



**HAL**  
open science

# Hierarchical exploration of continuous seismograms with unsupervised learning

Rene Steinmann, Leonard Seydoux, Eric Beaucé, Michel Campillo

► **To cite this version:**

Rene Steinmann, Leonard Seydoux, Eric Beaucé, Michel Campillo. Hierarchical exploration of continuous seismograms with unsupervised learning. 2021. hal-03241271v1

**HAL Id: hal-03241271**

**<https://hal.univ-grenoble-alpes.fr/hal-03241271v1>**

Preprint submitted on 28 May 2021 (v1), last revised 13 Jan 2022 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1           **Hierarchical exploration of continuous seismograms**  
2                                   **with unsupervised learning**

3           **René Steinmann<sup>1</sup>, Léonard Seydoux<sup>1</sup>, Éric Beaucé<sup>2</sup> and Michel Campillo<sup>1</sup>**

4           <sup>1</sup>ISTerre, équipe Ondes et Structures, Université Grenoble-Alpes, UMR CNRS 5375, 1381 Rue de la  
5                                   Piscine, 38610, Gières, France

6           <sup>2</sup>Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology,  
7                                   Cambridge, MA, USA

8           **Key Points:**

- 9           • Seismic data analysis  
10          • Unsupervised learning  
11          • Seismic waveform clustering

---

Corresponding author: René Steinmann, [rene.steinmann@univ-grenoble-alpes.fr](mailto:rene.steinmann@univ-grenoble-alpes.fr)

12 **Abstract**

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

29 **Plain Language Summary**

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

41 **1 Introduction**

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

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

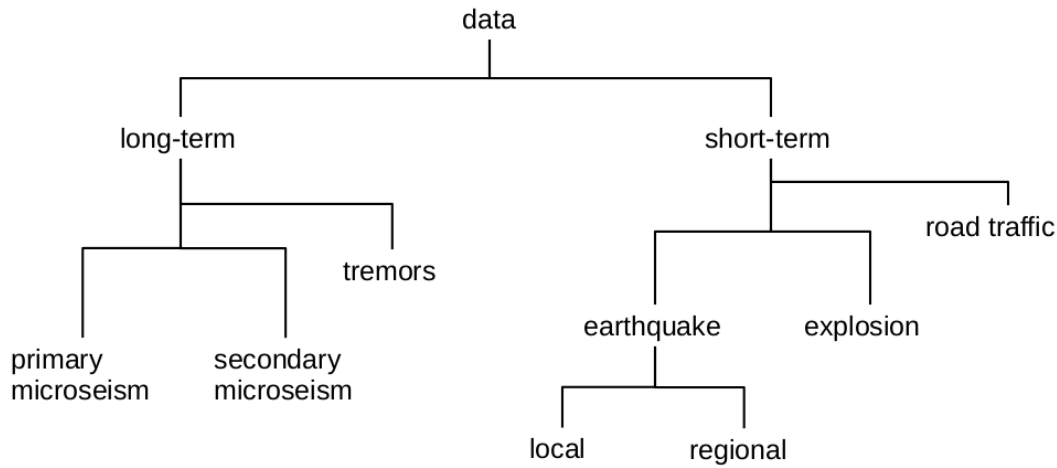
61 and can be done thanks to prior knowledge on the data or by defining proper algorithms to  
 62 learn the most relevant features. We distinguish classical machine-learning algorithms that  
 63 rely on human-defined features (Maggi et al., 2017; Malfante et al., 2018) or representation-  
 64 learning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

71 Once the features are defined, a model is trained to map the features to a certain output.  
 72 In supervised learning the model learns that mapping based on a labeled training data set  
 73 (Goodfellow et al., 2016). For instance, locating earthquakes with a single seismic station is  
 74 a non-trivial task to address with supervised learning. It has been recently shown that deep  
 75 neural networks can infer the position of earthquakes from single-station records (Perol et al.,  
 76 2018; Mousavi & Beroza, 2019). This experiment illustrates the range of applications that  
 77 supervised machine learning can find in the domain of seismology. Still, the limitations of  
 78 supervised learning lie within the ones of our expert knowledge. Indeed, supervised models  
 79 can only automatize a given task and therefore strongly rely on the quality of the labels  
 80 of the training data set. Moreover, supervised learning is limited to the labels we know  
 81 and cannot search for signatures or patterns with unknown properties in a data-exploration  
 82 fashion.

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

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

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



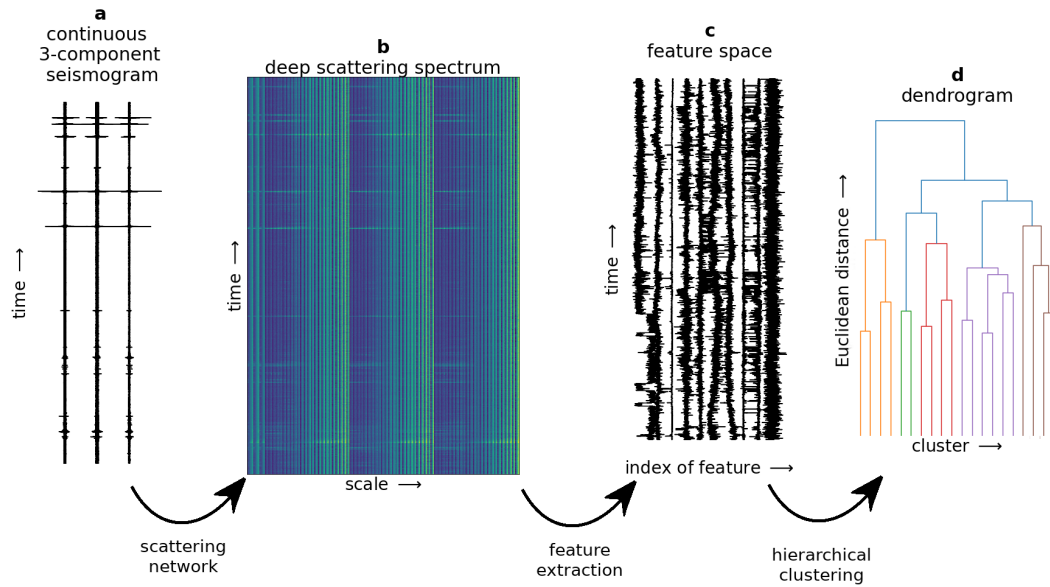
**Figure 1.** Illustration of possible hierarchy in seismic data. The different branches represent how a signal class splits into different subclasses depending on a given similarity measure. Here the different classes of events are thought in a hierarchical way, based on arbitrary signals properties (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.

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

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

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

139 In the next section, we present the workflow to build a dendrogram from continuous  
 140 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.

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

143 **2 Method**

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

155 **2.1 Deep scattering spectrum**

156 In this study, we aim to explore the data in an unsupervised and unbiased way, i.e.,  
 157 we want to assume as little as possible about the data themselves. For that purpose, it is  
 158 crucial to find a representation that is not a human-based selection of features. A time-  
 159 frequency representation such as the spectrogram is one way to create a set of features  
 160 without favoring any data characteristics. However, Andén and Mallat (2014) showed that  
 161 a spectrogram generated by the Fourier transform is not ideal for classification purposes since  
 162 it is not stable to time-warping deformations, especially at short periods compared with the

163 duration of the analyzing window. They introduce another time-frequency representation  
 164 called a deep scattering spectrum which is computed by a scattering network. This type of  
 165 network implements a cascade of convolutions with wavelet filters, modulus function, and  
 166 pooling operations. Deep scattering spectra are locally translation invariant and preserve  
 167 transient phenomena such as attack and amplitude modulation. These characteristics are  
 168 beneficial when it comes to classifying any time series data. In Andén and Mallat (2014) and  
 169 Peddinti et al. (2014), the authors have successfully classified audio data based on the deep  
 170 scattering spectrum. Seydoux et al. (2020) have brought that representation into seismology  
 171 and showed that small precursory signals of a landslide could be detected and classified in an  
 172 unsupervised fashion. Other successful deep-learning classifiers inspired by deep scattering  
 173 networks are presented in Balestrierio et al. (2018) and Cosentino and Aazhang (2020).

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

$$178 \quad S^{(\ell)} \left( t, f_{n_1}^{(1)}, f_{n_2}^{(2)}, \dots, f_{n_\ell}^{(\ell)} \right) = \max_{[t, t+dt]} \left| \phi^{(\ell)} \left( f_{n_\ell}^{(\ell)} \right) \star \left| \dots \star \left| \phi^{(2)} \left( f_{n_2}^{(2)} \right) \star \left| \phi^{(1)} \left( f_{n_1}^{(1)} \right) \star x \right| \right| \right|, \quad (1)$$

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

## 194 2.2 Features extraction from deep scattering spectrum

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

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

$$207 \quad \mathbf{x} = \mathbf{A}\mathbf{s}, \quad (2)$$

208 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  $C$  inde-  
 209 pendent components of the same dimension  $F$  as the observations and  $\mathbf{A} \in \mathbb{R}^{C \times N}$  is the



210 mixing matrix (that is, the mixing of the  $C$  independent components for every observation).  
 211 The observations are therefore a linear combination of the independent components. A test  
 212 of statistical independence is required to solve Equation 2 while ensuring the components  $s$   
 213 to be independent. Among the different strategies, we can look for a minimum of mutual  
 214 information, or similarly, a maximization of the non-Gaussianity. In our study, we apply  
 215 the `FastICA` algorithm from the `scikit-learn` Python library, which uses the negentropy  
 216 as a measure of non-Gaussianity (Hyvärinen & Oja, 2000). This analysis is similar to the  
 217 principal component analysis, with the difference that the independent components are not  
 218 orthogonal. In addition, there is no information about the variance explained by the different  
 219 independent components, and are therefore delivered unsorted by the algorithm.

## 220 **2.3 Hierarchical clustering**

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

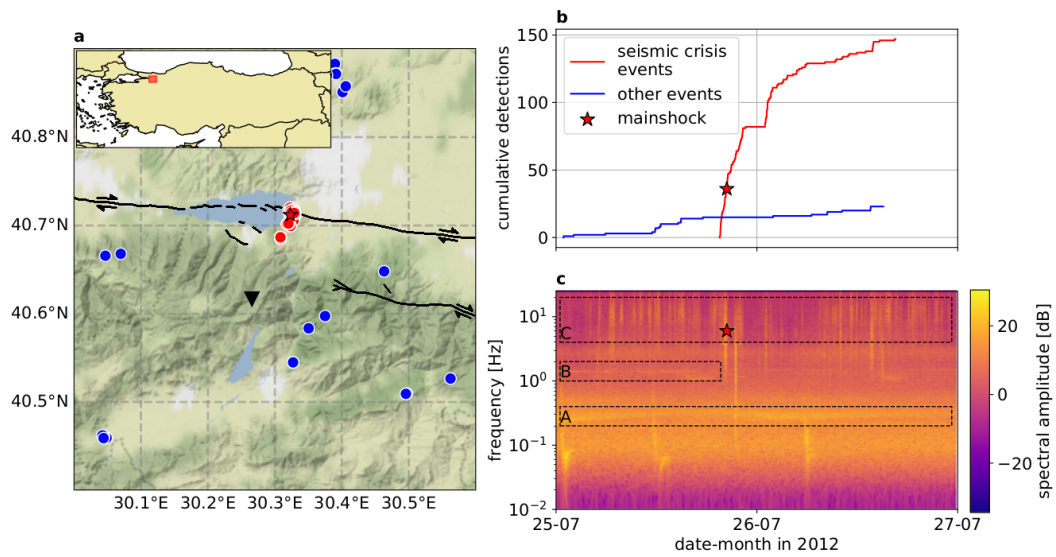
232 This study investigates how the data cluster in a hierarchical way with a bottom-up  
 233 approach, namely agglomerative clustering. At first, we briefly introduce the methodology.  
 234 Hierarchical clustering relies on a similarity matrix, which defines the similarity (e.g., a  
 235 specific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

241 The agglomerative clustering outcome depends mainly on the applied metric, which  
 242 drives the merging of the cluster. In our approach, we use the Ward’s method (Ward Jr,  
 243 1963). Given a distance  $d$  (here considered Euclidean), the Ward’s method aims at grouping  
 244 data samples  $x_i$  into clusters such as the within-cluster variance remains minimal after  
 245 merging different clusters. The within-cluster variance quantifies the spread  $\sigma$  of each within-  
 246 cluster data samples defined in Appendix B. By minimizing the overall variance,  $\sum_{c=1}^K \sigma_c$ ,  
 247 the Ward’s method allows for data clusters of variable population sizes and variances and  
 248 may highlight clusters of high density located in the vicinity of more spread, low-density  
 249 clusters. Therefore, Ward’s method is suitable for the expected seismic data partition.

## 250 **3 Data**

251 We test our proposed workflow on continuous three-component seismic data from the  
 252 station DC06 of the DANA experiment in Turkey (see for instance Poyraz et al., 2015, and  
 253 the map shown in Figure 3a). The sampling rate of the data is 50 Hz. We choose the data  
 254 set for mainly two reasons. First of all, the data set contains both seismic and anthropogenic  
 255 activity, which is a typical situation in most seismological studies. Second of all, an existing  
 256 template matching catalog provides labels for the seismicity in this area. The catalog was  
 257 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.

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

264 The spectrogram of the east component of station DC06 is presented in Figure 3c. The  
 265 oceanic microseism is visible around 0.2 Hz, where we can observe the dispersive nature of  
 266 the oceanic gravity waves. At around 1.5 Hz we can identify a nonstationary monochromatic  
 267 noise source, which seems to be more active during the first day. At frequencies higher than  
 268 3 Hz we can see increased activity during daytime, most likely induced by anthropogenic  
 269 noise sources. The main shock of the crisis during the evening of the 25th is also easy to  
 270 spot in the spectrogram.

## 271 4 Results

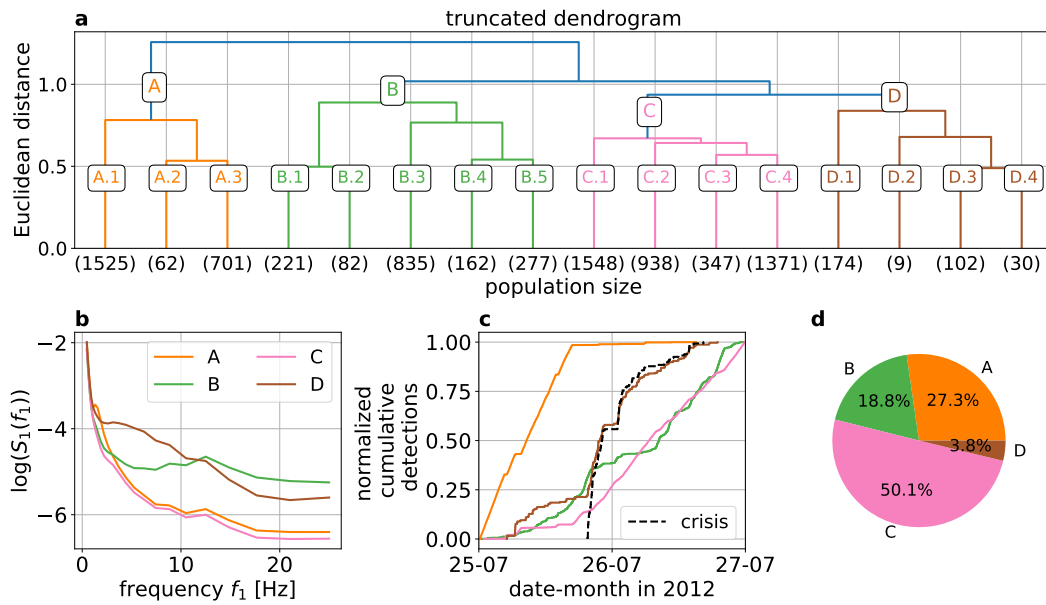
### 272 4.1 Feature space

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

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

### 300 4.2 Dendrogram

301 After transforming the continuous seismic data into a most relevant set of features, we  
 302 can use this representation to explore the data with hierarchical clustering. By controlling  
 303 the distance threshold, we can extract different numbers of clusters. In Figure 4a we selected  
 304 a distance threshold of 0.47 in order to show a truncated dendrogram stopping at 16 clusters.  
 305 At a distance of 0.9, we extract four main clusters labeled as A, B, C, and D. Figure 4b  
 306 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

328 an almost linear increase starting at the afternoon of the first day, exactly when cluster A  
 329 becomes almost inactive. The cluster size and frequency content suggest that cluster C is  
 330 related to samples containing only ambient noise.

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

## 337 5 Discussion

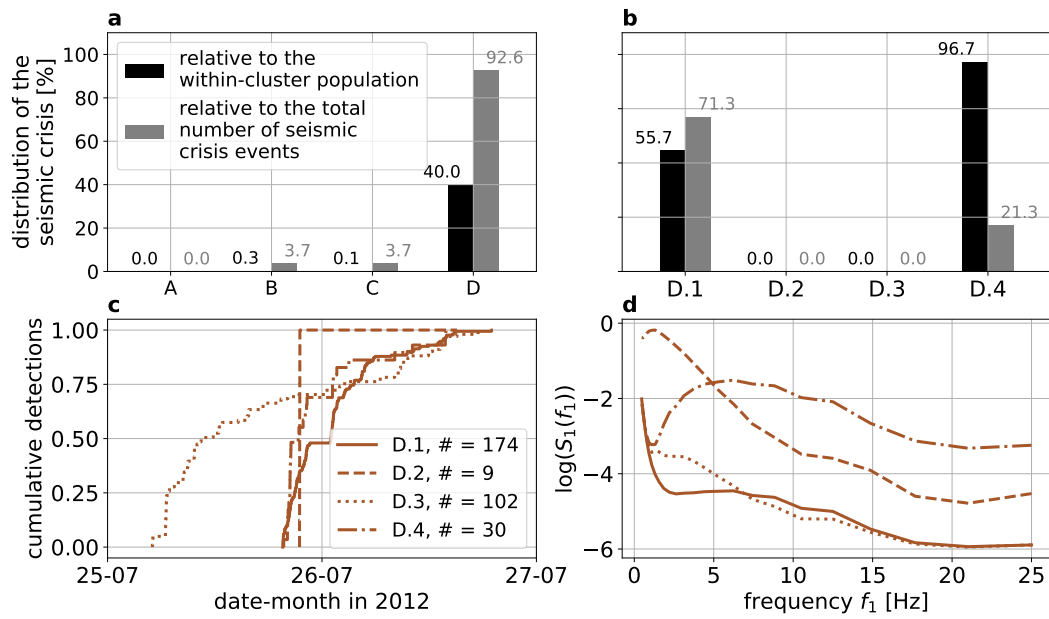
338 In this section, we will discuss and interpret the dendrogram's representation and its  
 339 clustering solution. While the main focus is on identifying how the seismic crisis occurs  
 340 in the dendrogram, we will also discuss how the general seismicity is observed through  
 341 this representation,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

### 343 5.1 Identification of the seismic crisis within the dendrogram

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

360 Firstly, we look at the distribution of the samples containing the seismic crisis across  
 361 the four subclusters in main cluster D. From Figure 5a, we know that more than 92 % of  
 362 the crisis was found in cluster D. We observe in Figure 5b that this amount splits into ca.  
 363 71.3 % in cluster D.1 and ca. 21.3 % in cluster D.4. The subclusters D.2 and D.3 contain  
 364 no earthquakes from the seismic crisis and will be discussed later. If we look at the cumu-  
 365 lative detection curve of each subcluster in D (Figure 5c), we see that cluster D.1 and D.4  
 366 share a very similar temporal pattern. The corresponding centroidal first-order scattering  
 367 coefficients (Figure 5d) explain why the crisis got split into two clusters: across almost all  
 368 frequencies the larger subcluster D.1 shows significantly smaller scattering coefficients than  
 369 the smaller subcluster D.4. Hence, the magnitudes of the events seem to be the character-  
 370 istics that separates the crisis into two clusters. Besides, we observe that 56 % of D.1 and  
 371 97 % of D.4 can be explained by the cataloged crisis. This observation raises the question:  
 372 what are the samples in D.1 and D.4 that cannot be related to the seismic crisis recorded  
 373 by the catalog? We can answer this question by looking at the waveforms representing the  
 374 corresponding data points of subclusters D.1 and D.4.

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



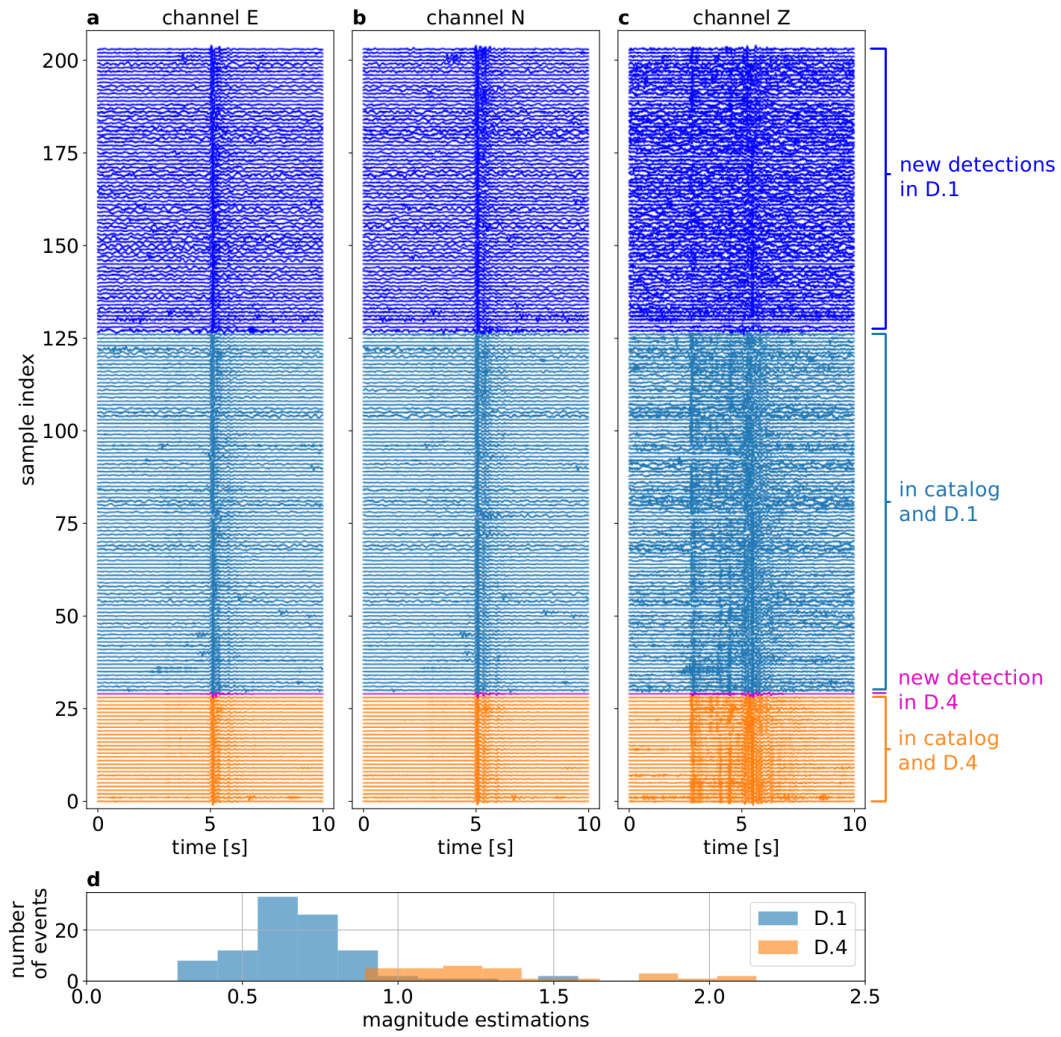
**Figure 5.** Identification of the seismic crisis within the main and subclusters. (a) The distribution of the seismic crisis across the four main clusters. (b) The distribution of the seismic crisis across the four subclusters in the main cluster D. (c) Normalized cumulative detection curves for the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

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

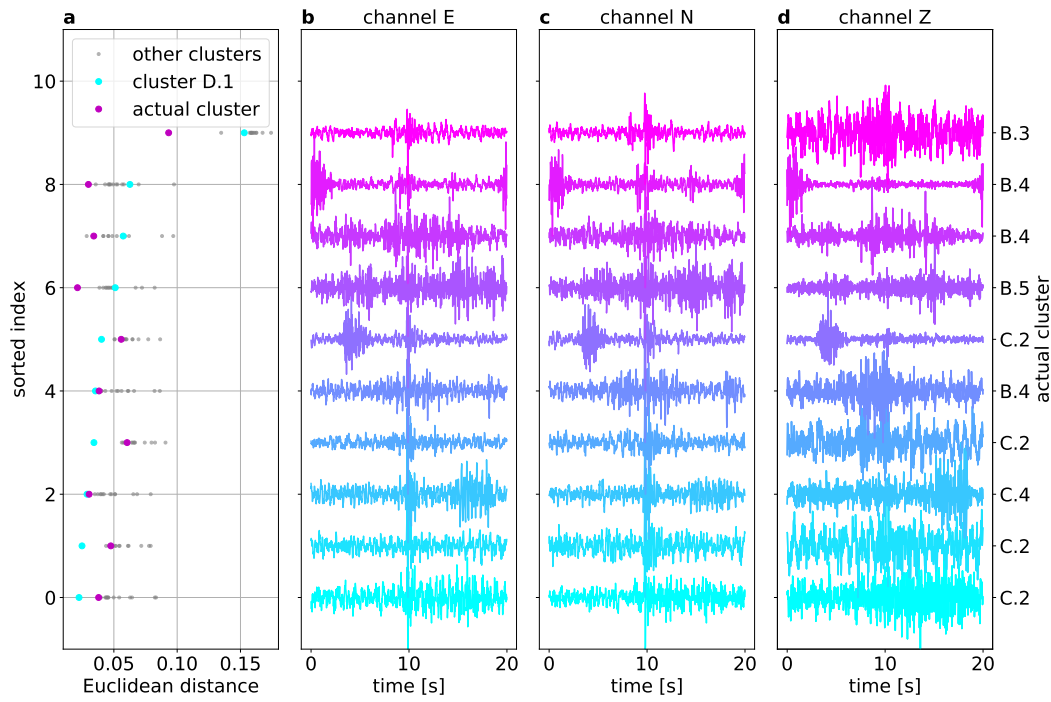
388 By investigating cluster D and its subclusters D.1 and D.4, we are able to identify two  
 389 subclusters representing the seismic crisis. While D.1 contains many events with smaller  
 390 magnitudes, D.4 contains fewer events with larger magnitudes. Together the two subclusters  
 391 contain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

396 However, 7.4% of the cataloged detections can not be found in subclusters D.1 or D.4.  
 397 In the following remarks, we want to analyze the misidentified 7.4% of cataloged events,  
 398 which equal ten over 135 events. First of all, we want to know where these events are  
 399 located in the feature space. Therefore, we calculate the Euclidean distance between the  
 400 misidentified events and the centroids of each cluster in the feature space (see Figure 7a).  
 401 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.

402 cyan, we highlight the distance between the sample and subcluster D.1 containing the low  
 403 magnitude events of the crisis. In gray, we highlight the distances to all other remaining  
 404 clusters as a comparison. We sorted the misidentified ten events according to the distance  
 405 to the centroid of D.1. We see that for the first six events, the distance to the centroid of D.1 is  
 406 smaller than to the centroid of its respective cluster. The corresponding waveform data also  
 407 offer explanations for the misidentification (Figure 7b to d). Indeed, the *P* and *S* arrivals  
 408 are noisy but visible for the first five events. Thus, some events might be misclassified  
 409 because samples are grouped with the Ward’s method, which solves iteratively an objective  
 410 function considering the Euclidean distance and the within-cluster variance. In other words,  
 411 clusters can agglomerate samples which might be closer to the centroids of other clusters  
 412 if we consider the pure Euclidean distance. After the first five events, when the distance  
 413 to its respective cluster becomes smaller than the distance to D.1., the *P* and *S* arrivals  
 414 are not visible anymore, or other large-amplitude events are present. Here the problem is  
 415 related to the representation of the data as a deep scattering spectrum or in the feature  
 416 space. Other large-amplitude transients can corrupt the representation since we perform a  
 417 maximum pooling to extract the scattering coefficients. This is not a specific problem of  
 418 maximum pooling but pooling in general since this operation reduces information in the  
 419 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.

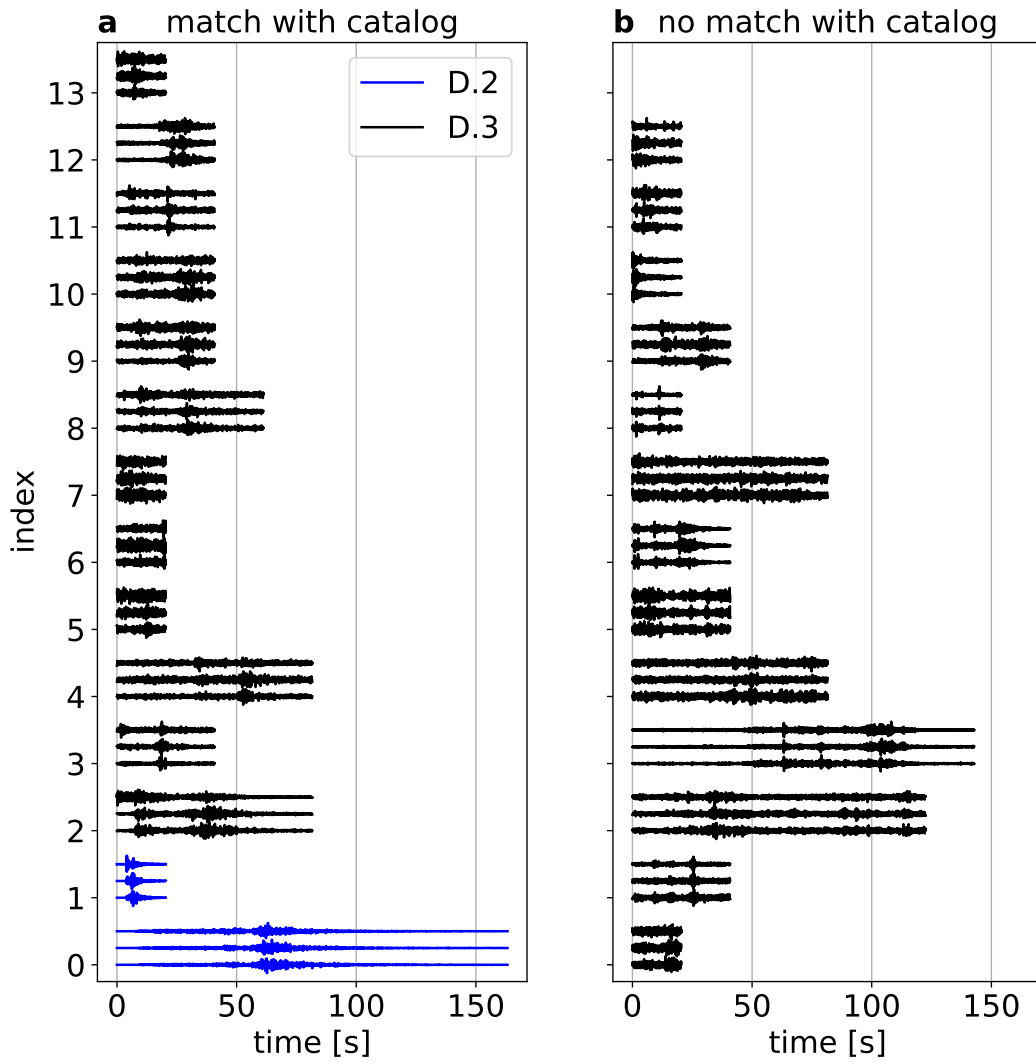
To verify this assumption, we can compare existing earthquake catalogs with the timestamps of the samples in the subclusters. We extend the local template matching catalog with a regional catalog limited to events within a radius of  $5^\circ$  around station DC06. The regional catalog is downloaded 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 20s 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. 20s 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  $\sigma_c$  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.

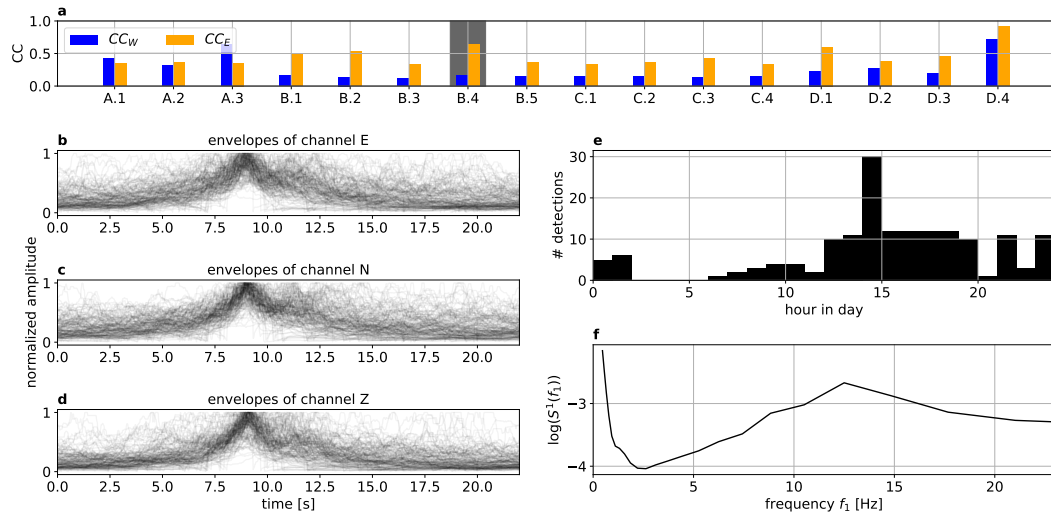
Moreover, we observe that the catalog can explain 67% of all samples of D.3. However, we only show some waveforms in black in Figure 8a. The other 33% are shown in Figure 8b, and some samples also show seismic phase arrivals (in particular, the seismograms shown at index six and nine). It is thus likely 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.

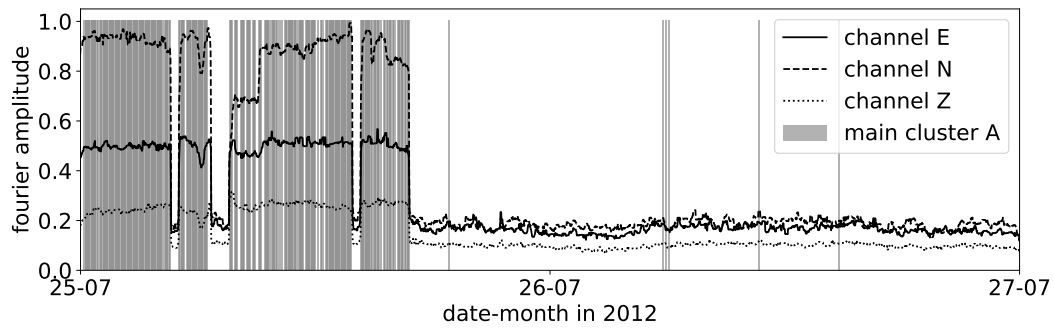


**Figure 9.** Interpretation of subcluster B.4. (a) Averaged correlation coefficient for the waveforms  $CC_W$  and for the envelopes  $CC_E$  for all 16 subclusters. (b,c,d) Aligned envelopes for the three channels for subcluster B.4. (e) Number of detections per hour for subcluster B.4. (f) Centroidal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

471 For that reason, we investigate the correlation coefficient of the waveform ( $CC_W$ ) and  
 472 the envelope ( $CC_E$ ) for all subclusters. Firstly, a template is defined by the closest sample  
 473 to the centroid representing the most typical waveform of a cluster. Then, we calculate  
 474 the correlation coefficient of the waveform data  $CC_W$  and the correlation coefficient of the  
 475 smoothed envelope  $CC_E$  between the template and the remaining samples. The envelope is  
 476 defined by the modulus of the analytical signal. The averaged results are depicted in Fig-  
 477 ure 9a. We firstly observe that  $CC_E$  is more significant than  $CC_W$  for most subclusters. In  
 478 particular, cluster B.4 shows the most significant discrepancy between  $CC_E$  and  $CC_W$ ; this  
 479 subcluster is part of cluster B, which we related to high-frequent urban noise. In Figure 9b  
 480 to d, we align the envelopes for each channel and each sample in B.4 to depict the shared  
 481 characteristics. We see a very symmetric envelope that lasts around 5s. The envelopes  
 482 look very similar on all three components. Figure 9e shows a histogram of detections over  
 483 the time of the day. We see that this cluster mostly appears during daytime with a clear  
 484 peak around 14:00 local time. Figure 9f shows the averaged first-order scattering coefficients  
 485 for all three channels. The frequencies above 5 Hz are very pronounced and peak between  
 486 10 and 15 Hz. In summary, we see that subcluster B.4 is related to non stationary urban  
 487 noise which produced similar envelopes lasting 5 s. Nearby road traffic could produce these  
 488 kind of signals.

#### 489 5.4 Long-lasting signals with low envelope correlation

490 As the last example, we want to draw attention towards clusters A and C. Both clusters  
 491 show relatively low correlation coefficients for the envelopes (see Figure 9). Cluster C  
 492 contains more than half of the data, and the average scattering coefficients are the lowest  
 493 for all frequencies compared to the other clusters (see Figure 4b and d). Moreover, the  
 494 subclusters of C have a relatively low distance to each other, and their within-cluster variance  
 495 is relatively low (see Figure B1 in Appendix B). This indicates that they contain similar  
 496 signals. Combining these facts, we conclude that this cluster contains ambient noise without  
 497 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

498 Cluster A seems to correlate with the monochromatic noise source around 1.5 Hz (see  
 499 Figure 3c and 4c). To prove that cluster A contains only data with increased activity around  
 500 1.5 Hz we depict the occurrence of cluster A and the Fourier amplitude of the three channels  
 501 filtered between 1.4 and 1.6 Hz as a function of time in Figure 10. In general, an increased  
 502 amplitude around 1.5 Hz correlates well with the appearance of cluster A. However, not all  
 503 samples with an increased monochromatic activity fall into cluster A. This can be explained  
 504 by the fact that a sample in the independent component space contains pooled information  
 505 of ca. 20 s of waveform data which can contain many different signals. For example, if two  
 506 different seismic data windows contain an increased monochromatic signal activity, but only  
 507 one of the two windows also contains an earthquake or road traffic, the representation in  
 508 the feature space will be different because of the pooling. Therefore, some samples with  
 509 increased activity around 1.5 Hz will not fall into cluster A because other signals happening  
 510 simultaneously will change their position in the independent component space. Moreover,  
 511 it is interesting to note that subcluster A.1 and A.3 show larger correlation coefficients for  
 512 the waveforms than for the envelopes (Figure 9a). This characteristic only applies to these  
 513 two subclusters and is related to the dominance of the monochromatic signal.

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

## 517 6 Conclusion

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

524 Firstly, we calculated time-frequency features with the scattering network, decreasing  
 525 the sampling period in time and increasing the number of dimensions. Due to the curse  
 526 of dimensionality, we reduced the data into a ten-dimensional data space with ICA. The  
 527 single independent components already revealed trends in the data set (see Appendix A). In  
 528 the reduced data space, we created the dendrogram based on the Ward's distance between  
 529 data points and clusters. The dendrogram was then used to navigate through the data set  
 530 and explore areas of interest. This approach is very different from conventional clustering,  
 531 where a certain number of clusters has to be defined beforehand. Here, the number of

532 clusters changes with the depth of the dendrogram. This approach can retrieve different  
 533 sized clusters, of which some would have been ignored by statistical analysis.

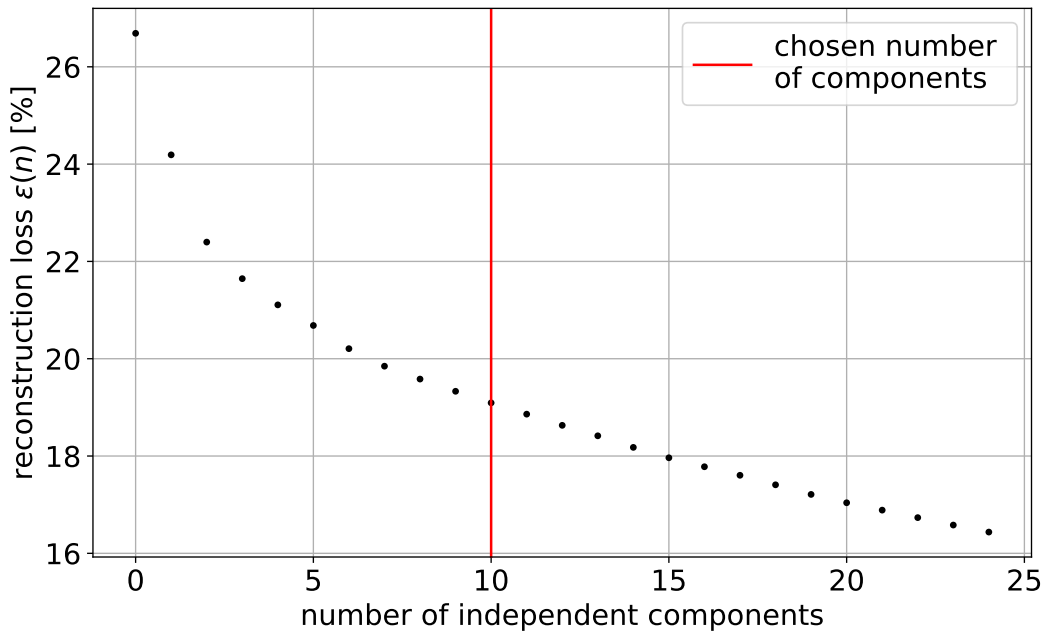
534 At a significant distance threshold, we extracted the four main clusters A, B, C, and D.  
 535 With the cluster size, the temporal detection, and averaged first-order scattering coefficients,  
 536 we delivered a rough interpretation of each cluster and obtained a rough overview of  
 537 the entire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

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

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

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

576 Moreover, the analysis of the seismic data showed its multi-label characteristics. Multi-  
 577 ple signals can arrive simultaneously and, thus, assigning a single label to a window does not  
 578 reflect the whole truth. Integrating this issue into clustering seismic data is an interesting  
 579 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  $n$ .

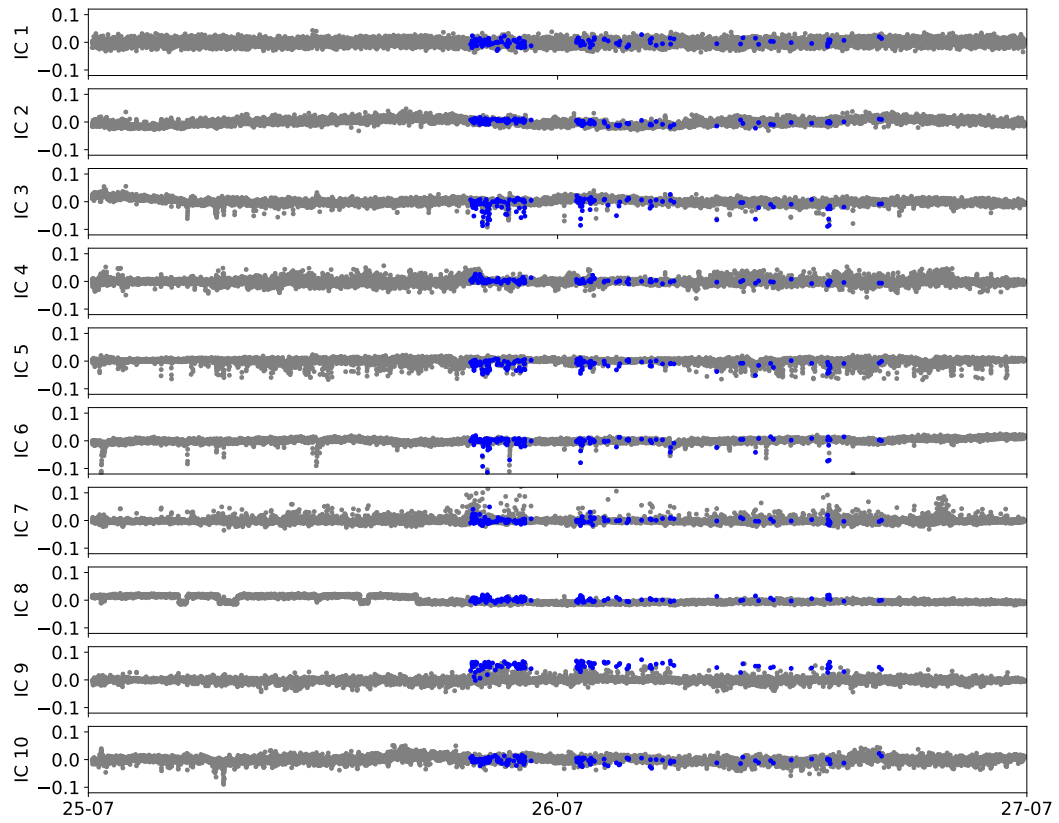
580 **Appendix A Number of relevant independent components**

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

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

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

593 The time series of the ten independent components calculated from the data set are  
 594 shown in Figure A2. To see if single components already show a clear distinction between  
 595 the seismic crisis and the rest of the data, we marked in blue the samples containing at least  
 596 one earthquake from the crisis. We see that all independent components show very different  
 597 trends. For example the ninth independent component seems to separate the seismic crisis  
 598 from the rest of the data. This observation raises the question if other trends, such as the  
 599 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.



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

## 606 Appendix B Within-cluster variance and inter-cluster distance

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

$$610 \quad d_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2, \quad (\text{B1})$$

611 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  $\hat{\mathbf{y}}_n \in \mathbb{R}^C$   
 612 belonging to cluster  $i$ , and where  $\|\cdot\|_2$  represents the  $L2$  norm. Similarly, the variance  $\sigma_i$   
 613 of cluster  $i$  is defined as:

$$614 \quad \sigma_i = \frac{1}{N_i} \sum_{n \in i} \|\hat{\mathbf{y}}_n - \boldsymbol{\mu}_i\|_2^2. \quad (\text{B2})$$

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

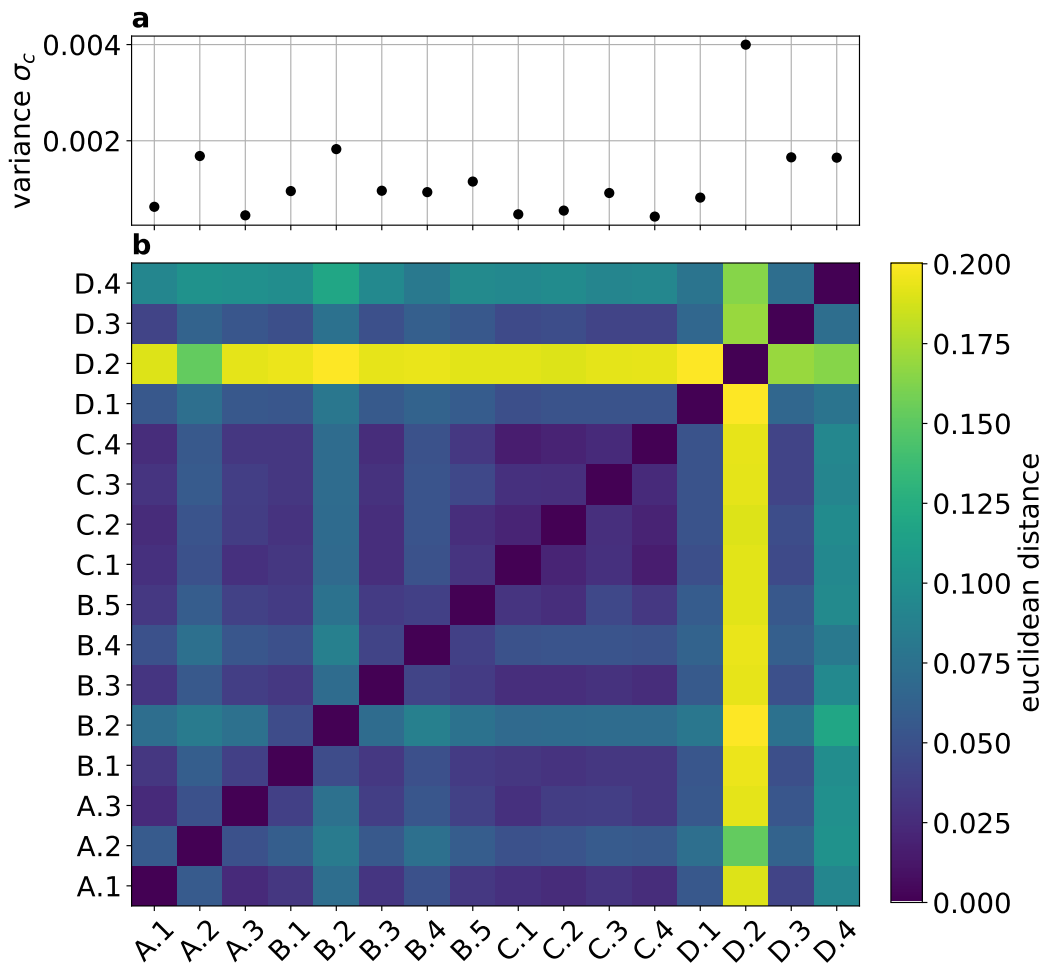
## 618 Appendix C Comparison with Single-station Template Matching

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

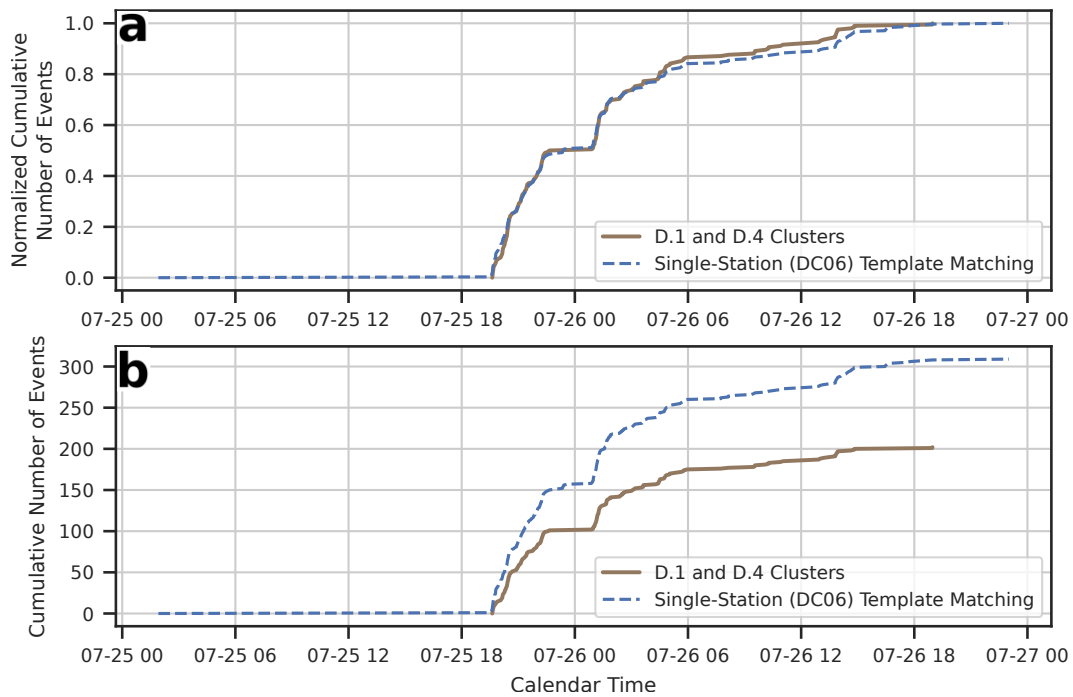
## 631 Acknowledgments

632 The authors acknowledge support from the European Research Council under the Euro-  
 633 pean Union Horizon 2020 research and innovation program (grant agreement no. 742335,  
 634 F-IMAGE). This work has also been supported by MIAI@Grenoble Alpes (ANR-19-P3IA-  
 635 0003). E.B. was also supported by funds associated with Robert D. van der Hilst's Schlum-  
 636 berger chair.

637 The facilities of IRIS Data Services, and specifically the IRIS Data Management Cen-  
 638 ter, were used for access to waveforms, related metadata, and/or derived products used  
 639 in this study. IRIS Data Services are funded through the Seismological Facilities for the  
 640 Advancement of Geoscience (SAGE) Award of the National Science Foundation under Co-  
 641 operative Support Agreement EAR-1851048. The data of the DANA array can be found  
 642 at DANA (2012). The scattering network which was used in this study can be found at  
 643 <https://github.com/leonard-seydoux/scatnet>. The python packages `ObsPy`, `SciPy` and  
 644 `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.

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

648 **References**

649 Andén, J., & Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal*  
 650 *Processing*, *62*(16), 4114–4128.  
 651 Balestrieri, R., Cosentino, R., Glotin, H., & Baraniuk, R. (2018). Spline filters for end-to-  
 652 end deep learning. In *International conference on machine learning* (pp. 364–373).  
 653 Beaucé, E., Frank, W. B., Paul, A., Campillo, M., & van der Hilst, R. D. (2019). Systematic  
 654 detection of clustered seismicity beneath the southwestern alps. *Journal of Geophysical*  
 655 *Research: Solid Earth*, *124*(11), 11531–11548.  
 656 Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.  
 657 Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for  
 658 data-driven discovery in solid earth geoscience. *Science*, *363*(6433).  
 659 Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is “nearest neighbor”  
 660 meaningful? In *International conference on database theory* (pp. 217–235).  
 661 Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010).  
 662 Obspy: A python toolbox for seismology. *Seismological Research Letters*, *81*(3), 530–  
 663 533.  
 664 Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*,  
 665 *36*(3), 287–314.  
 666 Cosentino, R., & Aazhang, B. (2020). Learnable group transform for time-series. In  
 667 *International conference on machine learning* (pp. 2164–2173).

- 668 DANA . (2012). *Dense array for north anatolia*. International Federation of Digital Seis-  
 669 mograph Networks. Retrieved from <http://www.fdsn.org/doi/10.7914/SN/YH.2012>  
 670 doi: 10.7914/SN/YH.2012
- 671 Deparis, J., Jongmans, D., Cotton, F., Baillet, L., Thouvenot, F., & Hantz, D. (2008).  
 672 Analysis of rock-fall and rock-fall avalanche seismograms in the french alps. *Bulletin*  
 673 *of the Seismological Society of America*, 98(4), 1781–1796.
- 674 Diaz, J. (2020). Church bells and ground motions. *Journal of Seismology*, 1–10.
- 675 Ebeling, C. W. (2012). Inferring ocean storm characteristics from ambient seismic noise: A  
 676 historical perspective. In *Advances in geophysics* (Vol. 53, pp. 1–33). Elsevier.
- 677 Emre, Ö., Duman, T., Özalp, S., Elmaci, H., & Olgun, S. (2011). 1: 250,000 scale active  
 678 fault map series of turkey. *Kayseri (NJ36-8) Quadrangle, Ankara*.
- 679 Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. ([http://](http://www.deeplearningbook.org)  
 680 [www.deeplearningbook.org](http://www.deeplearningbook.org))
- 681 He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on*  
 682 *knowledge and data engineering*, 21(9), 1263–1284.
- 683 Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and appli-  
 684 cations. *Neural networks*, 13(4-5), 411–430.
- 685 Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes  
 686 of seismic noise signals using unsupervised learning. *Geophysical Research Letters*,  
 687 47(15), e2020GL088353.
- 688 Kennett, B., & Engdahl, E. (1991). Traveltimes for global earthquake location and phase  
 689 identification. *Geophysical Journal International*, 105(2), 429–465.
- 690 Kodera, Y., & Sakai, S. (2020). An unsupervised automatic classification for continuous  
 691 seismic records: introducing an anomaly detection algorithm to solve the imbalanced  
 692 data problem. In *Agu fall meeting 2020*.
- 693 Köhler, A., Ohrnberger, M., & Scherbaum, F. (2010). Unsupervised pattern recognition in  
 694 continuous seismic wavefield records using self-organizing maps. *Geophysical Journal*  
 695 *International*, 182(3), 1619–1630.
- 696 Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A  
 697 survey on subspace clustering, pattern-based clustering, and correlation clustering.  
 698 *Acm transactions on knowledge discovery from data (tkdd)*, 3(1), 1–58.
- 699 Lacroix, P., & Helmstetter, A. (2011). Location of seismic signals associated with mi-  
 700 croearthquakes and rockfalls on the séchilienne landslide, french alps. *Bulletin of the*  
 701 *Seismological Society of America*, 101(1), 341–353.
- 702 LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.
- 703 Lin, W.-C., Tsai, C.-F., Hu, Y.-H., & Jhang, J.-S. (2017). Clustering-based undersampling  
 704 in class-imbalanced data. *Information Sciences*, 409, 17–26.
- 705 Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017).  
 706 Implementation of a multistation approach for automated event classification at piton  
 707 de la fournaise volcano. *Seismological Research Letters*, 88(3), 878–891.
- 708 Malfante, M., Dalla Mura, M., Mars, J. I., Métaixian, J.-P., Macedo, O., & Inza, A. (2018).  
 709 Automatic classification of volcano seismic signatures. *Journal of Geophysical Re-*  
 710 *search: Solid Earth*, 123(12), 10–645.
- 711 Meng, H., & Ben-Zion, Y. (2018). Characteristics of airplanes and helicopters recorded  
 712 by a dense seismic array near anza california. *Journal of Geophysical Research: Solid*  
 713 *Earth*, 123(6), 4783–4797.
- 714 Met Office. (2010 - 2015). Cartopy: a cartographic python library with a matplotlib interface  
 715 [Computer software manual]. Exeter, Devon. Retrieved from [http://scitools.org](http://scitools.org.uk/cartopy)  
 716 [.uk/cartopy](http://scitools.org.uk/cartopy)
- 717 Mousavi, S. M., & Beroza, G. C. (2019). Bayesian-deep-learning estimation of earthquake  
 718 location from single-station observations. *arXiv preprint arXiv:1912.01144*.
- 719 Obara, K. (2002). Nonvolcanic deep tremor associated with subduction in southwest japan.  
 720 *Science*, 296(5573), 1679–1681.
- 721 Peddinti, V., Sainath, T., Maymon, S., Ramabhadran, B., Nahamoo, D., & Goel, V. (2014).  
 722 Deep scattering spectrum with deep neural networks. In *2014 ieee international con-*

- 723 *ference on acoustics, speech and signal processing (icassp)* (pp. 210–214).
- 724 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others
- 725 (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning*
- 726 *research*, 12, 2825–2830.
- 727 Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake
- 728 detection and location. *Science Advances*, 4(2), e1700578.
- 729 Poyraz, S. A., Teoman, M. U., Türkelli, N., Kahraman, M., Cambaz, D., Mutlu, A., ...
- 730 others (2015). New constraints on micro-seismicity and stress state in the west-
- 731 ern part of the north anatolian fault zone: Observations from a dense seismic array.
- 732 *Tectonophysics*, 656, 190–201.
- 733 Riahi, N., & Gerstoft, P. (2015). The seismic traffic footprint: Tracking trains, aircraft, and
- 734 cars seismically. *Geophysical Research Letters*, 42(8), 2674–2681.
- 735 Ross, Z. E., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic
- 736 phase detection with deep learning. *Bulletin of the Seismological Society of America*,
- 737 108(5A), 2894–2901.
- 738 Rouet-Leduc, B., Hulbert, C., McBrearty, I. W., & Johnson, P. A. (2020). Probing slow
- 739 earthquakes with deep learning. *Geophysical research letters*, 47(4), e2019GL085870.
- 740 Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation
- 741 of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.
- 742 Seydoux, L., Balestrieri, R., Poli, P., De Hoop, M., Campillo, M., & Baraniuk, R. (2020).
- 743 Clustering earthquake signals and background noises in continuous seismic data with
- 744 unsupervised deep learning. *Nature communications*, 11(1), 1–12.
- 745 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ...
- 746 others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python.
- 747 *Nature methods*, 17(3), 261–272.
- 748 Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of*
- 749 *the American statistical association*, 58(301), 236–244.