Automatic noise exploration in urban areas

Fantine Huot, Yinbin Ma, Robert Cieplicki, Eileen Martin, and Biondo Biondi

ABSTRACT

Near-surface imaging with ambient noise has grown into an increasingly common tool over the past decade thanks to the virtual source method. However, if non-ideal noise sources are present, experts must manually analyze the noise to look for any issues they suspect, then design filters to remove these non-ideal noises. Up until now, the deployment and maintenance of the receiver array was the primary cost, but this is changing with advancements in Distributed Acoustic Sensing (DAS), an emerging technology that repurposes a fiber optic cable as a series of strain sensors. On the Stanford campus we have shown that we can record seismic waves with fiber optic cables sitting loosely in existing telecommunications conduits. As we look forward at the possibility of easily plugging into unused fibers in telecom bundles on-demand, it is clear that manual selection of non-ideal noise sources is the next bottleneck. Herein we show a variety of methods, mixing traditional signal processing and machine learning, to automatically assist geophysicists in analyzing the ambient noise recorded and selecting non-ideal noises. We demonstrate that we can identify different types of noise using clustering algorithms and that template matching can be used for detecting specific events.

INTRODUCTION

By measuring the speed of seismic waves propagating in the Earth's near-surface, we can image the top tens to hundreds of meters of the subsurface, with deeper features being resolved by lower frequencies. These seismic velocity images can be interpreted to evaluate earthquake or landslide risk, to detect permafrost, to find sinkholes or tunnels, or to track near-surface changes related to drilling activities. Additionally, in cases of complex near-surface conditions, resolving this complexity is a prerequisite to obtaining a high-quality image of the deeper subsurface.

By cross-correlating noise recorded at a selected receiver with noise recorded by all other receivers in an array, we can extract signals mimicking an active seismic survey with a source at the selected receiver, called its virtual source response function (Lobkis and Weaver, 2001; Lin et al., 2008; Wapenaar et al., 2010a,b). Thus, when active sources are too costly or logistically prohibitive, passive seismic can be a good option for near-surface imaging. However, the theory is limited by the assumption of homogeneous uncorrelated sources (Wapenaar et al., 2010a). Non-ideal sources

can cause artifacts in extracted velocities, but with careful processing experts can overcome these limitations (Bensen et al., 2007; Daskalakis et al., 2016; Zhan et al., 2013), even for anthropogenic sources occurring at higher frequencies (Girard and Shragge, 2016; Martin* et al., 2015; Martin et al., 2016; Nakata et al., 2011, 2015; Zeng et al., 2016). Acquisition cost is a further issue: nearly all of the past ambient noise studies on dense arrays have been temporary arrays which were labor-intensive to install and maintain (Lin et al., 2008; Nakata et al., 2011, 2015; Zeng et al., 2016).

Distributed acoustic sensing (DAS) is a new acquisition technology being increasingly adopted in the energy industry for microseismic monitoring (Webster et al., 2013) and time-lapse seismic surveys (Daley et al., 2013; Daley* et al., 2014; Mateeva et al., 2014; Bakku, 2015). DAS probes a fiber-optic cable with a laser interrogator unit (IU) to repurpose that fiber as a series of strain sensors. Motivated by avoiding the maintenance cost of node arrays, there have been several recent ambient noise experiments using fiber optics buried in shallow trenches directly coupled to the ground (Ajo-Franklin* et al., 2015; Martin* et al., 2015; Martin et al., 2016; Zeng et al., 2016). In fact, we can sacrifice some ground-to-sensor coupling in favor of easier installation by running fibers in existing telecommunications conduits (Martin et al., 2017). By running fibers in existing conduits, or even plugging into unused fibers in existing telecommunications bundles, easy, on-demand, repeatable seismic studies (even in urban areas) will soon be a reality.

With ambient noise data becoming increasingly easy to record, data volumes are increasing, and we can only extract their full value if we further automate the processing workflow. Herein we introduce some tools and metrics, mixing traditional signal processing and machine learning, to automatically assist geophysicists in analyzing the ambient noise recorded and selecting non-ideal noises.

We present this methodology in the context of a case study: a figure-eight-shaped array of 2.4 km of fiber optics lying loosely in existing telecommunications conduits underneath the Stanford University campus. This particular experiment is ideal because it demonstrates the wide variety of issues coming up on the horizon as we push for broader use of ambient noise. The array detects a wide variety of seismic noise sources that do not conform to the ideals of existing ambient noise theory: it sits in a seismically active region, 20 km from the Pacific ocean, 7 km from the San Francisco bay, with major highways on either side, a variety of roads with differing levels of traffic near the fiber, regular quarry blasts within 15 km, plumbing and HVAC systems throughout the site, multiple construction sites near the array, and foot and bicycle traffic throughout. With over 600 sensors continuously recording 50 samples per second since September 2016, manual inspection of most data is infeasible, making automation tools critical to extracting subsurface information from the data.

After providing a brief overview of the recorded data and the pre-processing steps, we present how unsupervised learning algorithms can help us identify the main types of seismic noise present in the data. We then focus on the high amplitude noise and use template matching to detect specific events.

THE STANFORD DAS ARRAY

The fiber optic cable is deployed in Stanford's telecommunication tunnels in a double loop pattern (Figure 1). Every 8 meters of cable acts as a receiver and records vibration at a sampling rate of 50 Hz, creating a data matrix of 600 channels distributed in space. As a result, we obtain neighboring time series which conveniently lend themselves to image processing (Figure 2). As the fiber optic cable is sitting loosely in existing telecommunications conduits, it is very poorly coupled with the ground. We balance the amplitudes by bandpass filtering (0.5 and 20 Hz) followed by automatic gain control (AGC). As a result, more weight is given to less noisy channels and we obtain balanced amplitudes across the whole array (Figure 2c).

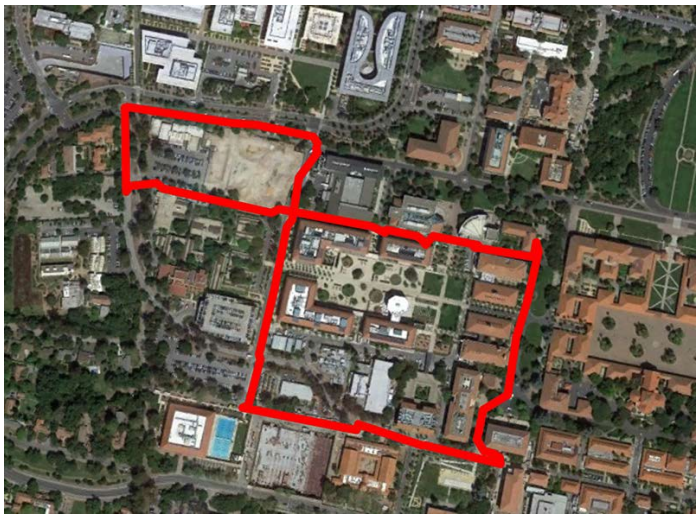


Figure 1: The fiber optic cable (red line) is deployed under Stanford campus in the telecommunication infrastructure tunnels. It forms a North-East and a South-West loop. [**NR**]

AUTOMATICALLY IDENTIFYING DIFFERENT TYPES OF SEISMIC NOISE

Statistical learning has become a highly explored field in many scientific areas as well as marketing, finance, and other business disciplines. In recent years, new and improved software packages have significantly eased the implementation burden for many statistical learning methods, providing scientists and practitioners with complete toolkits for training, testing, and deploying models with well-documented examples for all these tasks (Abadi et al., 2016; Pedregosa et al., 2011; Collobert et al., 2002; Jia et al., 2014; James et al., 2013). As a consequence, machine learning techniques are increasingly adopted to improve accuracy and speed up processing for seismic applications (Fisher et al., 2016; Yoon et al., 2015)

In order to identify the main types of noise in the DAS data, we performed unsu-

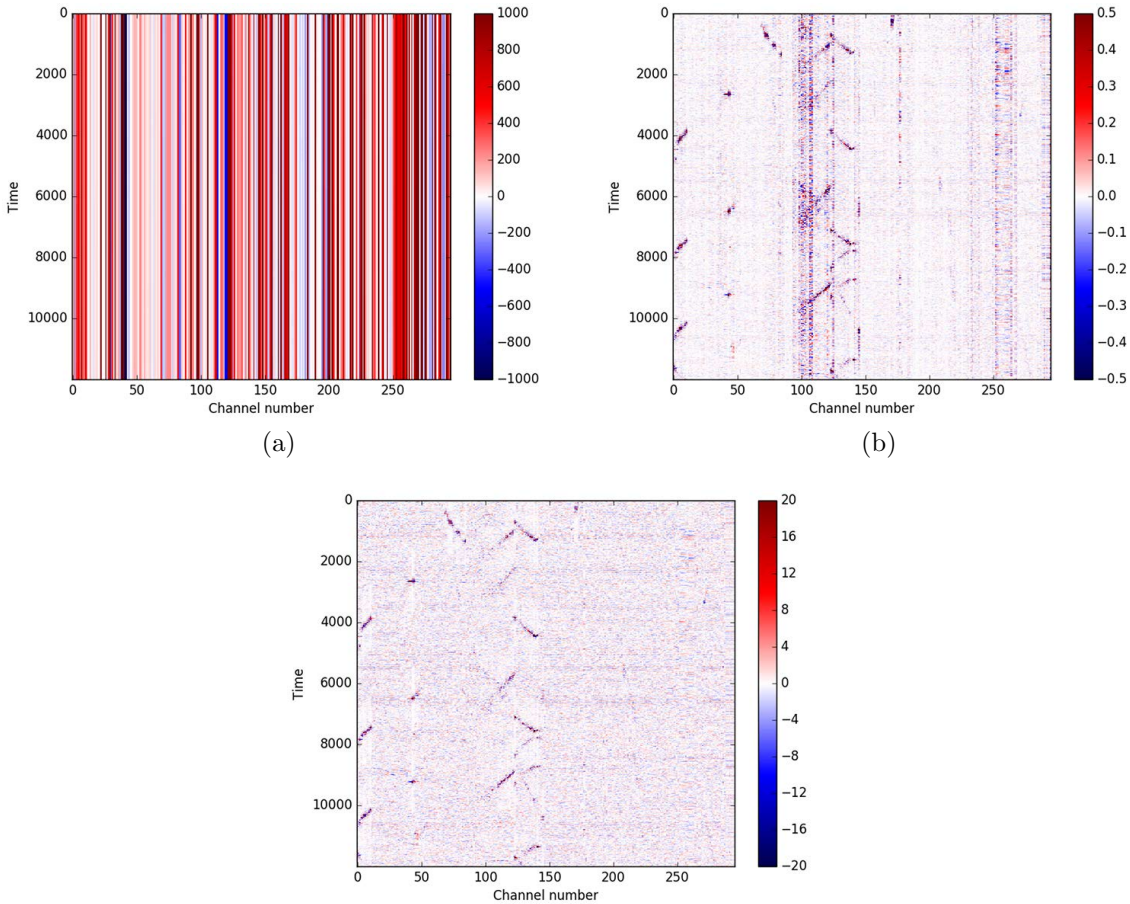


Figure 2: (a) 4 min of raw data (50 samples per second); (b) data after bandpass filter between 0.5 and 20 Hz; (c) data after AGC with a sliding window of 40 seconds. [**ER**]

pervised learning on a week-long subset of the data, encompassing the daily variations in the noise field. We applied continuous wavelet transforms (CWT) both along the time axis and the spatial axis to extract features. The CWT is commonly used in pattern recognition, as it has the ability to decompose complex patterns into elementary forms (Mallat, 1999; Sinha et al., 2005). CWT measures the similarity between a signal and an analyzing wavelet by comparing the input signal to shifted and compressed or stretched versions of the wavelet. For an input signal f, and a mother wavelet ψ, the CWT can be expressed as follows:

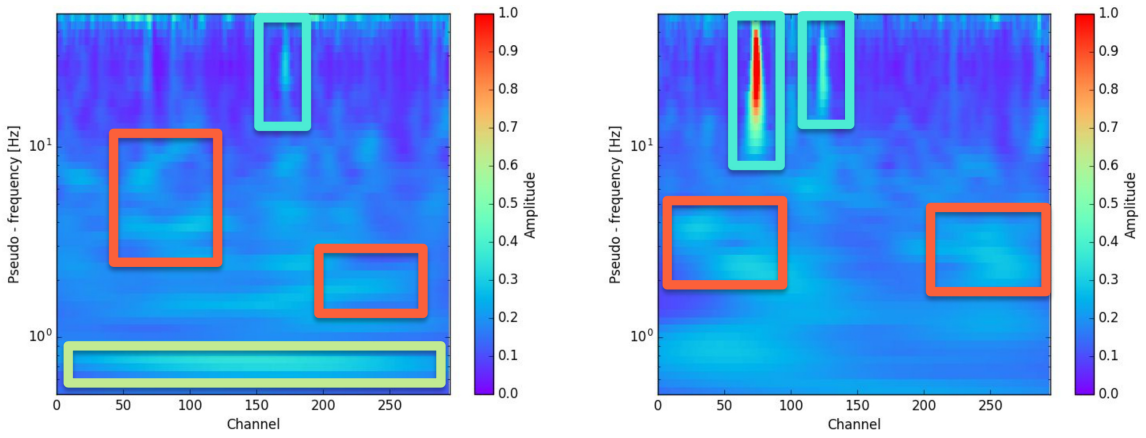
$$C(a,\, b\,;f(t),\, \psi(t)) = \int_{-\infty}^{\infty} f(t)\, \frac{1}{\sqrt{a}}\, \psi^* \left(\frac{t-b}{a} \right)\, dt\ ,$$

where a is the scaling factor, b the time shift, and $*$ denotes the complex conjugate.

We used the Morlet wavelet as the analyzing function and subsequently took the amplitude of the resulting complex numbers. For each sample we obtained 100 CWT scale factors, half of them resulting from decomposing the data along the time axis, the other half from decomposing along the spatial dimension. At this stage, we subsampled the features by averaging the CWT scales over windows of 0.5 s, as any seismic events that last less than 0.5 s would be hard to interpret physically. This resulted in a data matrix of 100 features for over a million samples.

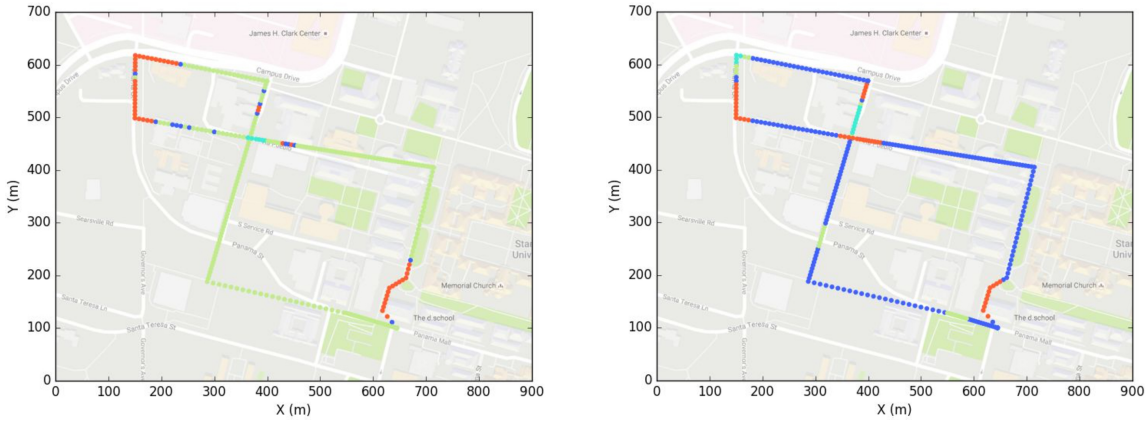
Common clustering methods for wavelet domain time series include K-means, agglomerative clustering and self-organizing maps (Liao, 2005; Köhler et al., 2009). Herein we use K-means for faster computation. Figure 4 presents the detected clusters over time, projected over the array's geometry at two different time stamps. We ran the algorithm for different numbers of clusters and qualitatively settled for 4 clusters, each capturing a different portion of the CWT scales, and therefore, easier to interpret. A higher number of clusters merely yielded subdivisions of these 4 main clusters. After examining each cluster's CWT and frequency components, the 4 main types of identified noise can be described as follows (Figure 4): Dark blue represents the incoherent ambient noise field; Light green corresponds to laser noise, responsible for disruptions with large spatial extent over the array; Cyan corresponds to traffic noise. Note how the clustering algorithm captured moving sources such as cars without being provided any information related to the geometry of the array; Orange indicates other sources of coherent noise of medium amplitude, mostly localized on the North-West part of the fiber optic array. These coherent noise sources could be linked to remote traffic noise (from cars that are not running exactly above the fiber), or construction noise (the area circled by the North-West loop is under construction).

While the exact sources of these main types of noise might be difficult to interpret, these results have an important consequence for seismic processing. By automatically detecting different types of coherent noise, we can selectively filter them out in CWT space, and apply inverse CWT to map the data back to the time domain to obtain ambient field data without non-ideal noise sources.



(a) (b)

Figure 3: Continuous wavelet transform (CWT) scales obtained over the spatial dimension at two different time stamps. The colored rectangles highlight the different types of seismic signature that were detected by the clusters. [**CR**]



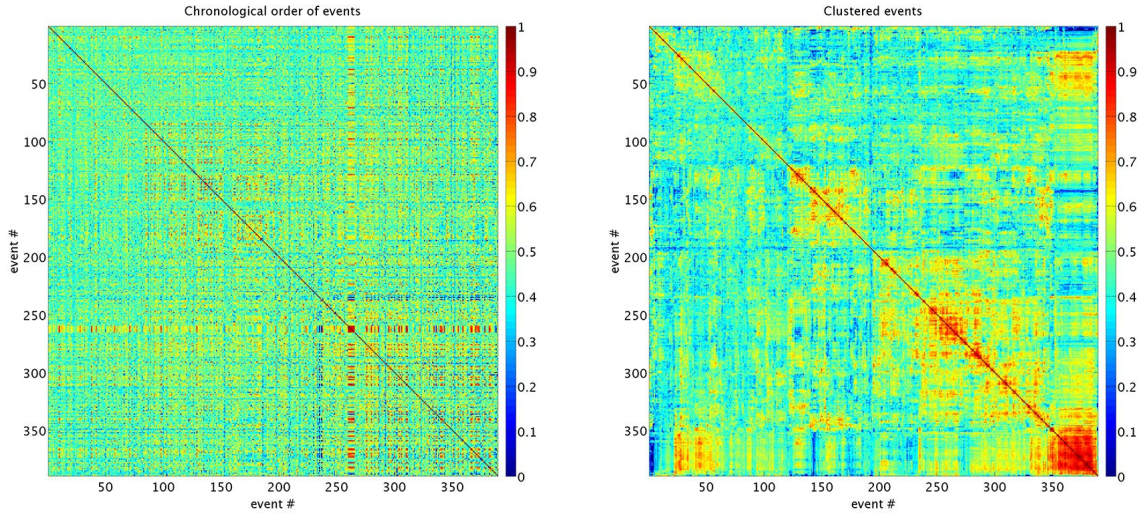
(a) (b)

Figure 4: Clustering results projected over the array's geometry at two different time stamps. The time stamps correspond to those from Figure 3. Dark blue represents the incoherent ambient noise field; Light green corresponds to laser noise, responsible for disruptions with large spatial extent over the array; Cyan corresponds to traffic noise. Note how the clustering algorithm captured moving sources such as cars without being provided any information related to the geometry of the array; Orange indicates other sources of coherent noise of medium amplitude, mostly localized on the North-West part of the fiber optic array. These coherent noise sources could be linked to remote traffic noise (from cars that are not running exactly above the fiber), or construction noise (the area circled by the North-West loop is under construction). [**CR**]

CLUSTERING OF HIGH-AMPLITUDE EVENTS

In this section, we narrow down our study to the high-amplitude events. We performed event detection by computing an edge density map using the gradient (Prewitt's operator) followed by convolution with a stencil of ones (200 samples long in time and 20 channels wide in space). We arbitrarily chose a conservative threshold, at the discretion of the interpreter, to select events that stand above the background noise.

Having selected potential events, we computed dot products between all possible pairs which gave us a measure of similarity used for clustering (Figure 5). We chose hierarchical trees to obtain clusters with varying number of events (James et al., 2013) (Figure 5). Although we assumed an input number of clusters between 10-20, there would generally be only a couple of clusters which captured most of the data variability. Figure 6 represents the four largest. Note visual consistency for inter-cluster events manifesting in the slope and the distribution of energy.



(a) (b)

Figure 5: (a) Similarity matrix between pre-selected high-energy events; (b) events from (a) sorted by a hierarchical clustering algorithm. Although we assumed an input number of clusters between 10-20, a small number of clusters already captures most of the data variability. [**CR**]

EVENT DETECTION USING TEMPLATE MATCHING

We then attempted to identify active sources automatically using template matching. As it is not feasible to manually label all the data, we selectively hand-picked several classes of templates: one class contained events along a road which is only open to school buses, as shown in Figure 7 (a), (b) and (c); another class contained events near a road crossing where the fiber optic cable changes direction, as shown in Figure 7

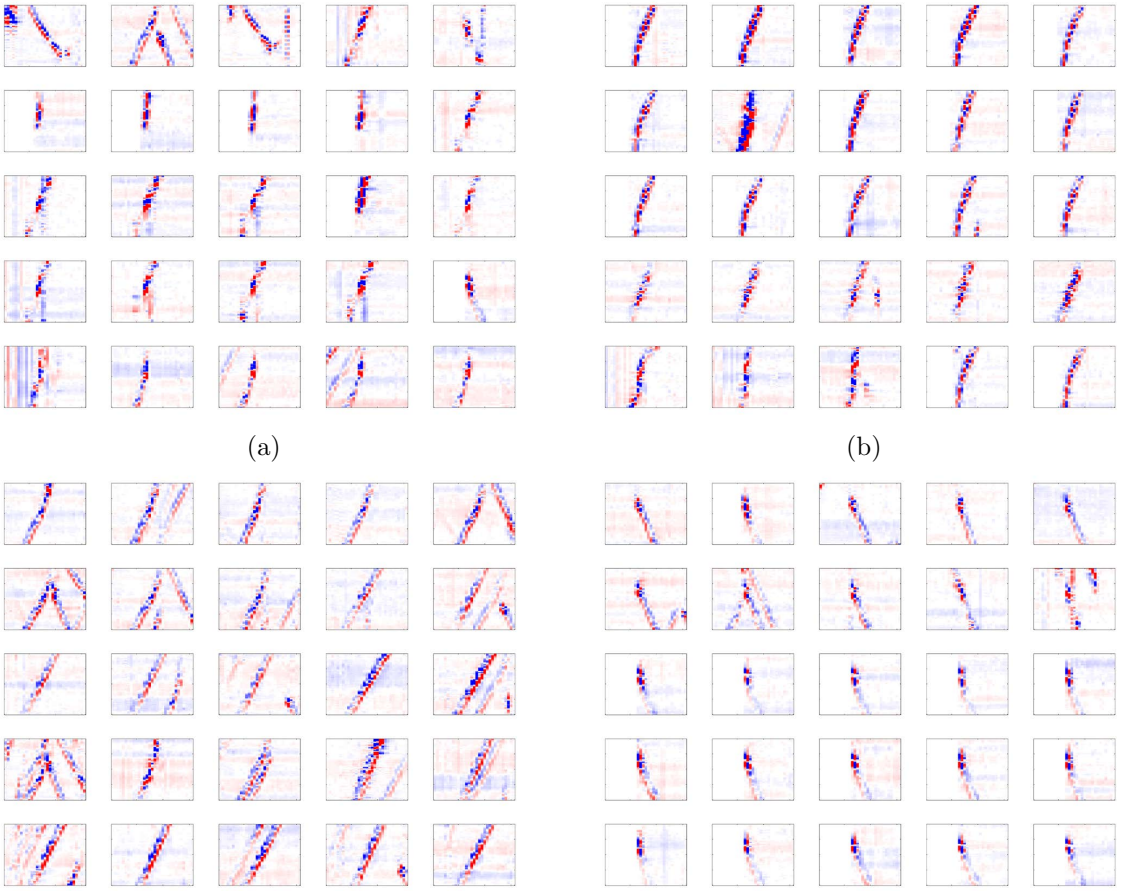


Figure 6: Example 25 events corresponding to each of the four largest identified clusters. Every subimage in this figure is 300 samples long in time (vertical axis) and 30 samples wide in space (horizontal axis). Note visual consistency for inter-cluster events manifesting in the slope and the distribution of energy. [**CR**]

(d), (e) and (f). The size of the templates is 100 seconds by 30 channels (240 m). We performed template matching using normalized cross-correlation (Lewis, 1995). The data is assigned to a particular class when the normalized cross-correlation with the templates from that class exceeds a certain threshold (0.5 in this study).

The detected sources are shown in the blue rectangles in Figure 8. The official schedule of the corresponding bus line is shown in the red rectangles. We see that using only a limited number of templates, we can detect events that approximately match the official bus schedule. The time discrepancy can be due to the bus not running exactly on time. The detected events that do not correspond to the schedule could be empty buses or delivery trucks.

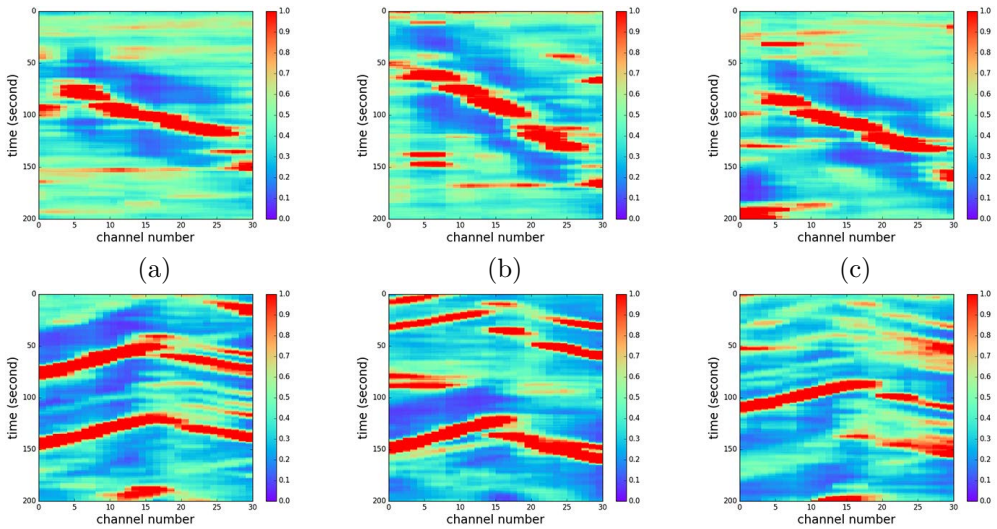
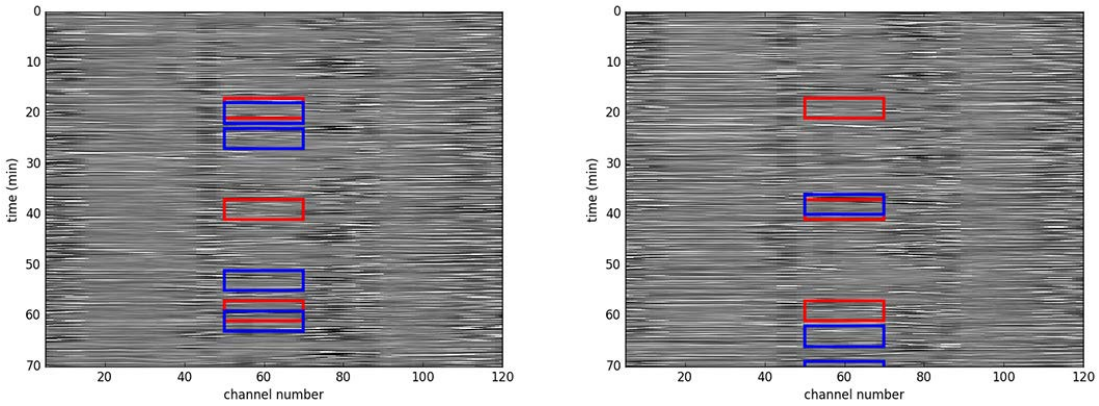


Figure 7: Manually picked events corresponding to vehicles at various locations on the Stanford University campus where only buses are allowed to circulate. [**CR**]

CONCLUSIONS

In this study we employed unsupervised and supervised learning techniques to characterize seismic data recorded by a fiber optics array deployed underneath the Stanford University campus. First, we borrowed conventional processing techniques from the field of seismology and applied clustering algorithms to distinguish between different types of noise, automatically separating noise generated by cars from incoherent background noise without requiring any information related to the geometry of the array. Clustering performed on the high amplitude events showed us we could automatically sort traffic noise depending on the car's travel direction. Template matching allowed us to detect active sources and compare them to the local bus schedule.



(a) (b)

Figure 8: Blue rectangles indicate the times of the identified active sources, while red rectangles correspond to the official timetable of the local bus line. We see that using only a limited number of templates, we can detect events that approximately match the official bus schedule. The time discrepancy can be due to the bus not running exactly on time. The detected events that do not correspond to the schedule could be empty buses or delivery trucks. [**CR**]

REFERENCES

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., 2016, Tensorflow: Large-scale machine learning on heterogeneous distributed systems: arXiv preprint arXiv:1603.04467.

Ajo-Franklin*, J., N. Lindsey, T. Daley, B. Freifeld, M. Robertson, C. Ulrich, S. Dou, E. Martin, and A. Wagner, 2015, A field test of distributed acoustic sensing for ambient noise recording, *in* SEG Technical Program Expanded Abstracts 2015, 2620–2624, Society of Exploration Geophysicists.

Bakku, S. K., 2015, Fracture characterization from seismic measurements in a borehole: PhD thesis, Massachusetts Institute of Technology.

Bensen, G., M. Ritzwoller, M. Barmin, A. Levshin, F. Lin, M. Moschetti, N. Shapiro, and Y. Yang, 2007, Processing seismic ambient noise data to obtain reliable broadband surface wave dispersion measurements: Geophysical Journal International, **169**, 1239–1260.

Collobert, R., S. Bengio, and J. Mariéthoz, 2002, Torch: a modular machine learning software library: Technical report, IDIAP.

Daley*, T., M. Robertson, B. Freifeld, D. White, D. Miller, F. Herkenhoff, and J. Cocker, 2014, Simultaneous acquisition of distributed acoustic sensing vsp with multi-mode and single-mode fiber optic cables and 3-component geophones at the aquistore co2 storage site, *in* SEG Technical Program Expanded Abstracts 2014, 5014–5018, Society of Exploration Geophysicists.

Daley, T. M., B. M. Freifeld, J. Ajo-Franklin, S. Dou, R. Pevzner, V. Shulakova, S. Kashikar, D. E. Miller, J. Goetz, J. Henninges, et al., 2013, Field testing of fiber-optic distributed acoustic sensing (das) for subsurface seismic monitoring: The

Leading Edge, **32**, 699–706.
Daskalakis, E., C. Evangelidis, J. Garnier, N. Melis, G. Papanicolaou, and C. Tsogka, 2016, Robust seismic velocity change estimation using ambient noise recordings: Geophysical Journal International, **205**, 1926–1936.
Fisher, W. D., T. K. Camp, and V. V. Krzhizhanovskaya, 2016, Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection: Journal of Computational Science.
Girard, A. and J. Shragge, 2016, Extracting body waves from ambient seismic recordings, *in* SEG Technical Program Expanded Abstracts 2016, 2715–2719, Society of Exploration Geophysicists.
James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013, An introduction to statistical learning, volume **6**: Springer.
Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, 2014, Caffe: Convolutional architecture for fast feature embedding: Proceedings of the ACM International Conference on Multimedia, 675–678.
Köhler, A., M. Ohrnberger, and F. Scherbaum, 2009, Unsupervised feature selection and general pattern discovery using self-organizing maps for gaining insights into the nature of seismic wavefields: Computers & Geosciences, **35**, 1757–1767.
Lewis, J., 1995, Fast normalized cross-correlation: Vision Interface.
Liao, T. W., 2005, Clustering of time series data: Pattern recognition, **38**, 1857–1874.
Lin, F.-C., M. P. Moschetti, and M. H. Ritzwoller, 2008, Surface wave tomography of the western united states from ambient seismic noise: Rayleigh and love wave phase velocity maps: Geophysical Journal International, **173**, 281–298.
Lobkis, O. I. and R. L. Weaver, 2001, On the emergence of the greens function in the correlations of a diffuse field: The Journal of the Acoustical Society of America, **110**, 3011–3017.
Mallat, S., 1999, A wavelet tour of signal processing: Academic press.
Martin*, E., J. Ajo-Franklin, S. Dou, N. Lindsey, T. Daley, B. Freifeld, M. Robertson, A. Wagner, and C. Ulrich, 2015, Interferometry of ambient noise from a trenched distributed acoustic sensing array, *in* SEG Technical Program Expanded Abstracts 2015, 2445–2450, Society of Exploration Geophysicists.
Martin, E., B. Biondi, M. Karrenbach, and S. Cole, 2017, Continuous subsurface monitoring by passive seismic with distributed acoustic sensors-the stanford array experiment: Presented at the First EAGE Workshop on Practical Reservoir Monitoring.
Martin, E., N. Lindsey, S. Dou, J. Ajo-Franklin, T. Daley, B. Freifeld, M. Robertson, C. Ulrich, A. Wagner, and K. Bjella, 2016, Interferometry of a roadside das array in fairbanks, ak, *in* SEG Technical Program Expanded Abstracts 2016, 2725–2729, Society of Exploration Geophysicists.
Mateeva, A., J. Lopez, H. Potters, J. Mestayer, B. Cox, D. Kiyashchenko, P. Wills, S. Grandi, K. Hornman, B. Kuvshinov, et al., 2014, Distributed acoustic sensing for reservoir monitoring with vertical seismic profiling: Geophysical Prospecting, **62**, 679–692.
Nakata, N., J. P. Chang, J. F. Lawrence, and P. Boué, 2015, Body wave extraction and tomography at long beach, california, with ambient-noise interferometry: Journal

of Geophysical Research: Solid Earth, **120**, 1159–1173.

Nakata, N., R. Snieder, T. Tsuji, K. Larner, and T. Matsuoka, 2011, Shear wave imaging from traffic noise using seismic interferometry by cross-coherence: Geophysics, **76**, SA97–SA106.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., 2011, Scikit-learn: Machine learning in python: The Journal of Machine Learning Research, **12**, 2825–2830.

Sinha, S., P. S. Routh, P. D. Anno, and J. P. Castagna, 2005, Spectral decomposition of seismic data with continuous-wavelet transform: Geophysics, **70**, P19–P25.

Wapenaar, K., D. Draganov, R. Snieder, X. Campman, and A. Verdel, 2010a, Tutorial on seismic interferometry: Part 1basic principles and applications: Geophysics, **75**, 75A195–75A209.

Wapenaar, K., E. Slob, R. Snieder, and A. Curtis, 2010b, Tutorial on seismic interferometry: Part 2underlying theory and new advances: Geophysics, **75**, 75A211–75A227.

Webster, P., J. Wall, C. Perkins, and M. Molenaar, 2013, Micro-seismic detection using distributed acoustic sensing, *in* SEG Technical Program Expanded Abstracts 2013, 2459–2463, Society of Exploration Geophysicists.

Yoon, C. E., O. OReilly, K. J. Bergen, and G. C. Beroza, 2015, Earthquake detection through computationally efficient similarity search: Science advances, **1**, e1501057.

Zeng, X., C. Thurber, H. Wang, D. Fratta, E. Matzel, and P. Team, 2016, High-resolution shallow structure revealed with ambient noise tomography on a dense array.

Zhan, Z., V. C. Tsai, and R. W. Clayton, 2013, Spurious velocity changes caused by temporal variations in ambient noise frequency content: Geophysical Journal International, ggt170.