. https://doi.org/10.1007/s10950-019-09834-1
50. Schimmel, M., Stutzmann, E., & Gallart, J. (2011). Using instantaneous phase coherence for signal extraction from ambient noise data at a local to a global scale: Ambient noise signal extraction. *Geophysical Journal International*, *184*(1), 494–506. https://doi.org/10.1111/j.1365-246X.2010.04861.x
51. Schippkus, S., Garden, M., & Bokelmann, G. (2020). Characteristics of the Ambient Seismic Field on a Large-N Seismic Array in the Vienna Basin. *Seismological Research Letters*, *91*(5), 2803–2816. https://doi.org/10.1785/0220200153

52. Shapiro, N. M., & Campillo, M. (2004). Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise: CORRELATIONS OF THE SEISMIC NOISE. *Geophysical Research Letters*, *31*(7), n/a-n/a. https://doi.org/10.1029/2004GL019491
53. Shen, H., & Shen, Y. (2021). Array-Based Convolutional Neural Networks for Automatic Detection and 4D Localization of Earthquakes in Hawai'i. *Seismological Research Letters*, *92*(5), 2961–2971. https://doi.org/10.1785/0220200419
54. Snieder, R. (2004). Extracting the Green's function from the correlation of coda waves: A derivation based on stationary phase. *Physical Review E*, *69*(4). https://doi.org/10.1103/PhysRevE.69.046610
55. Snover, D., Johnson, C. W., Bianco, M. J., & Gerstoft, P. (2021). Deep Clustering to Identify Sources of Urban Seismic Noise in Long Beach, California. *Seismological Research Letters*, *92*(2A), 1011–1022. https://doi.org/10.1785/0220200164
56. Soto, H., & Schurr, B. (2021). DeepPhasePick: a method for detecting and picking seismic phases from local earthquakes based on highly optimized convolutional and recurrent deep neural networks. *Geophysical Journal International*, *227*(2), 1268–1294. https://doi.org/10.1093/gji/ggab266
57. Suarez, A. L. A., & Given, P. (2021). Regional Earthquake Seismic Phase Picking Using Convolutional and Recurrent Neural Networks.
58. Sumy, D. F., Woodward, R., Barklage, M., Hollis, D., Spriggs, N., Gridley, J. M., & Parker, T. (2015). Sweetwater, texas large n experiment. In *AGU Fall Meeting Abstracts* (Vol. 2015, pp. S41A-2703).
59. Thangraj, J. S., Quiros, D. A., & Pulliam, J. (2020a). Using Ambient Noise Seismic Interferometry and Local and Teleseismic Earthquakes to Determine Crustal Thickness and Moho Structure of the Northwestern Gulf of Mexico Margin. *Geochemistry, Geophysics, Geosystems*, *21*(7), e2020GC008970. https://doi.org/10.1029/2020GC008970
60. Thangraj, Joseph Soloman, & Pulliam, J. (2021). Towards real-time assessment of convergence criteria in seismic interferometry: Selective stacking of cross-correlations at the San Emidio geothermal field. *Journal of Applied Geophysics*, *193*, 104426. https://doi.org/10.1016/j.jappgeo.2021.104426
61. Thomas, A. M., Inbal, A., Searcy, J., Shelly, D. R., & Bürgmann, R. (2021). Identification of Low-Frequency Earthquakes on the San Andreas Fault With Deep Learning. *Geophysical Research Letters*, *48*(13), e2021GL093157. https://doi.org/10.1029/2021GL093157
62. Vidal, C. A., Draganov, D., van der Neut, J., Drijkoningen, G., & Wapenaar, K. (2014). Retrieval of reflections from ambient noise using illumination diagnosis. *Geophysical Journal International*, *198*(3), 1572–1584. https://doi.org/10.1093/gji/ggu164
63. Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., & Woelfel, I. (2020). easyQuake: Putting Machine Learning to Work for Your Regional Seismic Network or Local Earthquake Study. *Seismological Research Letters*, *92*(1), 555–563. https://doi.org/10.1785/0220200226
64. Wang, K., Luo, Y., Zhao, K., & Zhang, L. (2014). Body waves revealed by spatial stacking on long-term cross-correlation of ambient noise. *Journal of Earth Science*, *25*(6), 977–984. https://doi.org/10.1007/s12583-014-0495-6

65. Wapenaar, K., Fokkema, J., & Snieder, R. (2005). Retrieving the Green's function in an open system by cross correlation: A comparison of approaches (L). *The Journal of the Acoustical Society of America*, *118*(5), 2783–2786. https://doi.org/10.1121/1.2046847
66. Wapenaar, K., Draganov, D., Snieder, R., Campman, X., & Verdel, A. (2010). Tutorial on seismic interferometry: Part 1 — Basic principles and applications. *GEOPHYSICS*, *75*(5), 75A195-75A209. https://doi.org/10.1190/1.3457445
67. Wapenaar, K., Thorbecke, J., van der Neut, J., Broggini, F., Slob, E., & Snieder, R. (2014). Green's function retrieval from reflection data, in absence of a receiver at the virtual source position. *The Journal of the Acoustical Society of America*, *135*(5), 2847–2861.
68. Weemstra, C., Westra, W., Snieder, R., & Boschi, L. (2014). On estimating attenuation from the amplitude of the spectrally whitened ambient seismic field. *Geophysical Journal International*, *197*(3), 1770–1788. https://doi.org/10.1093/gji/ggu088
69. Yang, S., Hu, J., Zhang, H., & Liu, G. (2020). Simultaneous Earthquake Detection on Multiple Stations via a Convolutional Neural Network. *Seismological Research Letters*, *92*(1), 246–260. https://doi.org/10.1785/0220200137
70. Yang, Y., Ritzwoller, M. H., Levshin, A. L., & Shapiro, N. M. (2007). Ambient noise Rayleigh wave tomography across Europe. *Geophysical Journal International*, *168*(1), 259–274. https://doi.org/10.1111/j.1365-246X.2006.03203.x
71. Zhan, Z., Ni, S., Helmberger, D. V., & Clayton, R. W. (2010). Retrieval of Moho-reflected shear wave arrivals from ambient seismic noise: SmS reflections from seismic noise. *Geophysical Journal International*, no-no. https://doi.org/10.1111/j.1365-246X.2010.04625.x

Figures

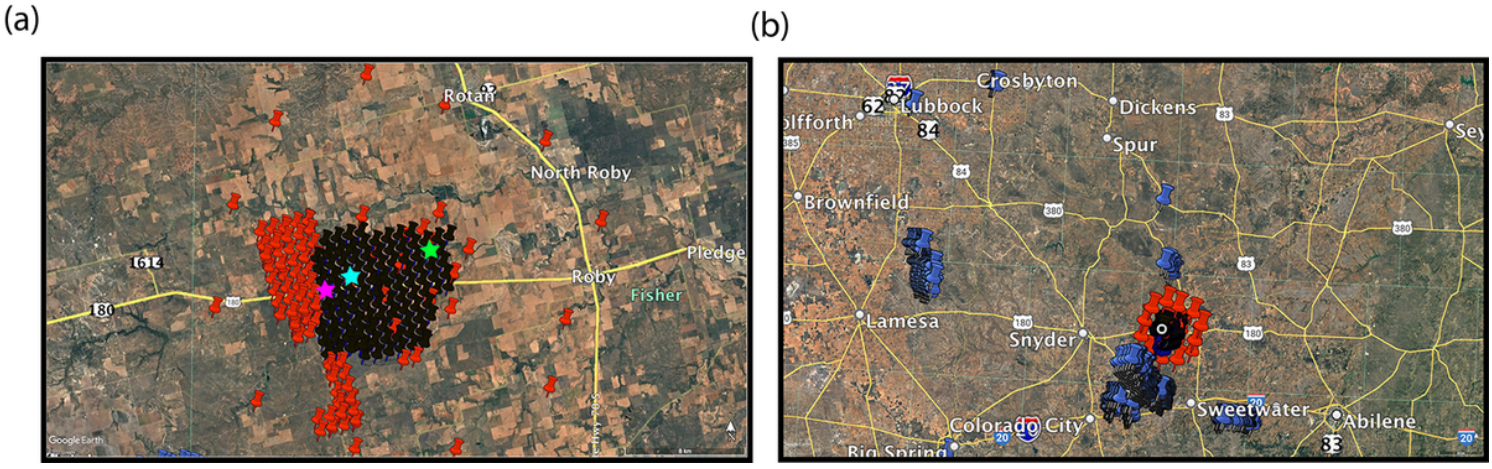

Figure 1

(a) Map of nodal seismic instruments (spaced closely towards the center of the array) and broadband seismic instruments (distributed in a circles). Nodal seismic instruments used in this study are indicated

by black pins. The green star marks the virtual source 6X542; the cyan star marks the location of station 4X506. See text for further details. (b) Map of wind turbines (blue pins) and cities that are possible sources of noise.

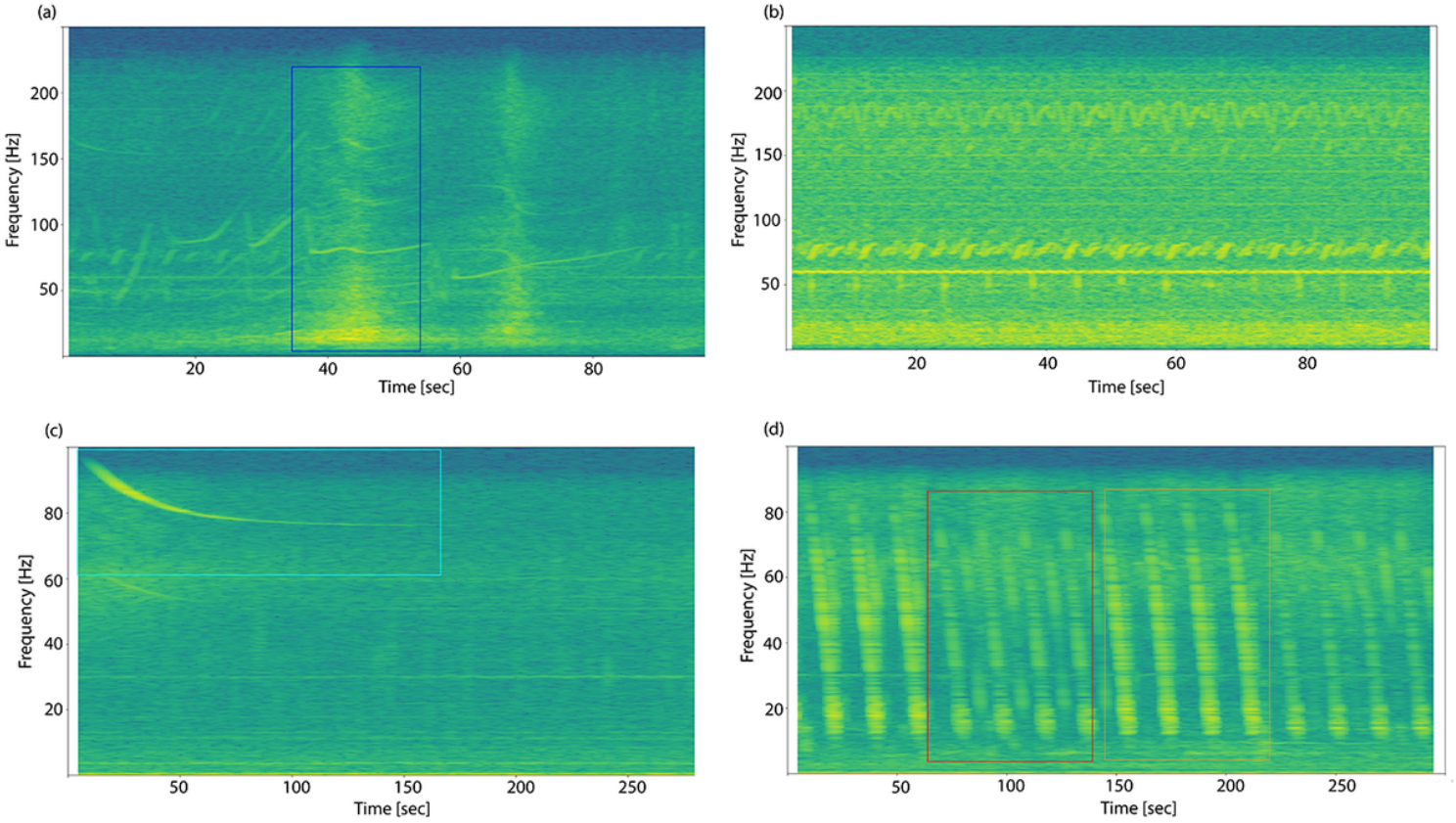

Figure 2

Spectrograms of (a) station 4X506 showing a car signal (blue rectangle), (b) an oil pump jack signal with a periodicity of 6.6s, (c) an airplane flying over broadband station C0104 (cyan rectangle), and (d) Vibroseis sweeps recorded by broadband station 6X542 (rectangle).

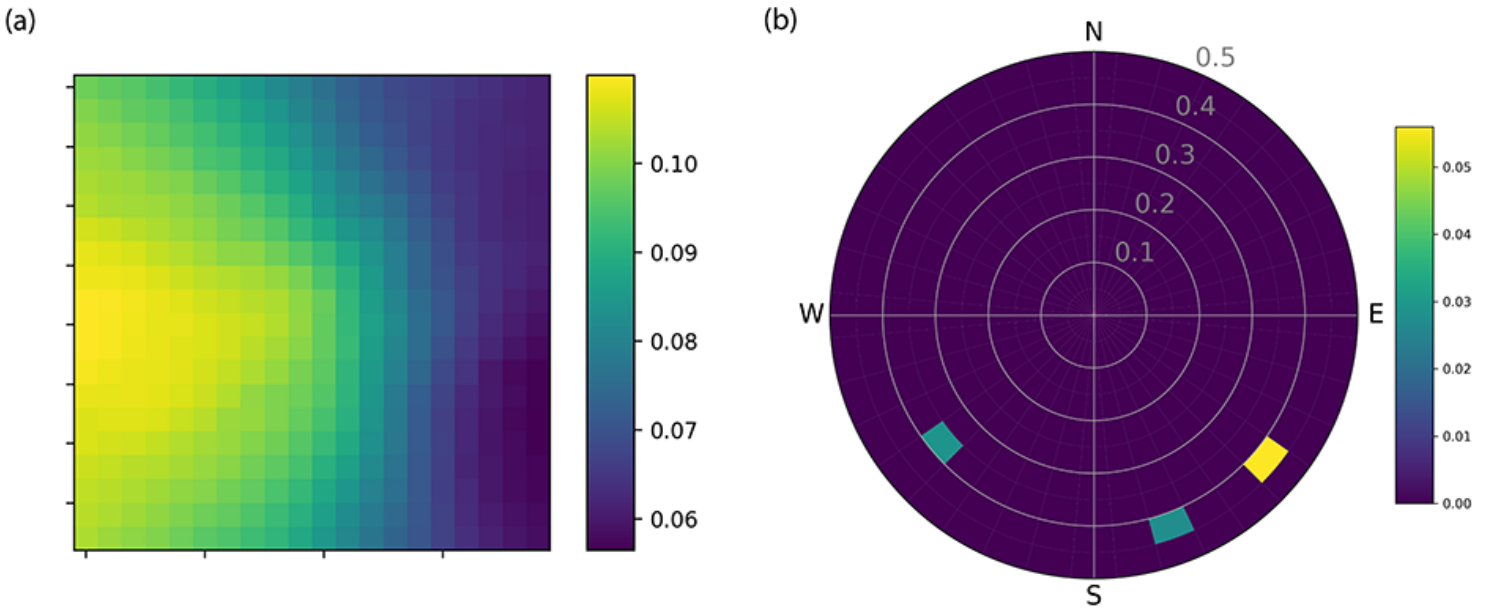

Figure 3

(a) Interferometric location features computed for a 20 x 20 grid and 20 s time window. (b) Beamforming features with 36 x 10 dimensionality computed for a 20 s time window.

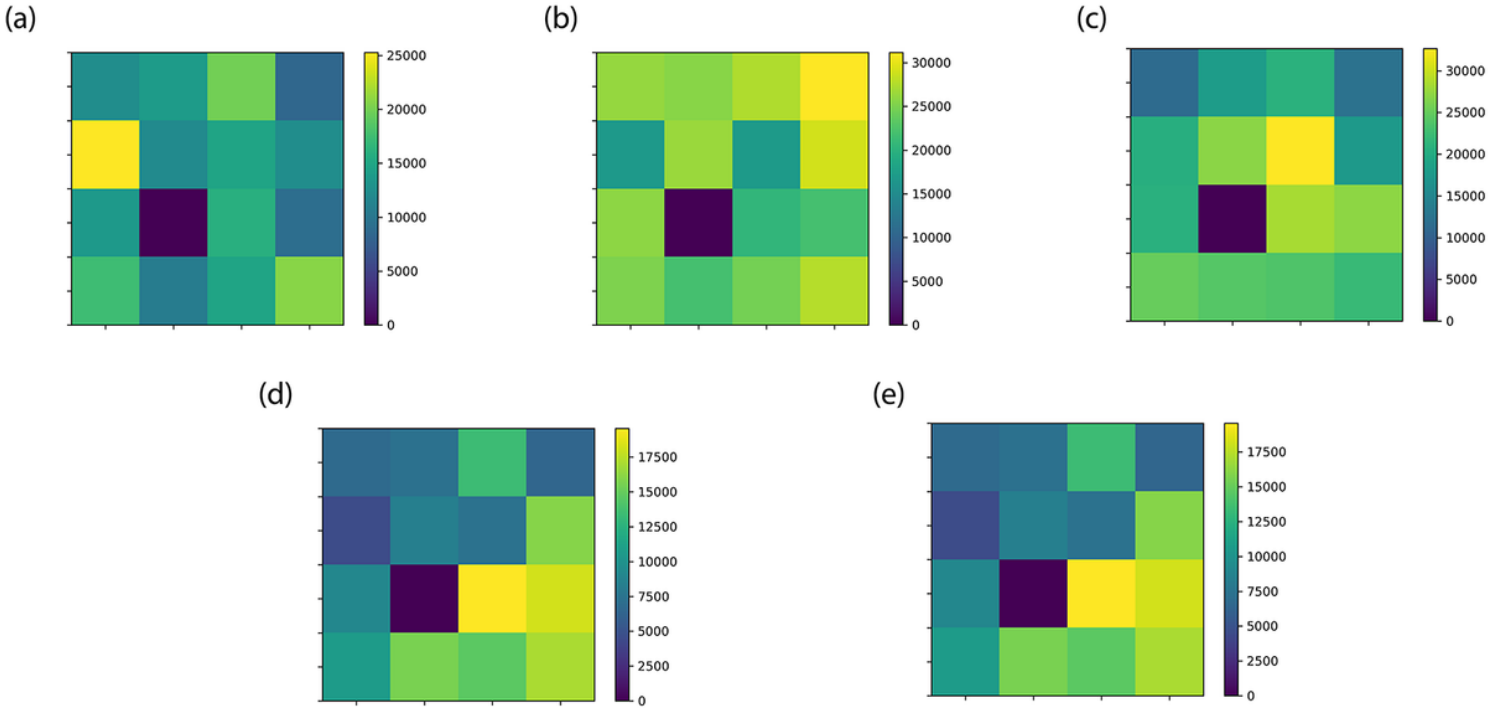

Figure 4

Frequency domain amplitude features for a 20 s time window for (a) 1-20 Hz (b) 20-40 Hz (c) 40-60 Hz (d) 60-80 Hz (e) 80-100 Hz.

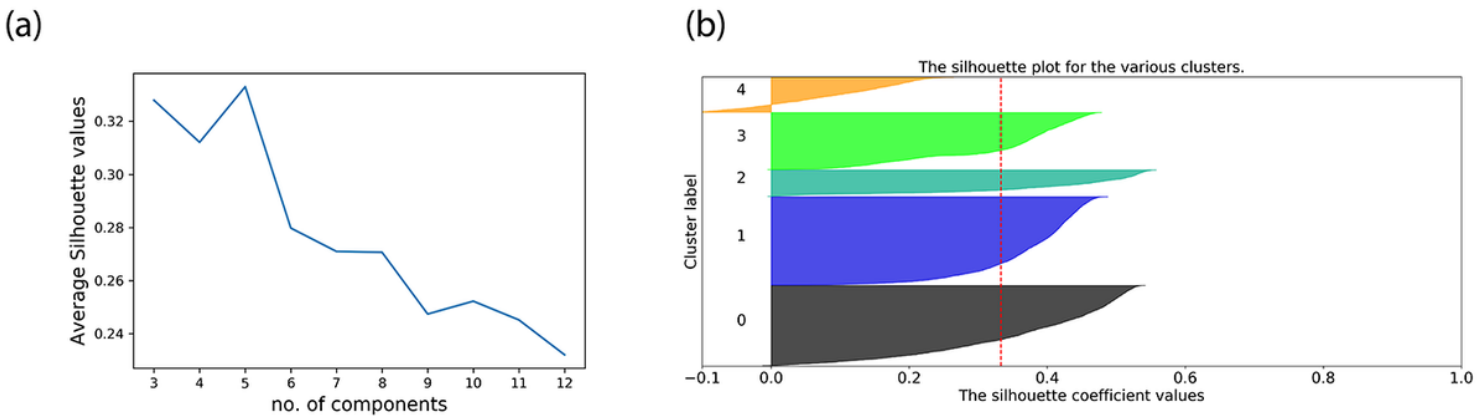

Figure 5

(a) Average silhouette variation with various number of clusters. (b) Cluster membership for five clusters.

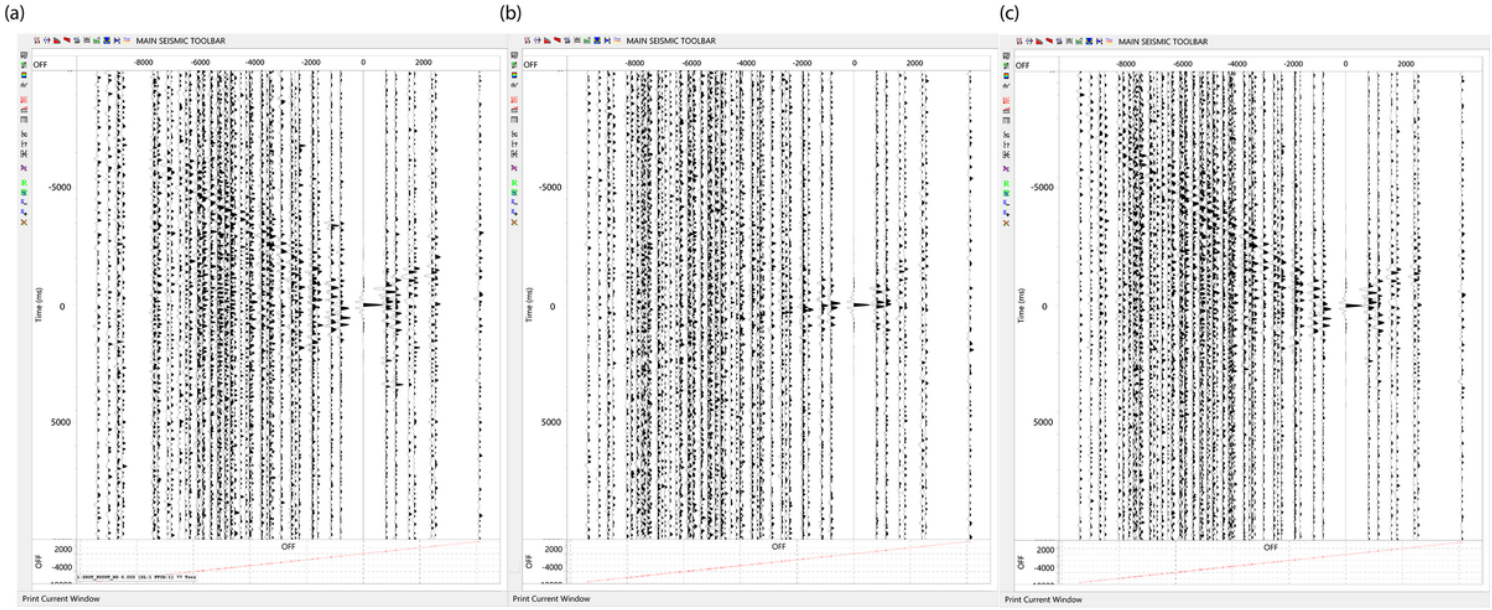

Figure 6

VSG 6X542 obtained by stacking within cluster (a) Group 0 (b) Group 4. (c) VSG 6X542 obtained by stacking all data.

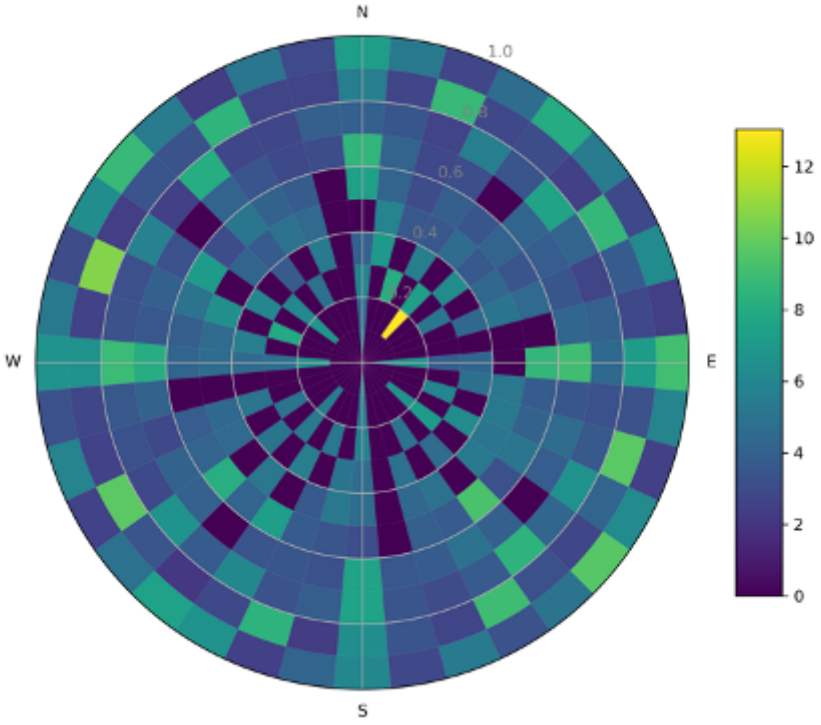

Figure 7

Sum of all beamforming features from 29th March to 30th March.

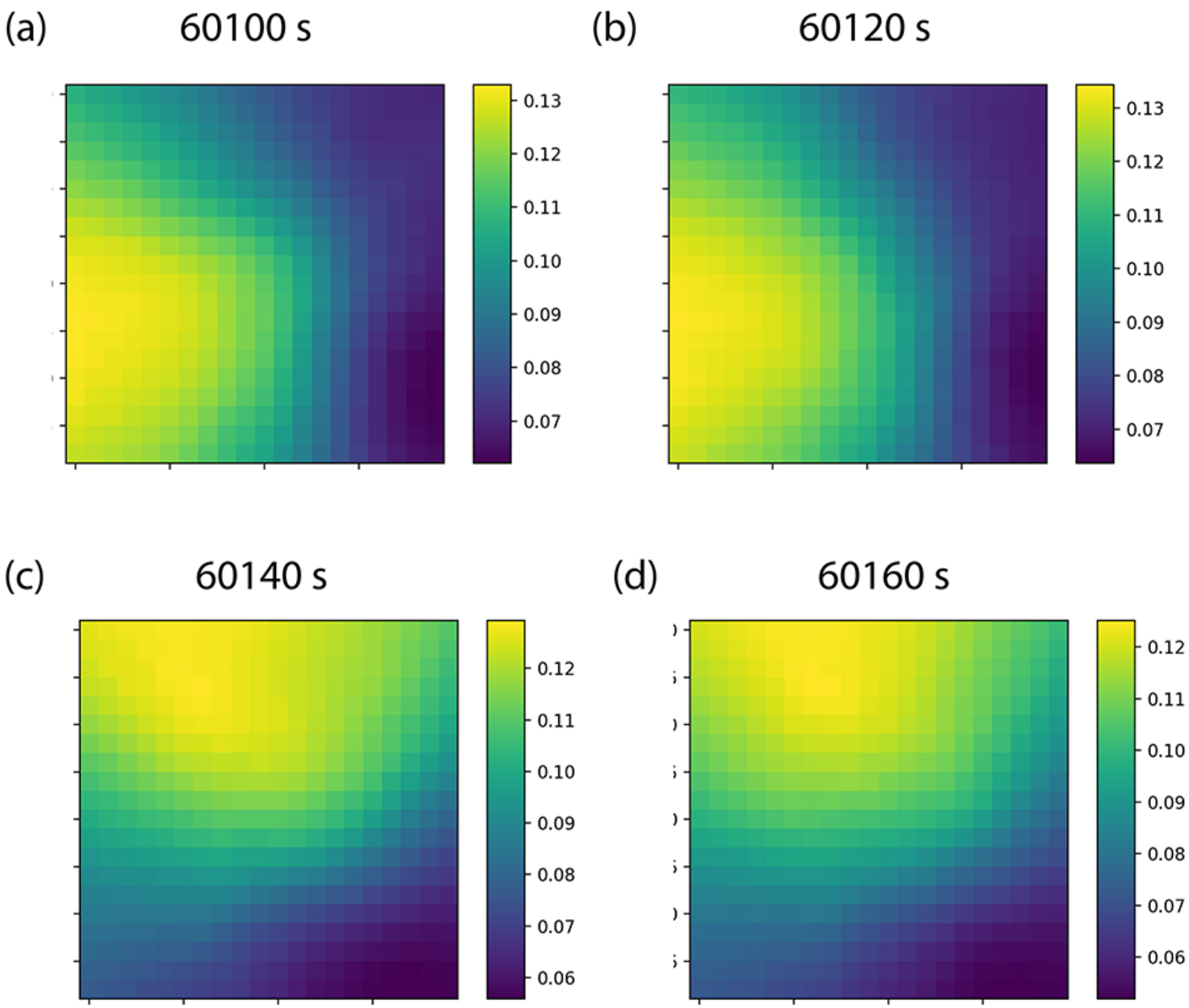

Figure 8

Two Vibroseis trucks operating on 29th March, 2014. The first Vibroseis truck is operating at (a) 60100 s and (b) 60120 s. The second Vibroseis truck is operating at (c) 60140 s and (d) 60160 s.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementryinfounsupervisedclusteringrevised1.docx