Hierarchical exploration of continuous seismograms with unsupervised learning

René Steinmann[1], Léonard Seydoux[1], Éric Beaucé[2] and Michel Campillo[1]

[1]ISTerre, équipe Ondes et Structures, Université Grenoble-Alpes, UMR CNRS 5375, 1381 Rue de la Piscine, 38610, Gières, France
[2]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Key Points:

- Seismic data analysis
- Unsupervised learning
- Seismic waveform clustering

Corresponding author: René Steinmann, `rene.steinmann@univ-grenoble-alpes.fr`

Abstract

We propose a strategy to identify seismic signal classes in continuous single-station seismograms in an unsupervised fashion. Our strategy relies on extracting meaningful waveform features based on a deep scattering network combined with an independent component analysis. We then identify signal classes from these relevant features with agglomerative clustering, which allows us to explore the data in a hierarchical way. To test our strategy, we investigate a two-day long seismogram collected in the vicinity of the North Anatolian fault in Turkey. We interpret the automatically inferred clusters by analyzing their occurrence rate, spectral characteristics, cluster size, and waveform and envelope characteristics. At a low level in the cluster hierarchy, we obtain three clusters related to anthropogenic and ambient seismic noise and one cluster related to earthquake activity. At a high level in the cluster hierarchy, we identify a seismic crisis with more than 200 repeating events and high-frequent signals with correlating envelopes and an anthropogenic origin. The application shows that the cluster hierarchy can be used to draw the focus on a certain class of signals and extract subclusters for further analysis. This is interesting, when certain types of signals such as earthquakes are under-represented in the data. The proposed method can be also used to discover new types of signals since it is entirely data-driven.

Plain Language Summary

Seismic data most likely contain a wealth of crucial information about active geological structures such as faults or volcanoes. The growing amount of seismic data collected nowadays cannot scale with manual investigation, suggesting automatic algorithms for scanning continuous data streams. We develop a strategy based on artificial intelligence to scan continuous seismic data and infer patterns automatically. We propose a hierarchical approach to gather similar signals into families since we expect the content of seismic data to be complex, dominated mainly by noise and with rare events such as explosions or earthquake signals. Our strategy relies on a particular neural network, the scattering network, to ease design and training. This paper analyzes two days of continuous seismic data collected in the vicinity of the North Anatolian fault. We compare and discuss our results with classical approaches for earthquake detection and noise description.

1 Introduction

When the first seismometers were developed and put in place, their primary purpose was to better understand earthquakes since they were a major hazard to human-kind. However, with time, seismologists found many kinds of signals on these recordings. It is only two decades ago that Obara (2002) discovered a new type of signal with tectonic origins called non-volcanic tremors. Other than tectonic signals seismometers also record the oceanic microseisms (see e.g. Ebeling, 2012, for a recent review), rockfalls and other mass movements (e. g. Lacroix & Helmstetter, 2011; Deparis et al., 2008), ground and air traffic (e. g. Riahi & Gerstoft, 2015; Meng & Ben-Zion, 2018) or other kind of human-induced sources (such as church bells in Diaz, 2020). The mixing of all these sources renders a complex seismic wavefield that makes the analysis and interpretation of seismic records difficult, especially if seismic data are the only data available. As a response to this problem, seismologists have developed many processing tools for exploring these complex seismic data. Nowadays, seismology benefits from artificial intelligence developments, bringing new machine-learning-based solutions for exploring seismic data, as we demonstrate in the present paper.

From a general perspective, machine learning defines a framework to solve tasks (such as recognizing patterns in a data set) when rule-based algorithms are not easy to formulate. In order to do so, machine-learning algorithms mostly rely on a set of data characteristics with which the task is easier to solve, instead of the data itself. These characteristics are called features. Finding the most relevant features should be done according to the task at hand

and can be done thanks to prior knowledge on the data or by defining proper algorithms to learn the most relevant features. We distinguish classical machine-learning algorithms that rely on human-defined features (Maggi et al., 2017; Malfante et al., 2018) or representation-learning algorithms where the features are learned from the data to optimize a given task (LeCun et al., 2015; Ross et al., 2018; Rouet-Leduc et al., 2020). While classical machine learning provides less accuracy in most cases, it provides interpretability since the features are known, which is an interesting aspect. Most algorithms that rely on representation learning are less easy to interpret since the features are more abstract, but they also provide more accurate results. In the present paper, we propose to use a hybrid approach between classical and representation learning algorithms that combines the advantages of both.

Once the features are defined, a model is trained to map the features to a certain output. In supervised learning the model learns that mapping based on a labeled training data set (Goodfellow et al., 2016). For instance, locating earthquakes with a single seismic station is a non-trivial task to address with supervised learning. It has been recently shown that deep neural networks can infer the position of earthquakes from single-station records (Perol et al., 2018; Mousavi & Beroza, 2019). This experiment illustrates the range of applications that supervised machine learning can find in the domain of seismology. Still, the limitations of supervised learning lie within the ones of our expert knowledge. Indeed, supervised models can only automatize a given task and therefore strongly rely on the quality of the labels of the training data set. Moreover, supervised learning is limited to the labels we know and cannot search for signatures or patterns with unknown properties in a data-exploration fashion.

Unsupervised learning methods such as clustering can overcome this particular problem as they provide tools to explore seismic data without labels, thus without any human biases (Bergen et al., 2019). In the particular case of waveform clustering, a classic approach consists of two steps: firstly, the continuous seismic data is projected onto a feature space, and then a clustering algorithm performs the identification of classes in this given feature space. We here review a few studies that applied this approach for the unsupervised classification of seismic signals.

In Köhler et al. (2010), the authors use self-organizing maps for a data-driven feature selection and clustering of seismic waveforms. With that approach they identify different long-term variations and short-term seismic events in the continuous data. They also mention that a human-based inspection should replace the automatic selection of the number of clusters based on a validity measure. In Johnson et al. (2020), the authors label continuous seismic data by performing a k-means clustering in a reduced spectral representation of the input seismic data recorded by a dense seismic array. They identify five clusters that mostly describe the weak ground motion but also contain signals from earthquakes. In Seydoux et al. (2020), the authors generate features with a deep scattering network and then use a Gaussian-mixture model for clustering. They blindly identify a recurrent precursory signal before a landslide in a daylong data set with this approach. However, if they increase the data set to 17 days, the class population imbalance becomes too large to recover the precursory signal into a single cluster. Instead, they identify two clusters related to seismic waves generated by storm systems in the Atlantic ocean. As a solution, they propose to perform a second-order clustering revealing more details in the first-order clusters and retrieving the precursory signal. In Kodera and Sakai (2020), the authors introduce an anomaly detector before clustering the seismic data in the spectral domain. The anomaly detector erases the class imbalance, and the clustering algorithm focuses only on the outliers. Detecting anomalies can be interesting if only anomalies are the target of interest, but it does not give a complete picture of the seismic data.

These studies have shown that extracting earthquake clusters from continuous seismic data is difficult mainly due to class imbalances, *i.e.* the large disparities between occurrences of different class of signals (He & Garcia, 2009). Indeed, we know that signals such as those produced by earthquakes inhabit a tiny part of the data, while seismic noise inhabits most

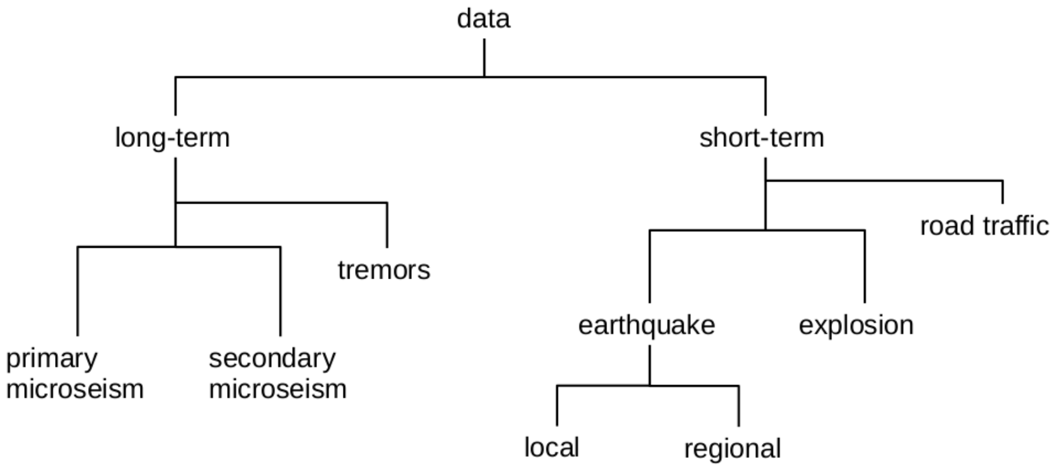

Figure 1. Illustration of possible hierarchy in seismic data. The different branches represent how a signal class splits into different subclasses depending on a given similarity measure. Here the different classes of events are thought in a hierarchical way, based on arbitrary signals properties (e.g. duration, frequency range or signal's structure). This scheme aims at illustrating the expected behavior of an optimal clustering algorithm, but does not depict the potential issues related to clustering such as overlapping between different classes of signals or imbalance between classes.

of the data. Therefore, the clustering algorithms such as k-means are very likely to find different classes of noise (Johnson et al., 2020).

In the present study, we introduce a new strategy that explores seismic data in an unsupervised fashion and finds classes with largely imbalanced population sizes. Our strategy follows the idea that seismic signals cluster in a hierarchy of classes following a specific metric, as schematized in Figure 1. Note that this illustration aims at sketching the concept rather than being complete or accurate. In the following lines, we consider the similarity between signal classes to be measured on a set of signal features that can be human-defined (such as mean frequency and signal duration) or learned with machine-learning tools, as we propose to do in the present paper. In the first place, one can imagine the seismic signal classes to split into long-term and short-term signals based on the duration of a signal (Figure 1). In the class of long-term signals, one could use a similarity measure based on frequency content to separate the primary from secondary microseism. We see that building a tree of classes lets us explore the data on different levels and that different features may be relevant at each node of the tree.

The sketch presented in Figure 1 also illustrates the problems of designing a data hierarchy by hand. The labels used in this sketch are the ones we created as seismologists based on our domain knowledge. That is problematic for those classes of signal that do not have a proper definition of signal and source properties, such as non-volcanic tremors. Moreover, some splittings, such as between earthquakes and explosions, ask for a more complex similarity measure which will be hard to design by hand. Hierarchical clustering produces precisely this kind of tree, called a dendrogram, based on the exploration of the similarity between samples in the feature space representation. Therefore, we propose to represent seismic data as a dendrogram and utilize it to explore the data in an unsupervised and unbiased way.

In the next section, we present the workflow to build a dendrogram from continuous single-station data. In section 3, we introduce a data set to apply and test the proposed

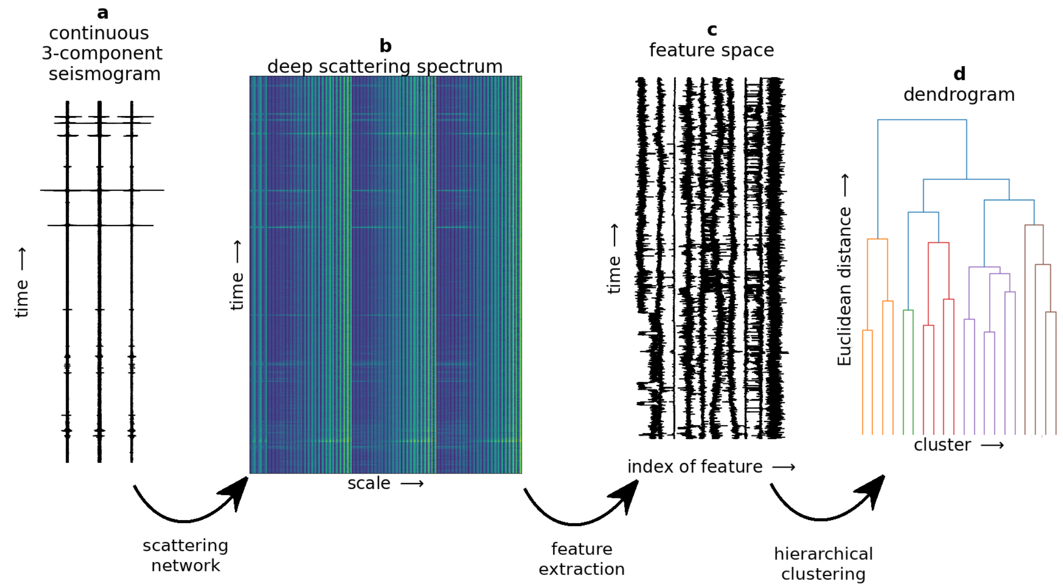

Figure 2. Proposed workflow for exploring continuous seismograms in a hierarchical way. **(a)** Input continuous 3-component seismograms, as detailed in Section 3. **(b)** Deep scattering spectrum of the seismograms, with a lower temporal resolution and a high number of dimensions, detailed in Section 2.1. **(c)** Independent features extracted from the deep scattering spectrum with independent component analysis, following the description in Section 2.2. **(d)** Dendrogram calculated from a similarity metric in the feature space, as explained in Section 2.3.

workflow. In section 4, we show and discuss briefly the resulting dendrogram. Section 5 is about navigating through the dendrogram and interpreting the clusters at different levels.

2 Method

A sketch of the hierarchical clustering workflow is depicted in Figure 2. To automatically infer classes in seismic data, one needs to project the time-domain data onto a feature space with invariant properties towards translation or small deformation (Andén & Mallat, 2014). For that purpose, we implement a scattering network (a convolutional neural network with wavelet filters) to calculate a deep scattering spectrum of the seismic data (Figure 2b). The large-dimensional representation provided by the deep scattering spectrum is not well-suited for the commonly used clustering algorithms (Beyer et al., 1999; Kriegel et al., 2009). Therefore, we extract the most relevant features out of the deep scattering spectrum with an independent component analysis (Comon, 1994), hereafter mentioned as the feature space (Figure 2c). Finally, we perform hierarchical clustering within the feature space and utilize the dendrogram for exploring the data and extracting clusters (Figure 2d).

2.1 Deep scattering spectrum

In this study, we aim to explore the data in an unsupervised and unbiased way, i.e., we want to assume as little as possible about the data themselves. For that purpose, it is crucial to find a representation that is not a human-based selection of features. A time-frequency representation such as the spectrogram is one way to create a set of features without favoring any data characteristics. However, Andén and Mallat (2014) showed that a spectrogram generated by the Fourier transform is not ideal for classification purposes since it is not stable to time-warping deformations, especially at short periods compared with the duration of the analyzing window. They introduce another time-frequency representation called a deep scattering spectrum which is computed by a scattering network. This type of network implements a cascade of convolutions with wavelet filters, modulus function, and pooling operations. Deep scattering spectra are locally translation invariant and preserve transient phenomena such as attack and amplitude modulation. These characteristics are beneficial when it comes to classifying any time series data. In Andén and Mallat (2014) and Peddinti et al. (2014), the authors have successfully classified audio data based on the deep scattering spectrum. Seydoux et al. (2020) have brought that representation into seismology and showed that small precursory signals of a landslide could be detected and classified in an unsupervised fashion. Other successful deep-learning classifiers inspired by deep scattering networks are presented in Balestriero et al. (2018) and Cosentino and Aazhang (2020).

We here use the strategy presented in Seydoux et al. (2020) for calculating the deep scattering spectrum. Considering the continuous input signal $x(t) \in \mathbb{R}^C$ (where C is the number of channels), the scattering coefficients $S^{(\ell)}$ of order ℓ are obtained from the following cascade of wavelet convolutions and modulus operations:

$$S^{(\ell)}\left(t, f_{n_1}^{(1)}, f_{n_2}^{(2)}, \ldots, f_{n_\ell}^{(\ell)}\right) = \max_{[t,t+dt]} \left|\phi^{(\ell)}\left(f_{n_\ell}^{(\ell)}\right) \star \left|\ldots \star \left|\phi^{(2)}\left(f_{n_2}^{(2)}\right) \star \left|\phi^{(1)}\left(f_{n_1}^{(1)}\right) \star x\right|\right|\right|\right|, \tag{1}$$

where $\star$ stands for the temporal convolution, $|\cdot|$ represents the modulus operator and $\phi^{(i)}(f_{n_i}^{(i)})$ is the wavelet filter at the layer i of the scattering network, with center frequency f_{n_i}. Here f_{n_i} refers to one of the center frequencies of the layer i indexed by $n_i = 1 \ldots N_i$, where N_i is the total number of wavelets at layer i; the number of wavelets per layer and frequency range of each layer is discussed later. While the authors in Seydoux et al. (2020) implement a learnable wavelet filter $\phi^{(i)}(f_{n_i}^{(i)})$ with respect to the clustering loss, we directly use a (non-learnable) Gabor filter, as originally presented in Andén and Mallat (2014). This choice was made principally because we do not perform a fixed cluster analysis in our study, but an exploration of the data instead where a loss function is harder to define. The maximum-pooling operation is performed over a time interval $[t, t+dt]$ of duration dt over the continuous data; the data sampling rate and the pooling operation control the final sampling rate of the deep scattering spectrum. While the first-order scattering coefficients resemble a spectrogram based on a wavelet transform, the second-order scattering coefficients contain information about the attack and modulation. For the interested reader we refer to Andén and Mallat (2014) and Seydoux et al. (2020).

2.2 Features extraction from deep scattering spectrum

The data in the scattering domain can have more than 1,000 dimensions and, thus, the conditions for clustering are not favorable (Kriegel et al., 2009). Indeed, distances in very high-dimensional spaces give little information about the structure of the data (the so-called curse of dimensionality Bellman, 1966). In addition, the representation is known to be highly redundant since the wavelet filters of the first scattering layer are often considered with a strong frequency overlap in order to provide a dense first-order representation. Therefore, it is recommended to reduce the dimensions before clustering. In our case, we use an independent component analysis (ICA) to reduce the dimension of the representation. In the following remarks, we explain the basic concept of ICA. For the interested reader we refer to (Comon, 1994).

ICA is introduced as a statistical tool for blind source separation and feature extraction. The generative model of the ICA can be described as:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^{N \times F}$ are the N observations of dimension F, $\mathbf{s} \in \mathbb{R}^{F \times C}$ are the C independent components of the same dimension F as the observations and $\mathbf{A} \in \mathbb{R}^{C \times N}$ is the mixing matrix (that is, the mixing of the C independent components for every observation). The observations are therefore a linear combination of the independent components. A test of statistical independence is required to solve Equation 2 while ensuring the components $\mathbf{s}$ to be independent. Among the different strategies, we can look for a minimum of mutual information, or similarly, a maximization of the non-Gaussianity. In our study, we apply the `FastICA` algorithm from the `scikit-learn` Python library, which uses the negentropy as a measure of non-Gaussianity (Hyvärinen & Oja, 2000). This analysis is similar to the principal component analysis, with the difference that the independent components are not orthogonal. In addition, there is no information about the variance explained by the different independent components, and are therefore delivered unsorted by the algorithm.

2.3 Hierarchical clustering

Cluster analysis is one way to assign labels to data samples in a given feature representation with unsupervised learning. The choice of the clustering algorithm depends mainly on the statistical characteristics of the data set. Seismometers are highly sensitive sensors over a wide range of frequencies. Most seismic records are dominated by ambient seismic noise, while seismic activity only inhabits a small part. In that sense, we expect classes of signals (or types of noises, as depicted in Figure 1) to be largely imbalanced; that is, most samples may belong to the background noise class while only a few data samples can relate to different classes of seismic events. Therefore, the k-means algorithm, which tends to identify clusters with similar population sizes and variances, can have difficulties detecting a repeating waveform pattern, with only a few occurrences, in a relatively large data set (Lin et al., 2017).

This study investigates how the data cluster in a hierarchical way with a bottom-up approach, namely agglomerative clustering. At first, we briefly introduce the methodology. Hierarchical clustering relies on a similarity matrix, which defines the similarity (e.g., a specific distance) between all data points in the data set. With a bottom-up approach, all data points start in a singleton cluster. The clusters start merging based on the similarity matrix until all data points unify in a single global cluster. This process is summarized in a dendrogram, revealing the hierarchical structure of the entire data set. Such a strategy fits very well the nature of seismic data, which records wavefields from different sources (Figure 1).

The agglomerative clustering outcome depends mainly on the applied metric, which drives the merging of the cluster. In our approach, we use the Ward's method (Ward Jr, 1963). Given a distance d (here considered Euclidean), the Ward's method aims at grouping data samples x_i into clusters such as the within-cluster variance remains minimal after merging different clusters. The within-cluster variance quantifies the spread σ of each within-cluster data samples defined in Appendix B. By minimizing the overall variance, $\sum_{c=1}^{K} \sigma_c$, the Ward's method allows for data clusters of variable population sizes and variances and may highlight clusters of high density located in the vicinity of more spread, low-density clusters. Therefore, Ward's method is suitable for the expected seismic data partition.

3 Data

We test our proposed workflow on continuous three-component seismic data from the station DC06 of the DANA experiment in Turkey (see for instance Poyraz et al., 2015, and the map shown in Figure 3a). The sampling rate of the data is 50 Hz. We choose the data set for mainly two reasons. First of all, the data set contains both seismic and anthropogenic activity, which is a typical situation in most seismological studies. Second of all, an existing template matching catalog provides labels for the seismicity in this area. The catalog was built following the methodology in Beaucé et al. (2019).

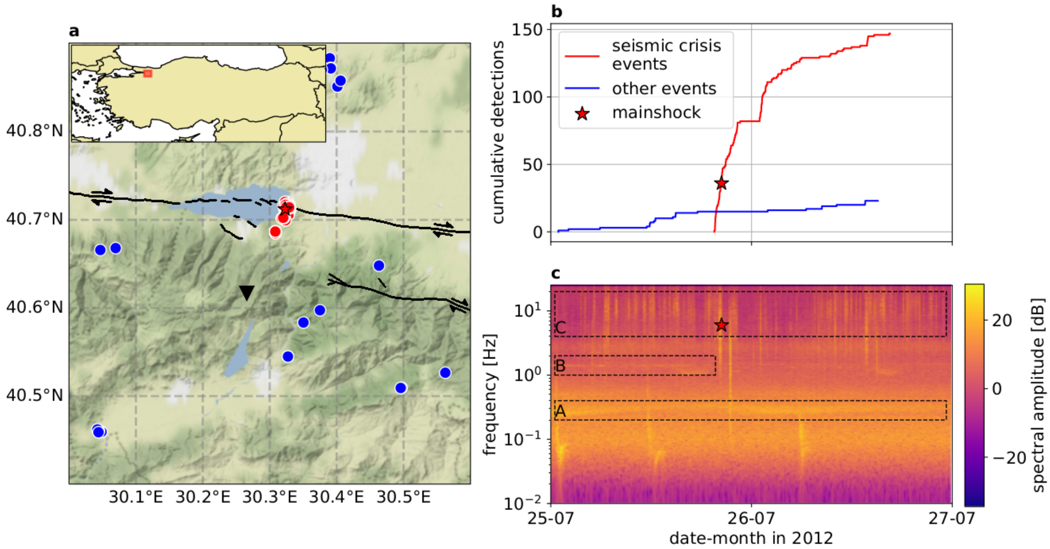

Figure 3. Geological context and seismic data used in the present study. **(a)** Map of the North Anatolian fault zone showing station DC06 (black triangle), the seismic crisis (red dots) including the identified mainshock (red star) and other seismic activity (blue dots); all detected with a template matching strategy. The geological faults that ruptured after 1900 (black lines) are adapted from Emre et al. (2011). **(b)** Cumulative detections of the seismic crisis (in red) and other seismic activity (in blue) obtained with template matching. **(c)** Continuous spectrogram of the east-component of station DC06, with a visual identification of (A) oceanic microseism, (B) a non-stationary monochromatic noise source, and (C) daily high-frequency activity.

We choose to analyze the seismic data from the 25th to the 27th of July 2012. During that period, a seismic crisis with 148 events occurred on and around the northern strand of the North Anatolian fault (see Figure 3a and b). The catalog explains the series of events with 17 templates having their hypocenters close to each other (Figure 3a, red dots). Since the seismic crisis resembles a repeating pattern with short time-warping deformations due to slight changes of the hypocenters, it is an interesting study case for our proposed method.

The spectrogram of the east component of station DC06 is presented in Figure 3c. The oceanic microseism is visible around 0.2 Hz, where we can observe the dispersive nature of the oceanic gravity waves. At around 1.5 Hz we can identify a nonstationary monochromatic noise source, which seems to be more active during the first day. At frequencies higher than 3 Hz we can see increased activity during daytime, most likely induced by anthropogenic noise sources. The main shock of the crisis during the evening of the 25th is also easy to spot in the spectrogram.

4 Results

4.1 Feature space

Firstly, we use the continuous three-component seismograms to calculate the deep scattering spectrum with the scattering network (as detailed in Equation 1). The network parameters are physics-driven and can be adjusted according to the goal. We use a two-layer network since Andén and Mallat (2014) argued that more layers do not necessarily introduce new valuable information. The first layer performs 24 wavelet transforms per channel starting at the Nyquist frequency (25 Hz) and going down to a frequency of 0.78 Hz. The second layer performs 15 wavelet transforms on each of the first-order wavelets transforms, from the Nyquist frequency down to a frequency of 0.19 Hz. With three channels as an input, 24 wavelet transforms at the first layer, and 15 wavelet transforms at the second layer, we have 1080 wavelet transforms. Since we are interested in detecting and classifying non-stationary events such as the seismic crisis, we use maximum pooling to downsample the scattering coefficients. If the focus of classification is the background noise, average pooling might be the better choice (as suggested in Seydoux et al., 2020). The scattering network transforms the three-channel continuous seismic data (3 x 8646001 data points) into the scattering coefficients (1080 x 8384 data points), i.e., we highly decreased the number of samples in time and highly increased the number of dimensions. The time resolution of the scattering coefficients is around 20.48 s.

For dimensionality reduction, we apply an independent component analysis using the `FastICA` algorithm from the `scikit-learn` Python library. In this study, we select the appropriate number of independent components according to the reconstruction loss between the original data and the reconstructed data after compression with an ICA (detailed in Appendix A). We emphasize that we look for a trade-off between keeping the most significant amount of information while using few independent components. From the study of the loss with increasing number of components shown in Appendix A and Figure A1 therein, we conclude that keeping ten independent components is a good compromise and constitute our choice in the present study. A visual representation of the ten independent components building the feature space is depicted in Figure A2 in Appendix A.

4.2 Dendrogram

After transforming the continuous seismic data into a most relevant set of features, we can use this representation to explore the data with hierarchical clustering. By controlling the distance threshold, we can extract different numbers of clusters. In Figure 4a we selected a distance threshold of 0.47 in order to show a truncated dendrogram stopping at 16 clusters. At a distance of 0.9, we extract four main clusters labeled as A, B, C, and D. Figure 4b shows the averaged first-order scattering coefficients of these four clusters. These first-

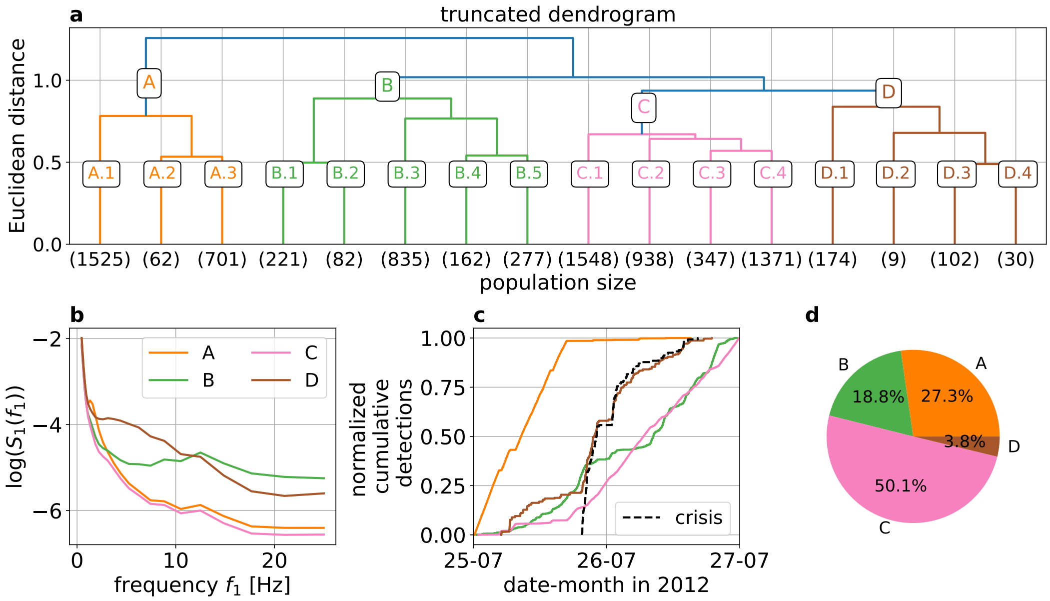

Figure 4. Dendrogram analysis and statistical characteristics of the different clusters. **(a)** Dendrogram calculated in the feature space (see Sec. 2.3 for explanations). The dendrogram is here truncated in order to form 16 clusters. The clusters marked with a letter are considered the main clusters, and the subclusters are indicated with numbers. The number of samples in each cluster indicates the numbers in the parenthesis. **(b)** Centroidal first-order scattering coefficients for main clusters A, B, C and D. **(c)** Normalized cumulative detections of main clusters A, B, C and D, and of the seismic crisis obtained from the multi-station template-matching catalog. **(d)** Relative size of the main clusters compared to the size of the entire data set.

order scattering coefficients describe the frequency characteristics of each cluster. Figure 4c presents the normalized cumulative detection rate of each cluster, with the seismic crisis detection rate indicated as a reference. The relative size of each cluster compared to the size of the entire data set is depicted in Figure 4d. In the following remarks, we will analyze each of the four main clusters from left to right.

Cluster A contains ca. 27 % of the data (Figure 4d) and is the first cluster to split from the whole data set, i.e., cluster A is the furthest away from the center of the data points (Figure 4a). Compared to the other clusters, its scattering coefficients for all frequencies are relatively low except for a local maximum around 1.5 Hz (Figure 4b). Looking at the corresponding cumulative detection curve (Figure 4c), we see that this cluster is active mainly during the first day until the late afternoon, which seems to correlate with the monochromatic signal around 1.5 Hz we have already identified in the spectrogram (Figure 3c).

Cluster B contains about 19 % of the data samples (Figure 4d) and has relatively large scattering coefficients for frequencies above 10 Hz (Figure 4b). The corresponding cumulative detection curve indicates that this cluster accumulates less detections during the beginning of a day than with later times of a day (Figure 4c). Combining these facts leads to the hypothesis that cluster B might be related to signals with an anthropogenic origin.

Cluster C is the largest cluster with more than 50 % of the data points (Figure 4d). Compared to the other clusters, it also has the lowest scattering coefficients at all frequencies (Figure 4b). Looking at the cumulative detection curve (Figure 4c), we see this cluster shows an almost linear increase starting at the afternoon of the first day, exactly when cluster A becomes almost inactive. The cluster size and frequency content suggest that cluster C is related to samples containing only ambient noise.

Finally, cluster D contains about 4 % of data set (Figure 4d) and is the smallest of the four clusters (Figure 4d). The corresponding first-order scattering coefficients show a local maximum around 5 Hz (Figure 4b). Its cumulative detection curve correlates well with the detections of the seismic crisis (Figure 4c), with additional detections before the seismic crisis starts. All these observations indicate that cluster D is probably related to nearby seismic activity in general.

5 Discussion

In this section, we will discuss and interpret the dendrogram's representation and its clustering solution. While the main focus is on identifying how the seismic crisis occurs in the dendrogram, we will also discuss how the general seismicity is observed through this representation, and interpret the remaining clusters with anthropogenic activity and ambient seismic noise properties. '

5.1 Identification of the seismic crisis within the dendrogram

Firstly, we identify all time segments containing onsets of the events of the seismic crisis and observe which clusters those time segments belong to. The template matching catalog contains 148 detections related to this seismic crisis. However, we only associate 136 samples in the feature space with the seismic crisis, since one sample represents about 20 s of waveform data and, thus, can contain multiple events. Figure 5a shows that a large majority of the samples, which contain arrivals of the seismic crisis, fall into cluster D (92.6 %). On the other hand, only 40 % of cluster D is related to the seismic crisis, underpinning the statement that this cluster is related to general seismic activity. Cluster B and C share the remaining 7.4 % of the crisis. Compared to the large population sizes of clusters B and C, the contribution of the crisis almost vanishes (0.3 and 0.1 %). Cluster A contains no detections of the crisis. While cluster D contains the majority of the seismic crisis, the interesting aspect is to understand what the remaining 60 % samples of this cluster are related to (earthquakes from the same source region, different signals, etc). To answer that question, we investigate the subclusters visible in Figure 4a obtained with a distance threshold of 0.47; in particular, we will narrow the focus on the subclusters of cluster D, namely the four subclusters D.1 to D.4.

Firstly, we look at the distribution of the samples containing the seismic crisis across the four subclusters in main cluster D. From Figure 5a, we know that more than 92 % of the crisis was found in cluster D. We observe in Figure 5b that this amount splits into ca. 71.3 % in cluster D.1 and ca. 21.3 % in cluster D.4. The subclusters D.2 and D.3 contain no earthquakes from the seismic crisis and will be discussed later. If we look at the cumulative detection curve of each subcluster in D (Figure 5c), we see that cluster D.1 and D.4 share a very similar temporal pattern. The corresponding centroidal first-order scattering coefficients (Figure 5d) explain why the crisis got split into two clusters: across almost all frequencies the larger subcluster D.1 shows significantly smaller scattering coefficients than the smaller subcluster D.4. Hence, the magnitudes of the events seem to be the characteristics that separates the crisis into two clusters. Besides, we observe that 56 % of D.1 and 97 % of D.4 can be explained by the cataloged crisis. This observation raises the question: what are the samples in D.1 and D.4 that cannot be related to the seismic crisis recorded by the catalog? We can answer this question by looking at the waveforms representing the corresponding data points of subclusters D.1 and D.4.

Figure 6a, b and c show the corresponding waveforms of all 204 data points of the two subclusters D.1 and D.4. For all waveforms we observe the P and S seismic phase

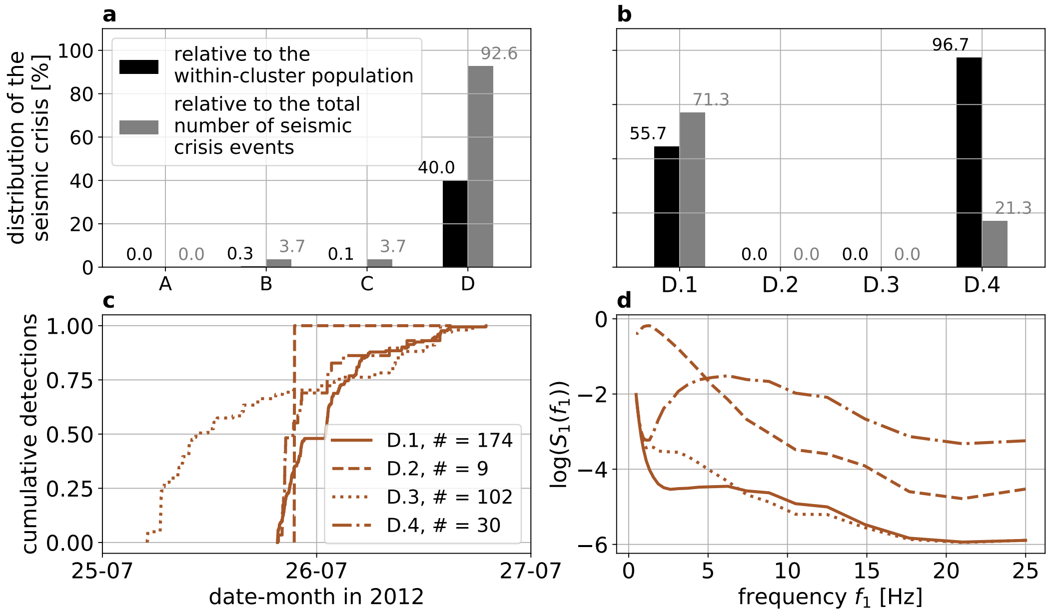

Figure 5. Identification of the seismic crisis within the main and subclusters. **(a)** The distribution of the seismic crisis across the four main clusters. **(b)** The distribution of the seismic crisis across the four subclusters in the main cluster D. **(c)** Normalized cumulative detection curves for the subclusters in the main cluster D. **(d)** Centroidal first-order scattering coefficients for the subclusters in the main cluster D.

arrivals of the earthquakes. The first 30 waveforms correspond to subcluster D.4. 29 of them are are also in the catalog (marked orange) while 1 of them is not in the catalog (marked magenta). The following 174 waveforms are from subcluster D.1. 98 of them are are also in the catalog (marked light blue) while 76 of them are not in the catalog (marked blue). The waveforms are very similar to each other on all three channels. This indicates that these new detections are coming from the same source area. Note also that the first 30 waveforms representing subcluster D.4 have a better signal-to-noise ratio than the following waveforms of subcluster D.1. This agrees with our assumption that the crisis is split into two subclusters due to magnitude differences. The magnitude estimations of the template matching catalog confirms this assumption (see Figure 6d). While most of the events located in D.1 range between M0.5 and M1, the events located in D.4 range between M1 and M2.2.

By investigating cluster D and its subclusters D.1 and D.4, we are able to identify two subclusters representing the seismic crisis. While D.1 contains many events with smaller magnitudes, D.4 contains fewer events with larger magnitudes. Together the two subclusters contain 92.6 % of the cataloged events and 77 new events, which have identical P and S wave arrivals as the cataloged ones. The new detections can be explained by the fact that we utilize a single station method and compare it to a catalog based on a multi station method. More details and a comparison with a single station template matching catalog based on station DC06 can be found in Appendix C.

However, 7.4 % of the cataloged detections can not be found in subclusters D.1 or D.4. In the following remarks, we want to analyze the misidentified 7.4 % of cataloged events, which equal ten over 135 events. First of all, we want to know where these events are located in the feature space. Therefore, we calculate the Euclidean distance between the misidentified events and the centroids of each cluster in the feature space (see Figure 7a). In magenta, we highlight the distance between the sample and its respective subcluster. In

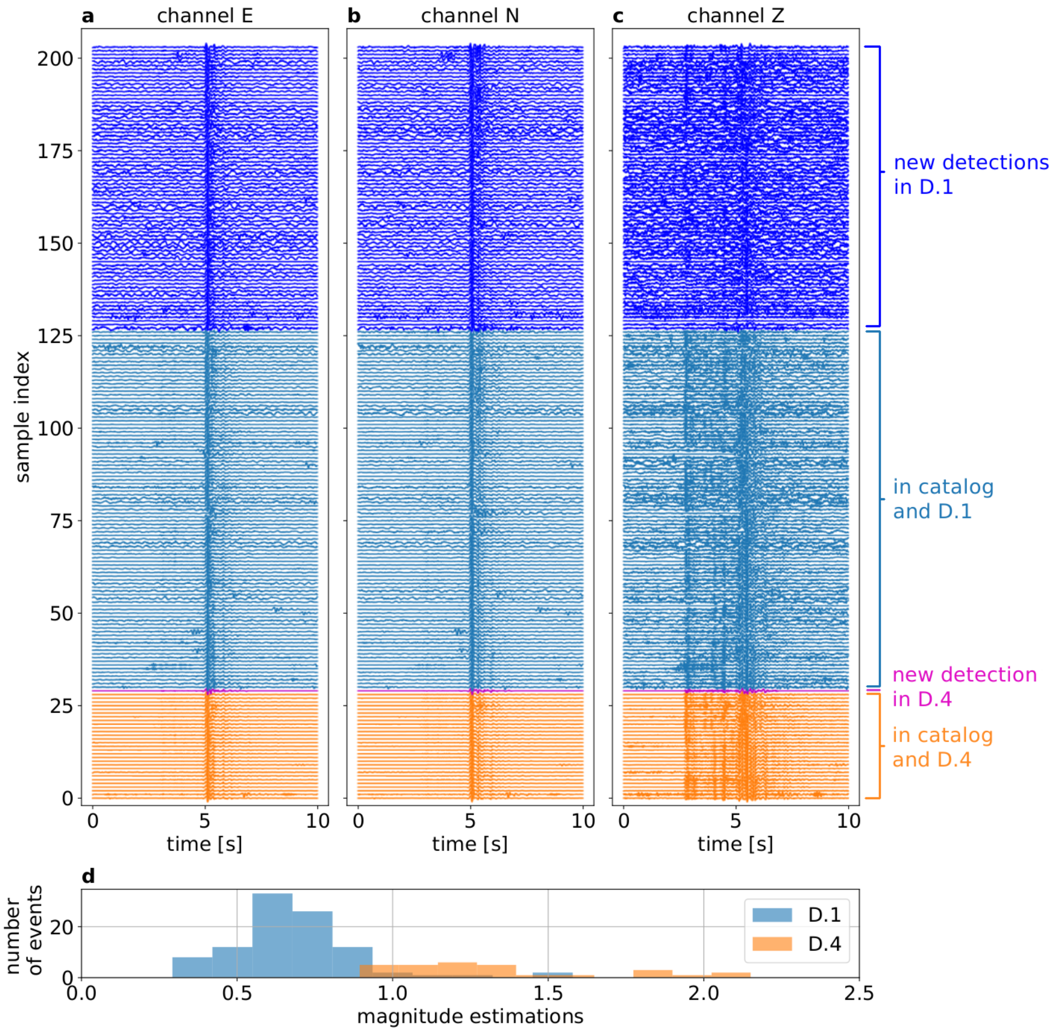

Figure 6. **(a,b,c)** Waveform data from subcluster D.1 and D.4. The color code indicates the according subcluster and if the event is mentioned by the catalog. **(d)** Magnitude estimations of the cataloged events of the seismic crisis found in subcluster D.1 and D.4.

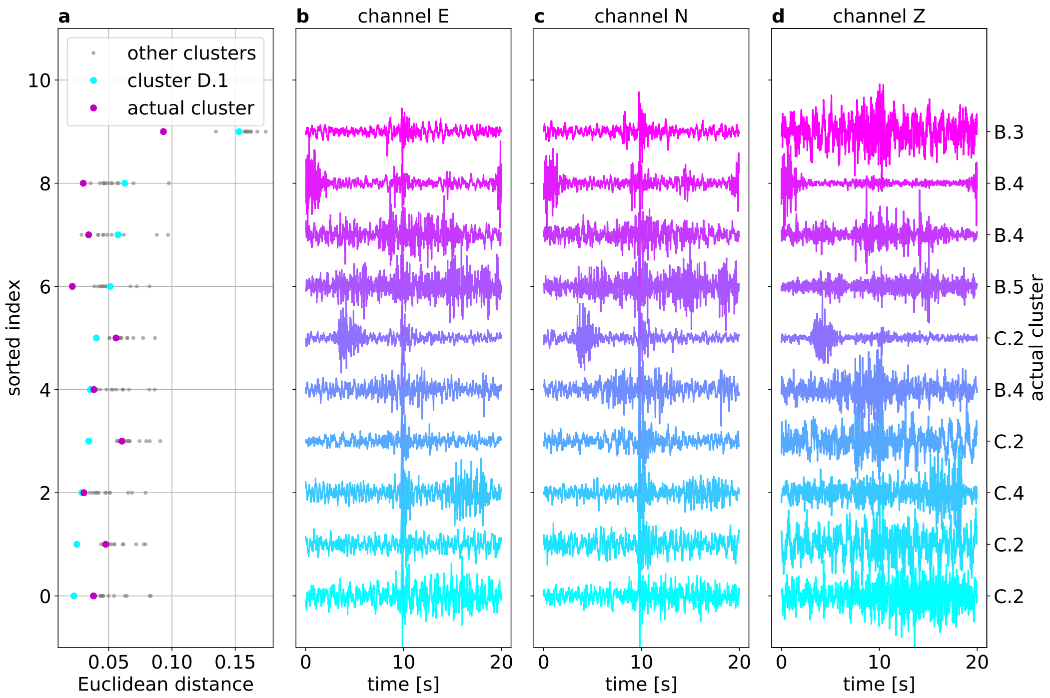

Figure 7. Analysis of the misidentified earthquake waveforms. **(a)** Distances between misidentified data points containing an event from the catalog and the centroids of all clusters. The magenta points show the distance between the data point and the centroid of its own respective subcluster. The cyan points show the distance between the data point and the centroid of D.1. The gray points show the distance between the data point and the centroids of the other 14 subclusters. **(b, c, d)** Corresponding aligned waveform data sorted according to the distance to the centroid of D.1 (respectively channels E, N, and Z). The color coding represents the distance to the centroid of subcluster D.1. A purple color indicates a larger distance than a light blue color.

cyan, we highlight the distance between the sample and subcluster D.1 containing the low magnitude events of the crisis. In gray, we highlight the distances to all other remaining clusters as a comparison. We sorted the misidentified ten events according to the distance to the centroid of D.1. We see that for the first six events, the distance to the centroid of D.1 is smaller than to the centroid of its respective cluster. The corresponding waveform data also offer explanations for the misidentification (Figure 7b to d). Indeed, the P and S arrivals are noisy but visible for the first five events. Thus, some events might be misclassified because samples are grouped with the Ward's method, which solves iteratively an objective function considering the Euclidean distance and the within-cluster variance. In other words, clusters can agglomerate samples which might be closer to the centroids of other clusters if we consider the pure Euclidean distance. After the first five events, when the distance to its respective cluster becomes smaller than the distance to D.1., the P and S arrivals are not visible anymore, or other large-amplitude events are present. Here the problem is related to the representation of the data as a deep scattering spectrum or in the feature space. Other large-amplitude transients can corrupt the representation since we perform a maximum pooling to extract the scattering coefficients. This is not a specific problem of maximum pooling but pooling in general since this operation reduces information in the data.

5.2 Neighboring clusters of the seismic crisis in the feature space

Having identified most of the seismic crisis in two neighboring subclusters already shows that the representation of the data and the distances between the data points are meaningful. As a next step, we want to analyze the neighborhood of these two subclusters to get a better understanding of the data representation. Since D.2 and D.3 share the same cluster with D.1 and D.4, we know that they are located next to each other in the feature space. This indicates that subcluster D.2 and D.3 might contain similar signals, such as seismic activity with a different origin than the seismic crisis.

To verify this assumption, we can compare existing earthquake catalogs with the timestamps of the samples in the subclusters. We extend the local template matching catalog with a regional catalog limited to events within a radius of 5° around station DC06. The regional catalog is downloaded from IRIS. For calculating the seismic phase arrivals at the station, we use the `TauP` module of `ObsPy` with the velocity model of Kennett and Engdahl (1991). We consider a sample related to an event of the catalog if the 20 s window of the sample overlaps with the window between the P wave arrival and the decaying coda.

The waveform data of D.2 and D.3 are presented in Figure 8. Figure 8a indicates the samples which can be explained by arrivals of a regional or local event, and Figure 8b shows the samples which can not be explained by arrivals of a regional or local event. Note that one sample represents ca. 20 s of waveform data and that consecutive samples are represented by one index. Subcluster D.2 contains only nine samples corresponding to two seismic events indicated in blue in Figure 8a. The first event represented by eight consecutive samples at index 0 is a relatively distant $M4$ event. The other event represented by a single sample is a quarry blast from a local mine mentioned by the template matching catalog. At first sight, it might seem unexpected that these two events are found in the same subcluster. However, subclusters D.2 shows the largest scattering coefficients for frequencies below 5 Hz (see Figure 5d), and its centroid is the furthest away from the remaining data set as we can see from the inter-cluster distance matrix presented in Figure B1 in Appendix B. Moreover, the within-cluster variance σ_c in the top panel of Figure B1 indicates that the samples of subcluster D.2 are the most spread out compared to the other subclusters, This suggests that both events are seen as outliers in the data space due to their high amplitudes at lower frequencies.

Moreover, we observe that the catalog can explain 67 % of all samples of D.3. However, we only show some waveforms in black in Figure 8a. The other 33 % are shown in Figure 8b, and some samples also show seismic phase arrivals (in particular, the seismograms shown at index six and nine). It is thus likely that the samples shown in Figure 8b contain uncataloged events. While subcluster D.1 and D.4 represent similar earthquakes from a similar source region, subcluster D.3 shows many kinds of signals, such as earthquakes with different magnitudes and distances to the station. We can interpret subcluster D.3 as an agglomeration of transient signals with increased energy between 1 and 5 Hz (see Figure 5d). Regional and local events also fall into this category. Thus, in the vicinity of the subclusters D.1 and D.4, related to the seismic crisis, other subclusters containing seismic activity can be found.

5.3 Anthropogenic signals with high envelope correlation

After identifying seismic activity in cluster D, we want to draw attention to the remaining part of the seismic data set. Seismic activity induces short-term signals with a characteristic waveform and envelope shape. However, if we want to classify other types of signals like tremors, anthropogenic noise, or ambient noise, correlating waveforms are unlikely to be suitable for this task. One key feature of the deep scattering spectrum is the representation of the waveform's envelope in the second-order scattering coefficients (Andén & Mallat, 2014). Consequently, we should find clusters with weakly correlating waveforms but strongly correlating envelopes.

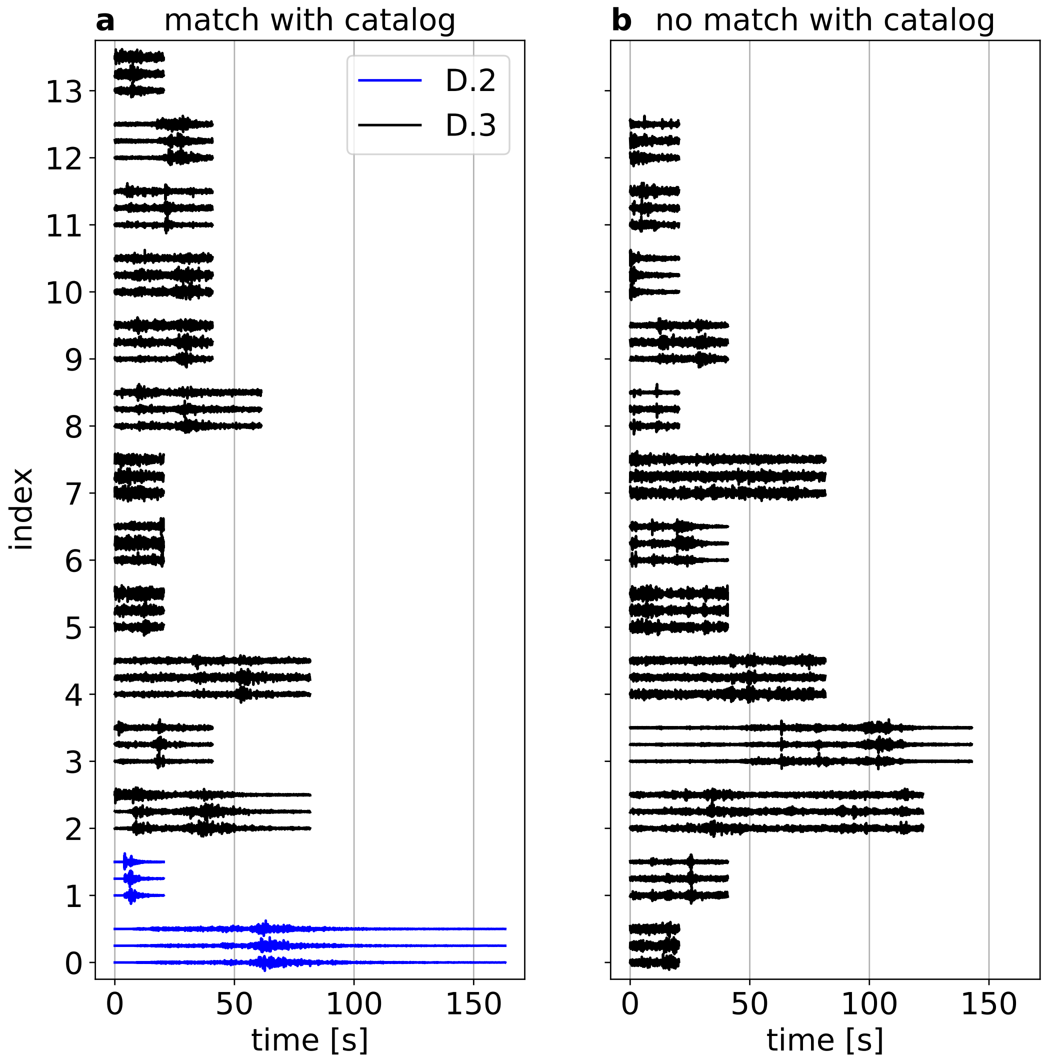

Figure 8. Seismic waveforms identified in subclusters D.2 and D.3. **(a)** waveform data of D.2 and D.3 where the phase arrivals match the merged catalog. **(b)** waveform data of D.3 which do not correspond to phase arrivals from the merged catalog.

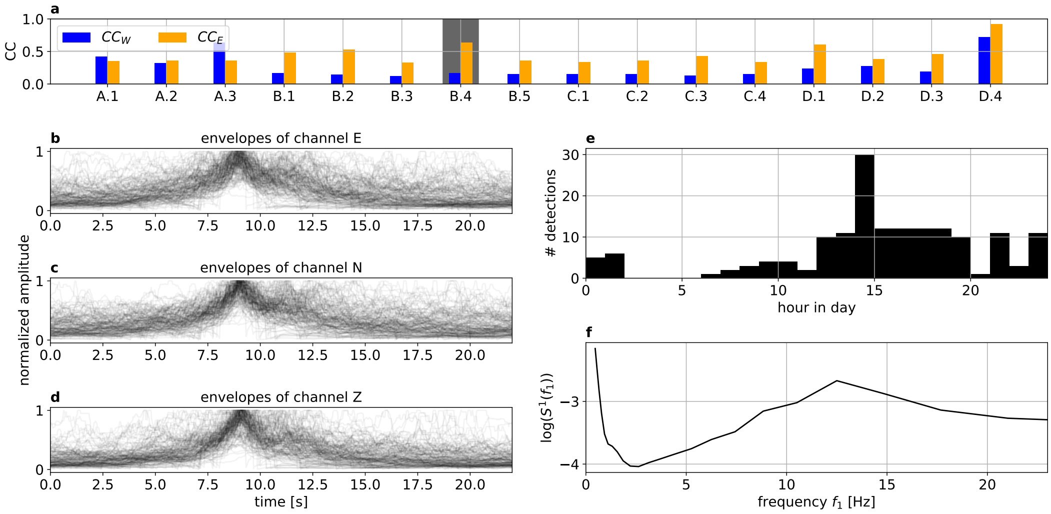

Figure 9. Interpretation of subcluster B.4. **(a)** Averaged correlation coefficient for the waveforms CC_W and for the envelopes CC_E for all 16 subclusters. **(b,c,d)** Aligned envelopes for the three channels for subcluster B.4. **(e)** Number of detections per hour for subcluster B.4. **(f)** Centroidal first-order scattering coefficients for subcluster B.4.

For that reason, we investigate the correlation coefficient of the waveform (CC_W) and the envelope (CC_E) for all subclusters. Firstly, a template is defined by the closest sample to the centroid representing the most typical waveform of a cluster. Then, we calculate the correlation coefficient of the waveform data CC_W and the correlation coefficient of the smoothed envelope CC_E between the template and the remaining samples. The envelope is defined by the modulus of the analytical signal. The averaged results are depicted in Figure 9a. We firstly observe that CC_E is more significant than CC_W for most subclusters. In particular, cluster B.4 shows the most significant discrepancy between CC_E and CC_W; this subcluster is part of cluster B, which we related to high-frequent urban noise. In Figure 9b to d, we align the envelopes for each channel and each sample in B.4 to depict the shared characteristics. We see a very symmetric envelope that lasts around 5 s. The envelopes look very similar on all three components. Figure 9e shows a histogram of detections over the time of the day. We see that this cluster mostly appears during daytime with a clear peak around 14:00 local time. Figure 9f shows the averaged first-order scattering coefficients for all three channels. The frequencies above 5 Hz are very pronounced and peak between 10 and 15 Hz. In summary, we see that subcluster B.4 is related to non stationary urban noise which produced similar envelopes lasting 5 s. Nearby road traffic could produce these kind of signals.

5.4 Long-lasting signals with low envelope correlation

As the last example, we want to draw attention towards clusters A and C. Both clusters show relatively low correlation coefficients for the envelopes (see Figure 9). Cluster C contains more than half of the data, and the average scattering coefficients are the lowest for all frequencies compared to the other clusters (see Figure 4b and d). Moreover, the subclusters of C have a relatively low distance to each other, and their within-cluster variance is relatively low (see Figure B1 in Appendix B). This indicates that they contain similar signals. Combining these facts, we conclude that this cluster contains ambient noise without any significant activity of transient signals.

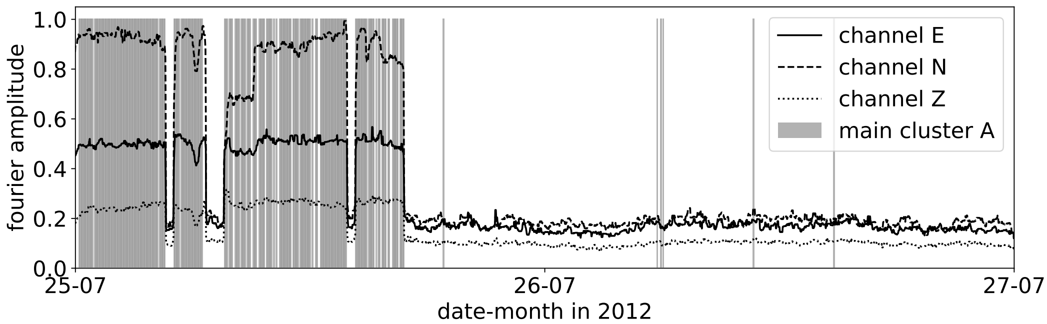

Figure 10. Fourier amplitude of all three channels calculated over 10 min windows in the frequency range of 1.4 to 1.6 Hz together with the activation of the main cluster A

Cluster A seems to correlate with the monochromatic noise source around 1.5 Hz (see Figure 3c and 4c). To prove that cluster A contains only data with increased activity around 1.5 Hz we depict the occurrence of cluster A and the Fourier amplitude of the three channels filtered between 1.4 and 1.6 Hz as a function of time in Figure 10. In general, an increased amplitude around 1.5 Hz correlates well with the appearance of cluster A. However, not all samples with an increased monochromatic activity fall into cluster A. This can be explained by the fact that a sample in the independent component space contains pooled information of ca. 20 s of waveform data which can contain many different signals. For example, if two different seismic data windows contain an increased monochromatic signal activity, but only one of the two windows also contains an earthquake or road traffic, the representation in the feature space will be different because of the pooling. Therefore, some samples with increased activity around 1.5 Hz will not fall into cluster A because other signals happening simultaneously will change their position in the independent component space. Moreover, it is interesting to note that subcluster A.1 and A.3 show larger correlation coefficients for the waveforms than for the envelopes (Figure 9a). This characteristic only applies to these two subclusters and is related to the dominance of the monochromatic signal.

Cluster A and C show that the dendrogram representation based on features from the deep scattering spectrum also finds cluster of noise sources without strong correlation of the waveforms or envelopes.

6 Conclusion

In this study, we proposed a new way of exploring seismic data hierarchically with a dendrogram based on features extracted from the deep scattering spectrum. A primary advantage of the workflow compared to other machine learning algorithms for classifying continuous seismic data is the interpretability at each step. For an application in this study, we chose a 2-day long data set containing a nearby seismic crisis with 148 cataloged events. These labels served as a sanity check for the algorithm.

Firstly, we calculated time-frequency features with the scattering network, decreasing the sampling period in time and increasing the number of dimensions. Due to the curse of dimensionality, we reduced the data into a ten-dimensional data space with ICA. The single independent components already revealed trends in the data set (see Appendix A). In the reduced data space, we created the dendrogram based on the Ward's distance between data points and clusters. The dendrogram was then used to navigate through the data set and explore areas of interest. This approach is very different from conventional clustering, where a certain number of clusters has to be defined beforehand. Here, the number of clusters changes with the depth of the dendrogram. This approach can retrieve different sized clusters, of which some would have been ignored by statistical analysis.

At a significant distance threshold, we extracted the four main clusters A, B, C, and D. With the cluster size, the temporal detection, and averaged first-order scattering coefficients, we delivered a rough interpretation of each cluster and obtained a rough overview of the entire data set. We identified cluster D as the cluster containing the seismic crisis. Inside cluster D, we found D.1 and D.4 containing 92.6 % of the seismic crisis. The main difference between the two subclusters is the magnitude of the events: D.4 contains events with a larger magnitude than D.1. 7.4 % (ten events) were found in subclusters of B and C due to poor signal-to-noise ratio or other significant amplitude signals in the pooling window. Here the problem is related to the pooling itself and the choice of similarity measure, which drives the iterative agglomeration. Nevertheless, we believe that Ward's method is an appropriate choice as a similarity measure for the agglomeration process, since it is adapted to the class imbalance within seismic data. Moreover, the misidentified ten events are outweighted by the 77 new events found in subcluster D.1 and D.4. The similarity of the waveforms suggests that they come from the same source area. The case of the seismic crisis has shown that we can identify a repeating pattern with slight variations of the waveforms in an unbalanced data set.

The other subclusters of D can also be primarily explained by seismic activity. D.2 is a minor outlier cluster containing a regional M4 event and a quarry blast from a nearby mine. 67 % of D.3 can be explained by a catalog containing local and regional events. These findings are very interesting when we talk about the meaning of neighborhood. Since we know that D.1 and D.4 contain the seismic crisis, we have reasons to assume that we can find similar types of signals (e.g., other types of earthquakes) in the neighborhood of these subclusters. However, we also need to keep in mind that subclusters from A, B, or C can also be in the vicinity of the subclusters D.1 and D.4. Further research needs to be done to understand better the meaning of neighborhood in this type of data representation.

At last, we also analyzed clusters that are not related to seismicity. B.4 contains samples with a low correlation coefficient for the waveform data but a high correlation coefficient for the envelopes. Here we found a characteristic envelope that was symmetric and lasted for 5 s. The traffic of a nearby road could be a possible source for this cluster. This case shows the possibility to detect patterns that do not share the same waveform but the same envelope. This is particularly interesting for the detection and classification of volcanic and tectonic tremors, which often show similar envelopes but no seismic phases. Moreover, we relate Cluster A to a monochromatic signal around 1.5 Hz and cluster C to the general ambient noise. These examples show that the workflow also finds clusters with low correlating waveforms and envelopes.

In general, the method can be used for various tasks. It is beneficial to get a general overview of an unknown data set. If there is a particular target of interest (e.g., earthquakes, urban noise sources, tremors), we can navigate the dendrogram and focus the analysis on a specific branch. The method can also be helpful to extract particular types of noise for performing ambient noise cross-correlation. We also believe that the dendrogram can reveal clusters/classes human expert knowledge could not reveal yet and expand the classes of signals we know so far.

Moreover, the analysis of the seismic data showed its multi-label characteristics. Multiple signals can arrive simultaneously and, thus, assigning a single label to a window does not reflect the whole truth. Integrating this issue into clustering seismic data is an interesting aspect for future work.

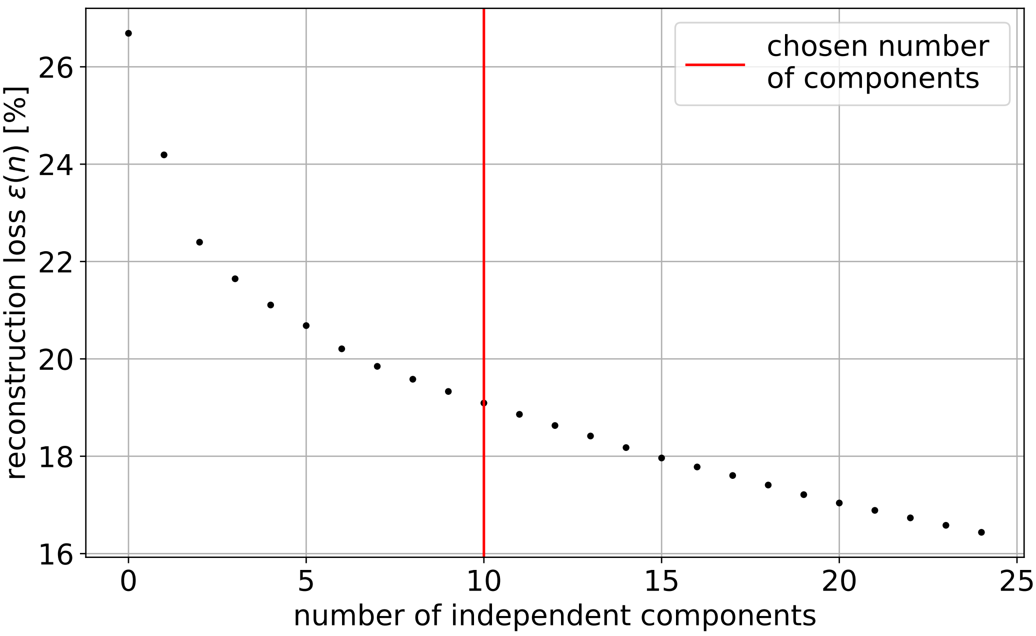

Figure A1. Reconstruction loss with independent component analysis from the deep scattering spectrum. The reconstruction loss $\epsilon(n)$ is calculated from Equation A1 as a function of the number of independent components n.

Appendix A Number of relevant independant components

For dimensionality reduction, we apply an independent component analysis using the `FastICA` algorithm from the `scikit-learn` Python library. Setting the number of dimensions in the reduced data space is always an exploratory task, and it is appropriate to estimate the information loss as a guideline for that. In this study, we use a reconstruction loss ϵ between the original data $\mathbf{x}$ and the reconstructed data $\hat{\mathbf{x}}^{(n)}$, obtained from Equation 2 with n independent components, as

$$\epsilon(n) = \frac{\sum_{i=0}^{N} |x_i - \hat{x}_i^{(n)}|}{N}. \tag{A1}$$

Figure A1 depicts the reconstruction loss $\epsilon(n)$ for an increasing number of independent components n. The reconstruction loss decreases rapidly with the first components. With a more significant number of components, the rate of error decrease becomes smaller. The choice of the number of dimensions in the reduced data space is a trade-off between keeping the dimensions low and retaining most of the information. Thus, ten independent components seem like a good compromise to us.

The time series of the ten independent components calculated from the data set are shown in Figure A2. To see if single components already show a clear distinction between the seismic crisis and the rest of the data, we marked in blue the samples containing at least one earthquake from the crisis. We see that all independent components show very different trends. For example the ninth independent component seems to separate the seismic crisis from the rest of the data. This observation raises the question if other trends, such as the background noise, can be correlated with specific independent components.

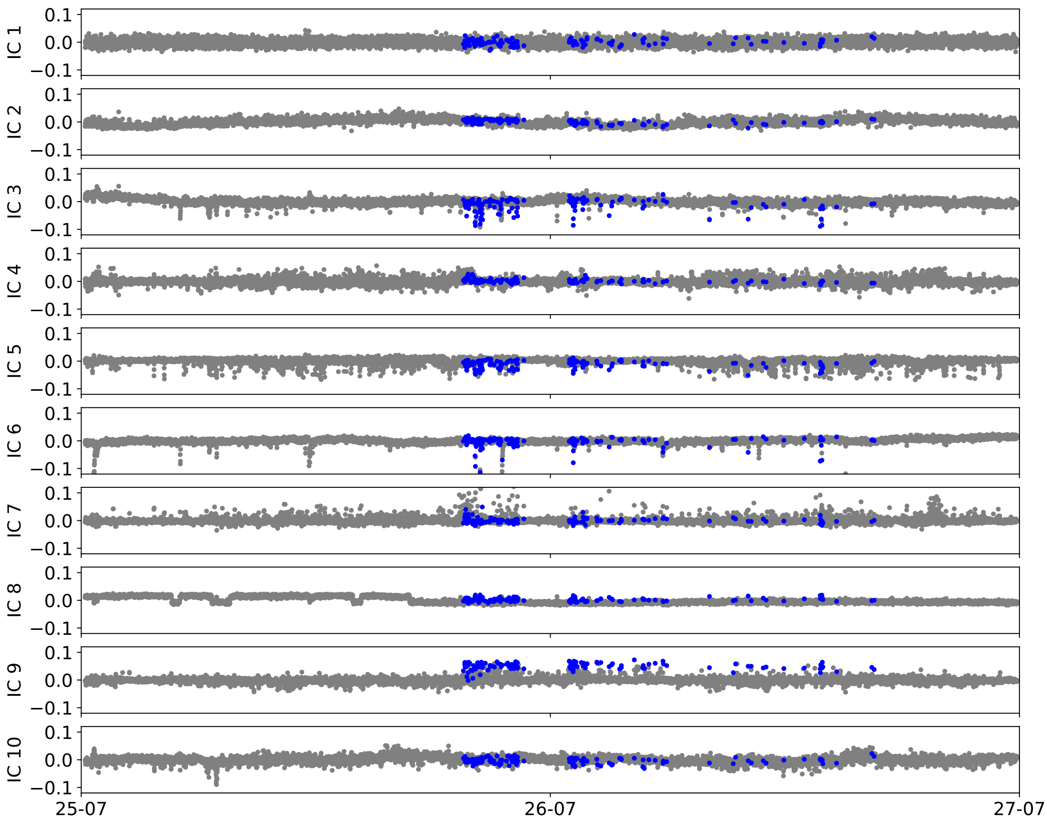

Figure A2. Time series of the ten independent components (IC) of the deep scattering spectrum for the overall seismic data set. The samples containing one or more arrivals of the earthquake from the nearby seismic crisis are highlighted with blue dots.

If we compare with the spectrogram of Figure 3c we see that the second independent component seems to correlate with the variations around 0.2 Hz and the eighth independent component seems to correlate with the monochromatic noise source around 1.5 Hz. This quick visual inspection shows us that the reduced data space can already be physically interpreted, and the ICA separates different signals on its different components, which is favorable for further analysis by clustering algorithms.

Appendix B Within-cluster variance and inter-cluster distance

This section presents the way we calculate the inter-cluster distance d_{ij} between clusters i and j and the within-cluster variance σ_i of cluster i. The inter-cluster distance are defined by the Euclidean distances between the centroids of the cluster:

$$d_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2, \tag{B1}$$

where $\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{n\in i}\hat{\mathbf{y}}_n$ represents the centroid of cluster i with the samples $\hat{\mathbf{y}}_n \in \mathbb{R}^C$ belonging to cluster i, and where $\|\cdot\|_2$ represents the $L2$ norm. Similarly, the variance σ_i of cluster i is defined as:

$$\sigma_i = \frac{1}{N_i}\sum_{n\in i}^{N_i}\|\hat{\mathbf{y}}_n - \boldsymbol{\mu}_i\|_2^2. \tag{B2}$$

This analysis is inspired from the silhouette analysis (Rousseeuw, 1987) and helps to understand better the clustering results. The within-cluster variances and the Euclidean distances between the centroids are depicted in Figure B1.

Appendix C Comparison with Single-station Template Matching

Station DC06 recorded higher signal-to-noise ratio S-waves from the seismicity crisis than the more proximal stations. Therefore, we are able to detect about twice more events by running the matched-filter search only on station DC06, with respect to the multi-station (ten stations) matched-filter search. The single-station template matching catalog captures a seismicity pattern similar to clusters D.1 and D.4, but reports about 50% more events (see Figure C1). Both the single-station and multi-station template matching catalogs were built with a detection threshold of eight times the root-mean-square of the correlation coefficient time series. The 20-second time resolution of the clustering method presented in this work sets a hard constraint on revealing the details of low magnitude seismicity. Nevertheless, we recall that producing a fine resolution earthquake catalog is not the first goal of our method, which instead aims at unraveling signals of different nature with no prior knowledge of the data set.

Acknowledgments
The authors acknowledge support from the European Research Council under the European Union Horizon 2020 research and innovation program (grant agreement no. 742335, F-IMAGE). This work has also been supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003). E.B. was also supported by funds associated with Robert D. van der Hilst's Schlumberger chair.

The facilities of IRIS Data Services, and specifically the IRIS Data Management Center, were used for access to waveforms, related metadata, and/or derived products used in this study. IRIS Data Services are funded through the Seismological Facilities for the Advancement of Geoscience (SAGE) Award of the National Science Foundation under Cooperative Support Agreement EAR-1851048. The data of the DANA array can be found at DANA (2012). The scattering network which was used in this study can be found at `https://github.com/leonard-seydoux/scatnet`. The python packages `ObsPy`, `SciPy` and `Scikit-learn` were heavily used for processing the data (Beyreuther et al., 2010; Virtanen

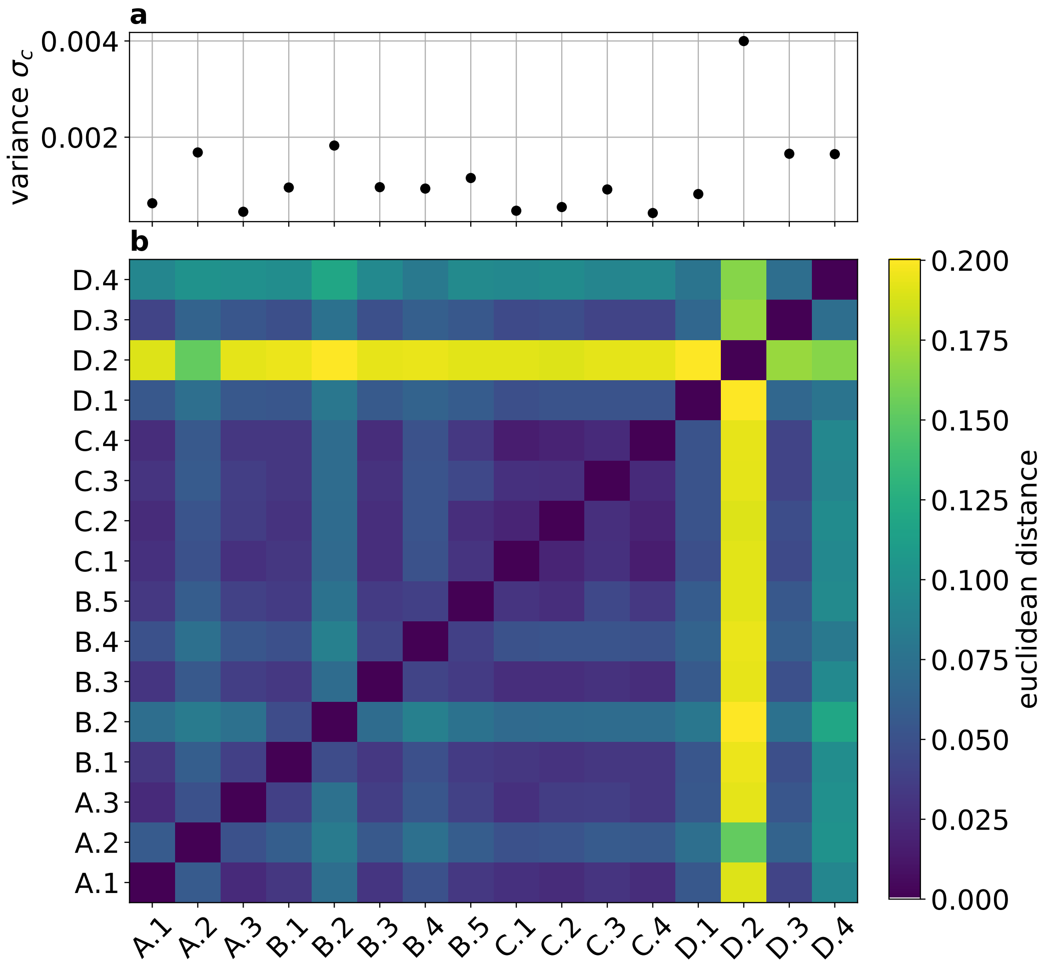

Figure B1. Inter-cluster distances and within-cluster variances. **(a)** Within-cluster variance according to equation B2 for all 16 subclusters. **(b)** Inter-cluster distance according to equation B1 between all 16 subclusters.

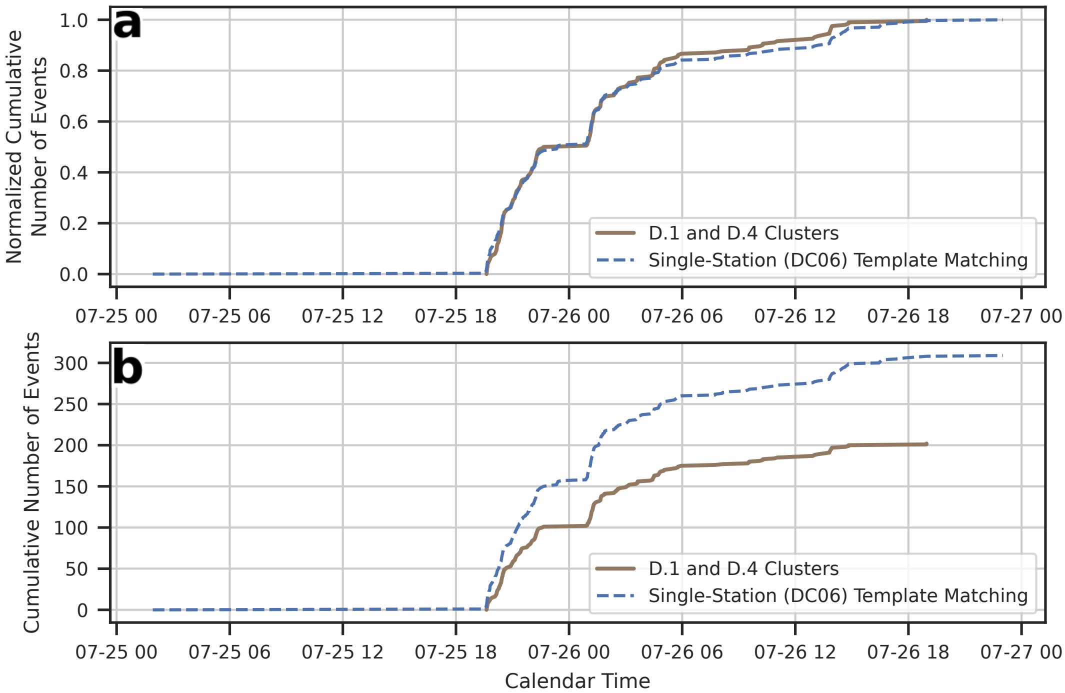

Figure C1. Comparison between the earthquake catalog from clusters D.1 and D.4 (thick brown line), and the single-station (DC06) template matching catalog (dashed blue line). **(a)** Normalized cumulative number of events. **(b)** Cumulative number of events. The single-station template matching catalog documents about 50% more events.

et al., 2020; Pedregosa et al., 2011). Maps were created with the python package `Cartopy` (Met Office, 2010 - 2015). We used map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

References

Andén, J., & Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal Processing*, *62*(16), 4114–4128.

Balestriero, R., Cosentino, R., Glotin, H., & Baraniuk, R. (2018). Spline filters for end-to-end deep learning. In *International conference on machine learning* (pp. 364–373).

Beaucé, E., Frank, W. B., Paul, A., Campillo, M., & van der Hilst, R. D. (2019). Systematic detection of clustered seismicity beneath the southwestern alps. *Journal of Geophysical Research: Solid Earth*, *124*(11), 11531–11548.

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid earth geoscience. *Science*, *363*(6433).

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International conference on database theory* (pp. 217–235).

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.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, *36*(3), 287–314.

Cosentino, R., & Aazhang, B. (2020). Learnable group transform for time-series. In *International conference on machine learning* (pp. 2164–2173).

DANA . (2012). *Dense array for north anatolia.* International Federation of Digital Seismograph Networks. Retrieved from http://www.fdsn.org/doi/10.7914/SN/YH_2012 doi: 10.7914/SN/YH_2012

Deparis, J., Jongmans, D., Cotton, F., Baillet, L., Thouvenot, F., & Hantz, D. (2008). Analysis of rock-fall and rock-fall avalanche seismograms in the french alps. *Bulletin of the Seismological Society of America*, *98*(4), 1781–1796.

Diaz, J. (2020). Church bells and ground motions. *Journal of Seismology*, 1–10.

Ebeling, C. W. (2012). Inferring ocean storm characteristics from ambient seismic noise: A historical perspective. In *Advances in geophysics* (Vol. 53, pp. 1–33). Elsevier.

Emre, Ö., Duman, T., Özalp, S., Elmaci, H., & Olgun, S. (2011). 1: 250,000 scale active fault map series of turkey. *Kayseri (NJ36-8) Quadrange, Ankara*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. (http://www.deeplearningbook.org)

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, *13*(4-5), 411–430.

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*(15), e2020GL088353.

Kennett, B., & Engdahl, E. (1991). Traveltimes for global earthquake location and phase identification. *Geophysical Journal International*, *105*(2), 429–465.

Kodera, Y., & Sakai, S. (2020). An unsupervised automatic classification for continuous seismic records: introducing an anomaly detection algorithm to solve the imbalanced data problem. In *Agu fall meeting 2020.*

Köhler, A., Ohrnberger, M., & Scherbaum, F. (2010). Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps. *Geophysical Journal International*, *182*(3), 1619–1630.

Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *Acm transactions on knowledge discovery from data (tkdd)*, *3*(1), 1–58.

Lacroix, P., & Helmstetter, A. (2011). Location of seismic signals associated with microearthquakes and rockfalls on the séchilienne landslide, french alps. *Bulletin of the Seismological Society of America*, *101*(1), 341–353.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lin, W.-C., Tsai, C.-F., Hu, Y.-H., & Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, *409*, 17–26.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017). Implementation of a multistation approach for automated event classification at piton de la fournaise volcano. *Seismological Research Letters*, *88*(3), 878–891.

Malfante, M., Dalla Mura, M., Mars, J. I., Métaxian, J.-P., Macedo, O., & Inza, A. (2018). Automatic classification of volcano seismic signatures. *Journal of Geophysical Research: Solid Earth*, *123*(12), 10–645.

Meng, H., & Ben-Zion, Y. (2018). Characteristics of airplanes and helicopters recorded by a dense seismic array near anza california. *Journal of Geophysical Research: Solid Earth*, *123*(6), 4783–4797.

Met Office. (2010 - 2015). Cartopy: a cartographic python library with a matplotlib interface [Computer software manual]. Exeter, Devon. Retrieved from http://scitools.org.uk/cartopy

Mousavi, S. M., & Beroza, G. C. (2019). Bayesian-deep-learning estimation of earthquake location from single-station observations. *arXiv preprint arXiv:1912.01144*.

Obara, K. (2002). Nonvolcanic deep tremor associated with subduction in southwest japan. *Science*, *296*(5573), 1679–1681.

Peddinti, V., Sainath, T., Maymon, S., Ramabhadran, B., Nahamoo, D., & Goel, V. (2014). Deep scattering spectrum with deep neural networks. In *2014 ieee international con-*

ference on acoustics, speech and signal processing (icassp) (pp. 210–214).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake detection and location. *Science Advances*, *4*(2), e1700578.

Poyraz, S. A., Teoman, M. U., Türkelli, N., Kahraman, M., Cambaz, D., Mutlu, A., ... others (2015). New constraints on micro-seismicity and stress state in the western part of the north anatolian fault zone: Observations from a dense seismic array. *Tectonophysics*, *656*, 190–201.

Riahi, N., & Gerstoft, P. (2015). The seismic traffic footprint: Tracking trains, aircraft, and cars seismically. *Geophysical Research Letters*, *42*(8), 2674–2681.

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.

Rouet-Leduc, B., Hulbert, C., McBrearty, I. W., & Johnson, P. A. (2020). Probing slow earthquakes with deep learning. *Geophysical research letters*, *47*(4), e2019GL085870.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Seydoux, L., Balestriero, R., Poli, P., De Hoop, M., Campillo, M., & Baraniuk, R. (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature communications*, *11*(1), 1–12.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.