Exploring continuous seismograms with machine learning

René Steinmann

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : STEP - Sciences de la Terre de lEnvironnement et des Planètes
Spécialité : Sciences de la Terre et de l'Univers et de l'Environnement
Unité de recherche : Institut des Sciences de la Terre

Exploration de sismogrammes continus par l'apprentissage automatique

Exploring continuous seismograms with machine learning

Présentée par :

René STEINMANN

Direction de thèse :

Michel CAMPILLO Professeur,	Directeur de thèse
Leonard SEYDOUX UGA	Co-encadrant de thèse

Rapporteurs :

Greg BEROZA
PROFESSEUR, Stanford University
Alessia MAGGI
PROFESSEUR DES UNIVERSITES, Université de Strasbourg

Thèse soutenue publiquement le **9 décembre 2022**, devant le jury composé de :

Michel CAMPILLO PROFESSEUR DES UNIVERSITES, UGA	Directeur de thèse
Greg BEROZA PROFESSEUR, Stanford University	Rapporteur
Alessia MAGGI PROFESSEUR DES UNIVERSITES, Université de Strasbourg	Rapporteure
Fabrice COTTON PROFESSEUR, Helmholtz Centre Potsdam, GFZ	Examinateur
William FRANK PROFESSEUR ASSISTANT, Massachusetts Institute of Technology	Examinateur
Céline HADZIIOANNOU PROFESSEUR ASSISTANT, Universität Hamburg	Examinatrice
Stephane GARAMBOIS PROFESSEUR DES UNIVERSITES, UGA	Président

Invités :

Leonard Seydoux
PROFESSEUR ASSISTANT, Université de Paris, Institut de physique du globe de Paris

Abstract

Continuous seismograms record time series of ground motion and are considered as a goldmine of information on active geological objects such as volcanoes or faults. However, the complexity and size of seismic data challenge the efficient and successful mining of the interesting information, hidden within a large amount of data. Automatic algorithms scanning continuous data streams can help overcome these challenges and might reveal new types of seismic signals, offering new insights about active geological objects. In this work, we develop a novel strategy based on machine learning, which infers meaningful and continuous patterns from seismograms and identifies groups of seismic signals in a data-driven fashion. The proposed strategy, which involves hierarchical waveform clustering, breaks up into three major steps: (1) a scattering network retrieves a rich and stable data representation of the continuous seismogram, (2) we lower the dimensionality of the data representation by extracting the most relevant features describing continuous temporal patterns, and (3) we perform hierarchical agglomerative clustering in the feature space, revealing hierarchical groups of similar signals in a tree-like structure. With this strategy we blindly identify a seismic burst of more than 200 similar low-magnitude earthquakes in the continuous seismogram recorded in a noisy urban environment. Besides identifying patterns and signal clusters related to various seismic sources, we are also able to infer medium changes due to freezing and thawing processes directly from the continuous seismogram of a single station. The continuous data-driven patterns describe also the stationarity of the seismic wavefield. An application to seismic data recorded in the vicinity of the Klyuchevskoy volcano, Russia, highlights the strong non-stationary character of seismic tremors, witnessing a constant change of the volcanic system. In general, hierarchical waveform clustering can deliver a quick and data-driven overview over the seismic signals and patterns present in the seismograms. Identifying blindly patterns related to medium changes seems possible and more studies and applications are needed for a generalization. We conclude that hierarchical waveform clustering might be a helpful tool in searching for tectonic background signals in vast amount of seismic time series.

Résumé

Les sismogrammes sont des séries temporelles du mouvement du sol considérées comme une mine d'informations sur les objets géologiques actifs tels que les volcans ou les failles. Cependant, la complexité et le volume de ces données rendent difficile une extraction efficace d'informations intéressantes. Des algorithmes automatiques appliqués aux données continues peuvent aider à surmonter ces difficultés et pourraient révéler de nouveaux types de signaux sismiques, offrant de nouvelles perspectives de recherche sur les objets géologiques actifs. Dans ce travail, nous développons une nouvelle stratégie basée sur l'apprentissage machine pour inférer des structures de signal significatives et continues à partir de sismogrammes, en particulier, des groupes de signaux sismiques. La stratégie proposée utilise le regroupement hiérarchique de formes d'onde, et comporte trois étapes principales : (1) un réseau diffusif permet une représentation riche et stable des données sismiques continues, (2) nous réduisons la dimensionnalité de la représentation des données en extrayant les éléments les plus pertinents décrivant les modèles temporels continus, et (3) nous effectuons un regroupement agglomératif hiérarchique à partir des données réduites, révélant des groupes hiérarchiques de signaux similaires dans une structure arborescente. Grâce à cette stratégie, nous montrons qu'il est possible de mettre en évidence des essaims sismiques de plus de 200 séismes similaires de faible magnitude dans des sismogrammes continus enregistrés dans un environnement urbain bruyant. Outre l'identification de groupes de signaux liés à diverses sources sismiques, nous déduisons également un changement de milieu dû à des processus de gel et de dégel directement à partir de données continues receuillies par une seule station. Ces caractéristiques continues basées sur les données fournissent également une excellente description du caractère stationnaire du champ d'ondes sismiques. Finalement, une application aux sismogrammes enregistrées à proximité du volcan Klyuchevskoy met en évidence le caractère fortement non stationnaire des trémors volcaniques, et témoigne d'une évolution constante du système volcanique. En général, le regroupement hiérarchique des formes d'onde peut fournir un aperçu rapide et orienté données des signaux sismiques et des structures présentes dans les sismogrammes. L'identification automatique de structures liées à des changements de propriétés du milieu semble possible et d'autres études et applications sont nécessaires pour une généralisation à d'autres cas d'étude. Le regroupement hiérarchique des formes d'onde s'avère être un outil utile pour la recherche de signaux tectoniques faibles dans les grandes séries temporelles sismiques enregistrées dans les observatoires sismologiques et volcaniques.

Acknowledgements

This work is the outcome of the recent three years of my life spent in Grenoble, France. Looking back in time, I am filled with immense gratitude and love for all the beautiful human beings I met and the opportunities I had. In the following I want to mention the people who had a great impact on me and on this work.

First of all: thank you Michel for giving me the opportunity to work on this thesis. You provided much more than an interesting subject to work on: your supervision was always supportive and full of trust, even in difficult times of lockdown and covid. You created an environment where I could develop and follow my own - often odd - ideas. You made me attend interesting summer schools and conferences where I learned many interesting things beyond my usual work. Thank you for all this.

Leonard, you are an amazing human being. The day I arrived in Grenoble you directly took me for a beer to "La Bobine" and at that point I knew that I made a good decision with saying yes to this thesis. From my point of view, you did a great job as a co-supervisor, providing many interesting discussions and a supportive environment for me to grow. You basically introduced me to the world of machine learning and your passion about it was definitely contagious. Thank you for all that and I wish you all the best for your future in Paris or wherever life will take you!

I also want to thank the thesis committee who took their time to read carefully my work and attended the thesis defense. The question round was very interesting for me and many of your questions are still resonating with me, perhaps opening new avenues for upcoming research. Thank you very much Greg Beroza, Alessia Maggi, William Frank, Celine Hadziioannou, Fabrice Cotton and Stephane Garambois.

The actual start of my journey in seismology started with my supervisor of my master's project: Celine Hadziioannou. From the retrospective, I feel very lucky that you started your tenure track at Hamburg right before I had to make a choice for my master's project. I'm very grateful to have had you as a mentor, introducing me to this strange and fascinating world of ambient seismic noise and all its magic tricks. I was really happy to have seen you again in Bordeaux and I hope that we keep crossing paths in the future. I am also amazed by what you have already built up in Hamburg and I wish you all the best for whatever the future holds for you.

Ein grosses Dankeschön auch an meine Eltern und meine Familie, die mir trotz der Distanz immer nah waren. Doch dieses Dankeschön reicht weiter als die letzten drei Jahre: in jeglichen Momenten meines Lebens bot sie mir Liebe und Geborgenheit und unterstützte mich in all meinen Entscheidungen, auch wenn sie noch so bizarr erschienen. Insbesondere möchte ich meinem Vater danken, der mir in allen Situationen vertraute und mich bedingungslos unterstützte. Ohne seine Liebe und

Unterstützung wäre all das nicht möglich gewesen.

Déménager loin de son pays d'origine et apprendre une nouvelle langue est un défi. L'endroit que l'on appelle chez-soi est loin et il n'est pas certain que le nouvel endroit se transforme en quelque chose similaire. Grenoble est définitivement devenue quelque chose qui signifie chez-soi pour moi et c'est principalement grâce aux gens. J'ai fait 13 heures de voiture de Hambourg à Grenoble et, à mon arrivée, j'ai été directement accueilli avec une bière fraîche par Félicien et Ildut, deux de mes nouveaux colocataires. C'est l'endroit où j'ai rencontré l'univers français de la pétanque, des tartiflettes et des apéros, en particulier pendant les périodes de covid. Même si cet endroit a vu beaucoup de gens venir et partir, il a toujours été un lieu de confort, un endroit que je suis heureux d'appeler chez moi grâce à Félicien, Sophie, Inès, Hélène, Jeanne, Audrey, Karine, Alix et Hugo. Merci d'avoir été si patients avec mon apprentissage du français et mes habitudes de dîner tôt.

Besides the coloc, there is another great community in Grenoble, which made me call Grenoble my new home: it is the community of ISTerre – or as a certain person would say: mes chers compatriotes. This place is full of life, interesting discussions and lovely people, I enjoyed every lunch and coffee break with all of you. Here I also met the people with whom I shared beautiful hikes, cold skiing days and long nights of dancing. In particular, I want to mention Lara, Gino, Dilruba, Hester, Malcon and Camila. You guys enriched my life in so many ways, I can't describe how grateful I am to have met you.

Contents

List of Figures

List of Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
ICA	**I**ndependent **C**omponent **A**nalysis
IC	**I**ndependent **C**omponent
ML	**M**achine **L**earning
NAF	**N**orth **A**natolian **F**ault
PCA	**P**rincipal **C**omponent **A**nalysis
PC	**P**rincipal **C**omponent
PDF	**P**robability **D**ensity **F**unction
STFT	**S**hort **T**ime **F**ourier **T**ransform
SVD	**S**ingular **V**alue **D**ecomposition

Chapter 1

Introduction

Since the first observed teleseismic earthquake in 1889, seismology went through many transformations regarding instrumentation, experiment designs, data processing and driving research questions (for a historical review see e.g. Dewey and Byerly, 1969). In the early days, the instruments had limited sensitivity and the recorded data were investigated and processed by hand (Lay and Wallace, 1995). Often, only earthquake recordings were considered of interest and the recorded ground motion before or after an event were ignored. For instance, only a few studies before 1950 addressed the nature of the ambient seismic wavefield due to limitations in instrumentation and processing techniques (for a review see e.g. Bonnefoy-Claudet, Cotton, and Bard, 2006). A major milestone was the introduction of the World Wide Standardized Seismographic Network (WWSSN) in the 1960s monitoring nuclear tests (Oliver and Murphy, 1971). This global network of seismographs established procedures which are common nowadays such as data sharing and the continuous recording of ground motion. Observations provided by this network elucidated the plate tectonic theory and the structure of the Earth's crust (Isacks, Oliver, and Sykes, 1968). In the following decades, many different types of network and array designs of different scales have been proposed such as the global network GEOSCOPE or the rolling array USArray. On the one hand the continuous recordings of ground motion and the different array designs improved our knowledge around classical seismological research questions regarding the physical process of an earthquake. On the other hand it also created unforeseen applications of seismology with new research questions such as environmental seismology (see e.g. Larose et al., 2015) or cryoseismology (see e.g. Podolskiy and Walter, 2016).

1.1 Knowledge gain by identifying known signals

The capability to derive knowledge and models from observational data is an important factor for many scientific disciplines including seismology. In the early days the amount of data was so little that the data treatment was manageable by hand. Moreover, seismic data was mainly used to study earthquakes which can be easily recognized in seismograms by human experts. The introduction of continuous recordings and arrays produced quantities of data that demanded automatized tools helping seismologists to retrieve the relevant information for the task at hand (e.g. earthquake detection for building earthquake catalogs). The design of these automatized tools are often based on what we currently know about the signal of interest. For instance, the short-term-average long-term average (STA/LTA) method is a tool for earthquake detection by utilizing the transient and large-amplitude character of earthquakes which has been observed for many years beforehand (Allen, 1978).

Joswig (1990) proposed an alternative approach to the STA/LTA, exploiting different known signal characteristics of earthquakes. Recently, supervised deep learning gained much attention in identifying earthquake signals thanks to the large amount of training data provided by earthquake catalogs (see Mousavi and Beroza, 2022, for a recent review). The continuous improvement of those methods helped in building very detailed catalogs from a growing amount of data, resulting in a knowledge gain regarding earthquake processes. In turn, the knowledge gain and new observations helped to improve the automatic tools.

1.2 Knowledge gain by identifying unknown signals

Besides gaining knowledge by exploiting the known signals, the identification of new and unknown signals reveal new insights, too. An interesting case is the detection of non-volcanic tremors in Japan in the early 2000s by Obara (2002). Due to its similar signal characteristics to volcanic tremors, the new signal was named non-volcanic tremor and soon similar signals were reported around other subduction zones and transform faults (see e.g. Rubinstein, Shelly, and Ellsworth, 2009). Generally, this signal class can be described as a continuous long-lasting vibration of low-amplitude, mostly observable in the frequency range of 2 to 8 Hz. The discovery of non-volcanic tremors shed new light on the physical processes occurring around plate boundaries and fault zones, since these signals originate from the deeper part of the faults (Shelly, Beroza, and Ide, 2007). Still today, the underlying mechanism is poorly understood. Non-volcanic tremors have been reported for some transform faults such as the Alpine Fault and the San Andreas Fault (Nadeau and Dolenc, 2005; Wech et al., 2012). For other transform faults such as the North Anatolian Fault, studies searched for non-volcanic tremors in seismic time series and reported null results (Pfohl et al., 2015; Bocchini et al., 2021). There could be many reasons for the absence of tremors in the data. It could be simply the case that the physical constraints of the North Anatolian Fault does not support tremor sources or that the instruments are not recording tremor signals due to its weak amplitude. It is also likely that the processing tools scanning the seismic time series for tremors are not well adapted to the general detection of tremors, since the design of these tools are based on our limited understanding of tremors and the few observations recorded at other fault zones. The poor signal characteristics led even to the confusion of train generated signals with tremors (Hutchison and Ghosh, 2017; Inbal et al., 2018). The example of non-volcanic tremors motivates the need of methods which are able to identify unknown patterns in continuous seismograms in a data-driven fashion (that is solely derived from the data itself). We are confident that the increasing amount of data and the increasing sensitivity of the instruments provide overseen but relevant information, improving our understanding of active geological objects such as fault systems or volcanoes.

The scope of the presented work is to design and test tools which identify interesting patterns in seismic time series in a data-driven fashion. Chapter 2 introduces the proposed strategy for seismic data exploration. Chapter 3 discusses the first application where we reveal the hierarchical structure of seismic signal classes recorded in an urban environment close to a fault zone (Steinmann et al., 2022). Chapter 4 shows the second application where we identify blindly the seismic imprint of a medium change in the continuous seismogram (Steinmann, Seydoux, and

Campillo, 2022). Chapter 5 analyzes a seismic dataset recorded in the vicinity of a volcano and uncovers the ever-changing nature of a tremor-dominated wavefield. Chapter 6 summarizes the main outcome of this thesis, leading to new ideas and research questions.

Chapter 2

Towards clustering seismic waveform data

The main question which arises is: *how to design a tool which identifies signals and patterns we might have never seen before?* Compared to the example of STA/LTA for earthquake detection, we do not have any information on how a relevant or interesting signal could look like. Therefore, the tool should be able to pick up something "interesting" in a data-driven fashion. A common approach would be to compare all the recorded signals of a given dataset and group them based on their similarity or dissimilarity. A method of machine learning (ML), called clustering, groups objects based on similarity measurements. Objects are grouped in the same cluster, if the similarity is high and objects are grouped in different clusters, if the similarity is low. Usually the objects are described with a set of characteristics – called features – and the similarity measurement takes place in the feature space. This similarity measurement could be any type of distance $d \in \mathbb{R}$ between two objects $x, y \in \mathbb{K}$ in the feature space $\mathbb{K}$:

$$d(x, y) = |x - y|, \tag{2.1}$$

where $|\cdot|$ denotes a norm, usually the Euclidean norm. If we apply that idea to seismic time series, we could define a sliding time window and a cluster analysis assigns clusters to all the windows according to the similarity measurement. Ideally, windows containing similar seismic signals are then grouped in the same cluster. This data-driven analyses – which we call from now on waveform clustering – would reveal reoccurring patterns in the seismic time series we might have never seen before. The concept sounds simple but the devil lies in the details and, therefore, the present chapter is centered on how exactly we design this approach.

We start with the general concept of clustering and how this task can be solved in different ways. Then, we motivate the need to transform the seismic waveform data into a representation adapted to the task of clustering. For this purpose, we introduce the concept of a scattering network and dimensionality reduction methods. At this point, we covered all the details of our proposed strategy (shown in Figure 2.1) and we put it in the context of similar approaches, which have been developed in the past. At last, we introduce shortly manifold learning techniques, which offer an interesting and promising way to explore continuous seismograms a bit differently to the waveform clustering approach.

2.1 The world of clustering

The aim of clustering is to gain new knowledge about a given data set by revealing and exploring its underlying structure. Given the general description and vaguely

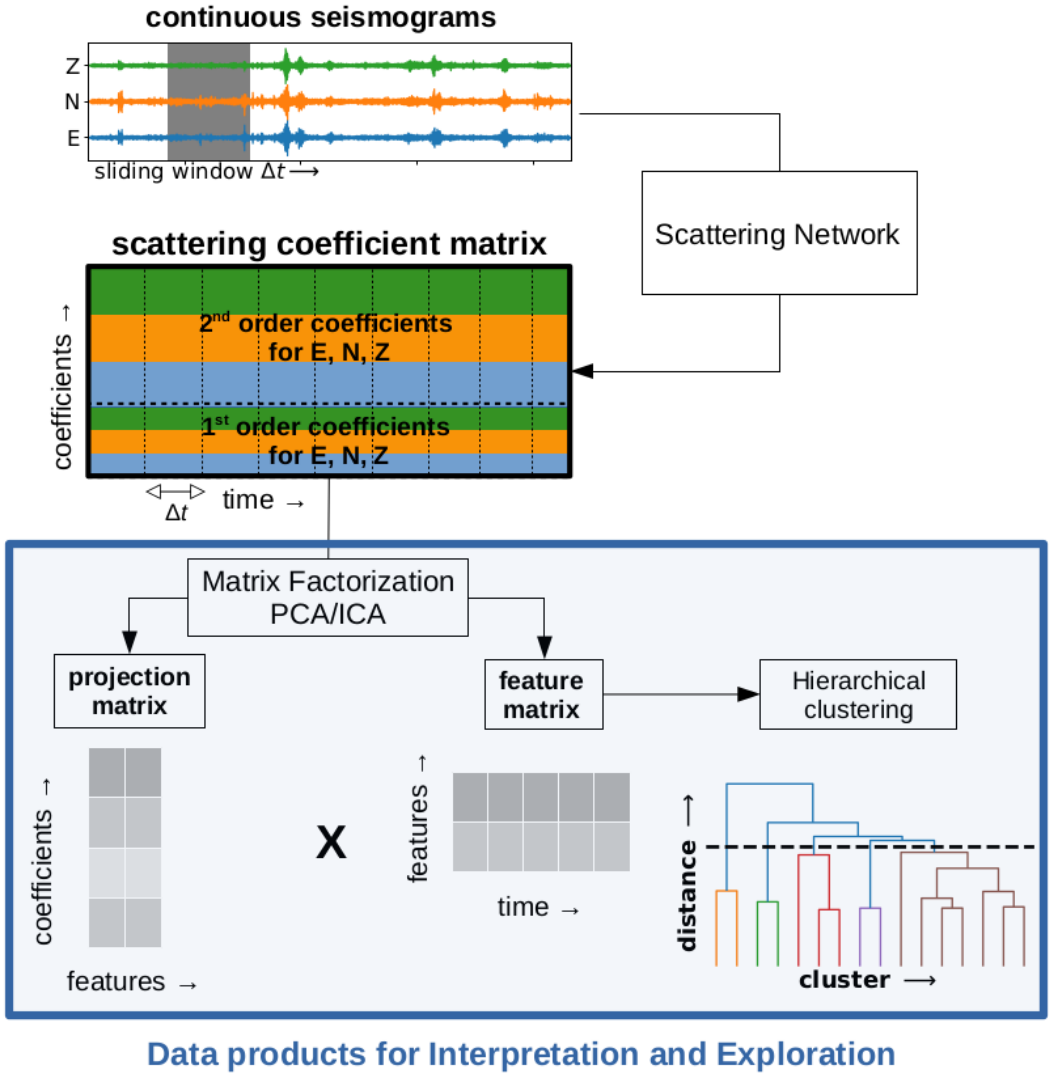

FIGURE 2.1: The proposed strategy for exploring continuous seismograms. For detailed explanation consider Section 2.2 and Figure 2.5 for the scattering network, Section 2.3 for the matrix factorization and Section 2.1.2 for hierarchical clustering.

defined goal of clustering, it is evident that the realization of such a task is ambiguous and subject to many details. Therefore, a whole zoo of clustering algorithms exists nowadays with a variety of distance measurements and objective functions. Choosing the appropriate clustering algorithm is a challenging task and requires a minimum amount of domain and data knowledge such as the underlying distribution of the data. However, there is no established general guideline in choosing the one and only clustering algorithm and, often, it is recommended to test different types of algorithms. Clustering is a tool for data exploration and knowledge discovery. It is not an automated end-to-end approach but rather an interactive multi-objective optimization task based partially on trial and error. In a similar mindset, Vladimir Estivill-Castro wrote in a positioning paper about clustering: "Clustering is in the eye of the beholder" (Estivill-Castro, 2002). In the following lines we will introduce some different strategies for clustering and argue why we choose agglomerative hierarchical clustering.

2.1.1 Many ways to solve the clustering task

Figure 2.2 from Ezugwu et al. (2022) shows a possible hierarchical taxonomy of the most common clustering algorithms. At the top clustering can be divided into hierarchical and partitional clustering approaches. Hierarchical clustering builds a hierarchy of clusters with either a bottom-up (agglomerative) or top-down approach (divisive). In agglomerative hierarchical clustering, clusters consists of single objects and are then iteratively merged until all objects are unified in a single cluster. It is the opposite case for the divisive approach: a cluster containing all objects is iteratively divided into smaller clusters until each cluster contains one object. For both approaches, the decision to merge two clusters is based on a so-called linkage criterion which measures the dissimilarities between all clusters at each iteration. Most commonly, the two clusters showing the smallest dissimilarity are merged. We will discuss different linkage criteria and their properties more in detail later. A so-called dendrogram visualizes the clustering process and reveals which clusters are merged at what distance. By applying a distance threshold to the dendrogram, one can obtain all merged clusters until the set threshold.

As a visual example of agglomerative hierarchical clustering consider Figure 2.3. Seven objects (A to F) are described by a two-dimensional feature space and we assume that the distance in this feature space represents a similarity measurement. By eye, we could identify three clusters of different population sizes: A and B; C; and D, E and F. The right hand-side of Figure 2.3 shows a sketch of a possible dendrogram revealing the hierarchical clustering process. Instead of providing a single clustering solution, the dendrogram shows us the underlying structure of the whole data set. For example, the dendrogram tells us that A and B are very close to each other since they merge at a very low distance. Since A and B are far away from the other objects, they merge at a very large distance with the remaining data. Note that the sketch aims at visualizing the general concept of agglomerative hierarchical clustering and it does not take into account the different types of linkage criteria and distances.

Compared to hierarchical clustering, partitional clustering provides a single partitioning of the data set without producing a hierarchical structure of the data such as the dendrogram. Therefore, it returns less information than hierarchical clustering. However, it can handle a larger amount of data, since the construction of the dendrogram is computationally expensive (Jain, Murty, and Flynn, 1999). The most

famous example for partitional clustering is k-Means which finds k clusters with objects belonging to the clusters with the nearest mean (centroid). The number of clusters k has to be pre-determined, which is one of the main drawbacks of k-means. Based on a given number of clusters k, k-Means initializes randomly k centroids and then adjusts these centroids iterativaly by minimizing the within-cluster variances and maximizing the variability between the clusters. The random initialization of the centroids is another drawback, since the final result can vary largely depending on the initial positions. The iteration stops when the centroids are stable in space or a stopping criterion is met. k-means is considered a hard partitioning strategy where the data is divided into distinct clusters and each object belongs to only one cluster. Fuzzy clustering provides a different strategy where clusters can overlap and objects are assigned to multiple clusters with a degree of membership. The fuzzy version of k-Means is called fuzzy C-means. Fuzzy approaches are an interesting choice if the dataset does not show clear boundaries between different sets of objects.

For a more complete overview of different clustering approaches and how they compare, we refer to Jain, Murty, and Flynn (1999), Jain (2010), and Ezugwu et al. (2022). For this work, we use agglomerative hierarchical clustering since the dendrogram seems to be an interesting and helpful tool to explore the content and structure of the data. Moreover, seismic data is known for its large class imbalances such as between earthquakes and the ambient seismic wavefield. Therefore, k-means which obtains clusters of similar sizes would be an inappropriate choice. Nevertheless, we want to emphasize that hierarchical clustering is a choice we made and other approaches might reveal also interesting and perhaps different patterns. However, it is out of the scope of this work to explore and compare all possible clustering approaches.

2.1.2 About agglomerative hierarchical clustering

The linkage criterion is the main driving parameter behind agglomerative hierarchical clustering. It measures the dissimilarity $D(U, V)$ between all cluster $U = \{u_1, \dots u_N\}$ and $V = \{v_1, \dots v_M\}$ and merges the two cluster U and V for which $D(U, V)$ is minimal. At each step cluster U and V with the smallest D are merged and a new cluster replaces U and V. Due to the newly formed cluster, the similarity measurement D between all clusters has to be updated and again the two closest clusters are merged. In the agglomerative approach this process is repeated until all objects are unified in a single cluster. In the following we will introduce only the most common examples for the many different realizations of the linkage criterion.

The single linkage criterion

The most simple and straightforward approach is the single linkage criterion. Given two clusters U and V with its respective objects u_i and v_j and a distance measurement d, the cluster dissimilarity D of the single linkage approach is given as:

$$D(U, V) = \min_{u_i \in U, v_j \in V} d(u_i, v_j) \tag{2.2}$$

In other words, the distance between the two clusters U and V is given by the smallest distance of all pairwise distances of its objects. Therefore, the single linkage criterion is also often called nearest neighbor method. In the context of agglomerative clustering, this criterion merges the two clusters with the nearest neighbor. This

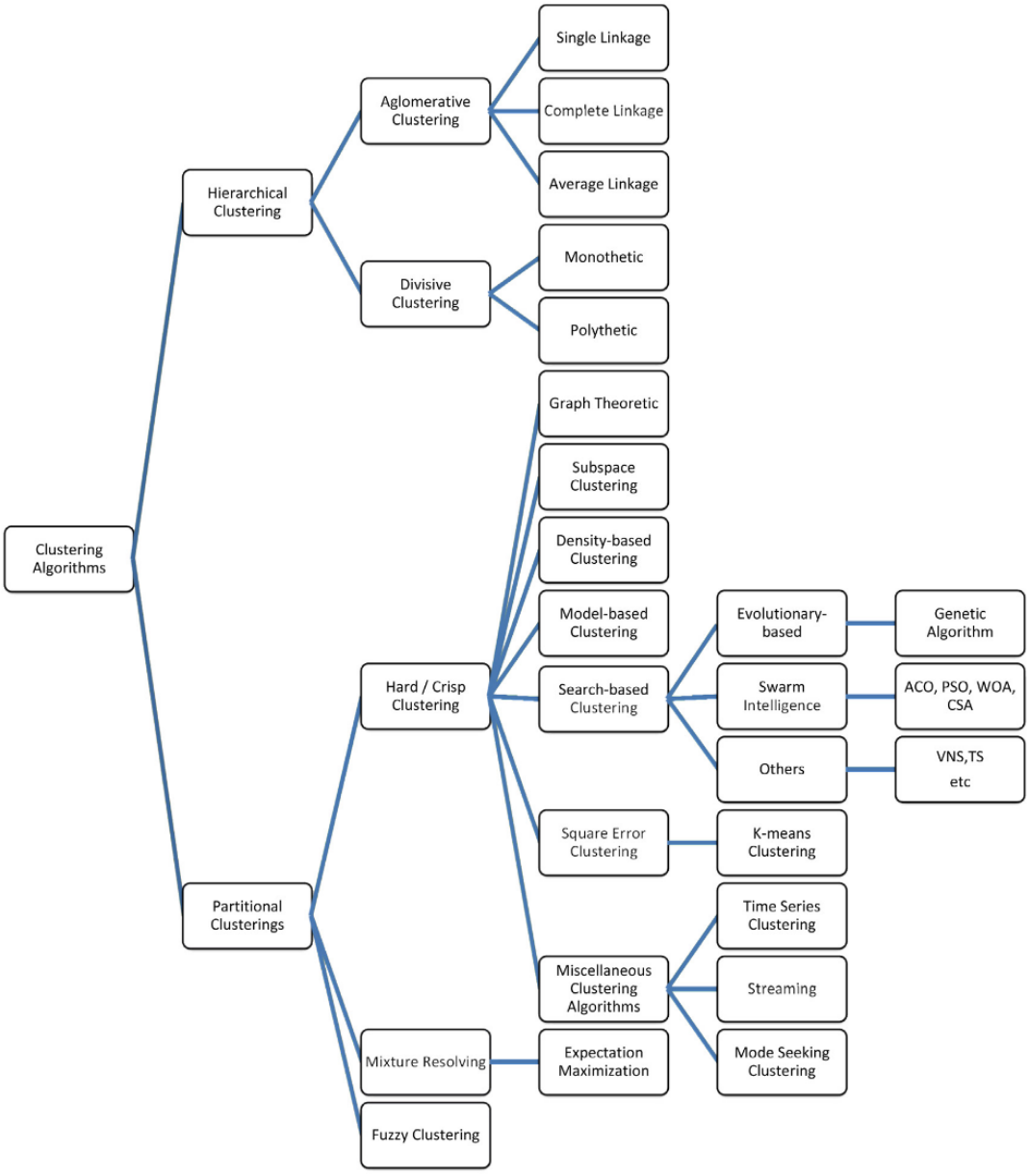

FIGURE 2.2: Taxonomy of clustering after Ezugwu et al. (2022)

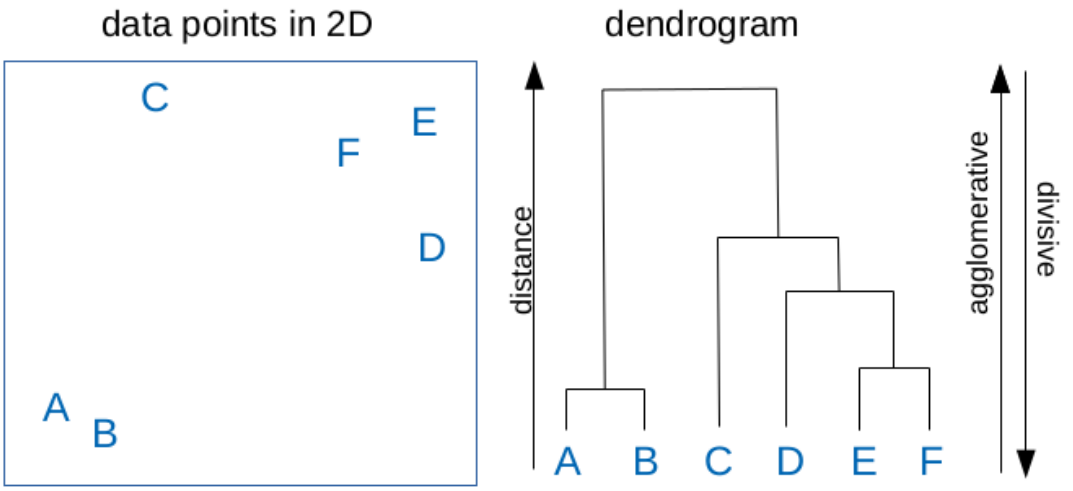

FIGURE 2.3: An illustration of hierarchical clustering without considering the linkage criterion or type of distance measurement

definition and equation 2.2 show already the main drawback of this method. Two large and distinct clusters would merge as soon as one member of one cluster would be close to one member of the other cluster - hence, the naming: single linkage. This causes poorly separated clusters to be chained together (the well-known chaining effect of single linkage) and long and thin clusters are constructed. Due to these drawbacks and its lack of robustness, the use of single linkage is often not recommended (see e.g. Baker, 1974; Milligan and Isaac, 1980).

The complete linkage criterion

The opposite realization of the single linkage criterion would be the complete linkage where the dissimilarity between two clusters is defined by the largest distance of all pairwise distances of its objects. Therefore, it is often called the furthest neighbor method. The dissimilarity measurement can be written as:

$$D(U,V) = \max_{u_i \in U, v_j \in V} d(u_i, v_j) \tag{2.3}$$

$D(U,V)$ can be also described as the diameter of the newly formed cluster S resulting from the merge of U and V. Compared to single linkage which finds elongated clusters, the complete linkage approach builds many clusters with small within-cluster dissimilarities. A major drawback of this method is that two similar clusters are lately merged when one of their pairwise distances is large. Therefore, complete linkage is often referred to be space-dilating and single linkage is referred to be space-contracting. The two extreme end-members of a linkage criterion show that a compromise between the two might be the better choice for most applications. The centroid and Ward's method would fall into that category. In the following we will introduce the Ward's method, since this is our choice for the presented work.

The Ward's method

The Ward's method, which has been introduced in Ward Jr (1963), merges two clusters at each iteration with a more complex objective function. This linkage criterion merges two clusters which result in the smallest amount of increase of the within-cluster variances including the newly formed cluster. Therefore, the Ward's method has to calculate at each step all possible cluster merges and its within-cluster variance. In that sense, it shares similarities with the objective function of k-means but it is applied in a hierarchical and not partitional clustering process. According to Müllner (2011), the objective function which minimizes the change of variance can be written as:

$$D(U,V) = \sqrt{\frac{2\#U\#V}{\#U+\#V}}\|c_U - c_V\|_2 \tag{2.4}$$

with $\#X$ being the cardinality of a cluster X, c_X being the centroid of a cluster X and the Euclidean distance $\|\cdot\|_2$. Note that this criterion gives a measurement of the variance of the newly formed cluster based on the Euclidean distance between the centroids of the merged clusters. According to Kuiper and Fisher (1975), the Ward's method is well-suited for data with spherical multivariate normal distributions. However, it runs into difficulties if the clusters have unequal diameters or if the cluster's shape is more ellipsiodal than spherical (Kaufman and Rousseeuw, 2009).

The introduction of the different linkage criteria has shown that all approaches come with different characteristics making them suitable for different types of data. Depending on the type of data, one criterion might perform better than the others and the criterion itself also imposes a certain structure on the data. However, single and complete linkage are rarely a good choice and mostly a criteria between the two extremes such as the Ward's method is more appropriate. In this work, we only apply the Ward's method, however, we want to point out that other interesting linkage criteria such as the centroid's method also exist and have not been tested in this work.

2.2 Waveform data representation with the scattering network

Until this point, we discussed the general idea of clustering and the concept of agglomerative hierarchical clustering – our choice of clustering for this work. In the specific case of waveform clustering, we would like to assign clusters with a sliding window to continuous seismograms (as indicated by the sliding window in Figure 2.1). Unfortunately, the waveform data itself is not adapted to the task of clustering for two main reasons. Firstly, the waveform data is a representation sensitive to translation, i.e. the representation contains information about the position of a signal in time. This is problematic for clustering, since this property results in large distance measurement for the same signal shifted in time. Usually, a signal shifted in time is considered still the same signal and should ideally result in a distance measurement of zero, i.e. they are identical and located in the same cluster. Secondly, the waveform data is an unstable representation regarding small signal deformations. A distance measurement taken on two signals, where one is a slight compressed or dilated version of the other, would return large values and the two highly similar signals would be located in different clusters. For these two reasons, it is crucial to

find a translation invariant representation which is stable towards small deformations and, thus, adapted to the task of clustering. The amplitude spectrum of the Fourier transform is translation invariant but not stable towards small deformations of the signal. Due to its non-localized sine waves the Fourier transform is unstable towards deformations, in particular for higher frequencies (Bruna and Mallat, 2013). The wavelet transform replaces the sine waves by localized waveforms and, thus, the wavelet coefficients deliver a stable representation regarding the deformation of the signal. However, the wavelet transform is not translation invariant. By adding non-linear averaging operators to the wavelet transform, we can create an architecture, which resembles a Convolutional Neural Network (CNN) and outputs a translation invariant representation. However, the non-linear averaging operator destroys important information about the signal. By repeating the wavelet transform in combination with the averaging non-linear operators, we can recover most of the lost information in higher order coefficients and create a representation which is translation invariant and stable to small deformations. The described architecture is called a scattering network and has been mainly introduced in Bruna and Mallat (2013) and Andén and Mallat (2014). In this work, we choose to represent seismic waveform data as scattering coefficients produced by the scattering network.

2.2.1 The mathematical design of the scattering network

In the following lines, we will define mathematically the scattering network. Considering a wavelet $\psi(t)$, we can define a set of filter bank $\psi_\lambda(t) = \lambda\psi(\lambda t)$ by dilating the original wavelet $\psi(t)$ - also called mother wavelet - with a set of dilation factors $\lambda \in \mathbb{R}$. In the frequency domain the set of wavelet banks would be $\hat{\psi_\lambda}(\omega) = \hat{\psi}(\omega/\lambda)$. The dilation factor λ can then be defined as

$$\lambda = 2^{\frac{k}{Q}}, k = \{0, 1, ..., JQ - 1\} \tag{2.5}$$

with $Q \in \mathbb{N}$ being the number of wavelets per octave and $J \in \mathbb{N}$ being the number of octaves. This definition of the dilation factor provides a logarithmic grid of the center frequencies for the set of wavelet filter banks.

By convolving a time series $x(t) \in \mathbb{R}$ with a set of wavelet filter banks $\psi_\lambda(t)$ and taking the modulus, we obtain a real-valued time-frequency representation $W_\lambda(t)$ of the time series called a scalogram:

$$W_\lambda(t) = |x(t) \star \psi_\lambda(t)| \tag{2.6}$$

This is the first convolutional layer of the scattering network with the convolution operator $\star$. In Andén and Mallat (2014) the authors introduce a low-pass filter $\phi(t)$ to retrieve the first-order scattering coefficients:

$$S_1 x(t, \lambda) = W_\lambda(t) \star \phi(t) = |x(t) \star \psi_\lambda(t)| \star \phi(t) \tag{2.7}$$

The low pass filter smooths the representation and makes it more stable to small deformation of the signal. However, it also removes other small scale structures of the signal which might be important for classification or clustering tasks. This information is recovered by repeating the convolution and modulus operation, retrieving higher-order scattering coefficients. Note that the set of dilation factors λ can differ with the layer of the scattering network. With two sets of wavelet filter banks, $\psi_{\lambda_1}(t)$

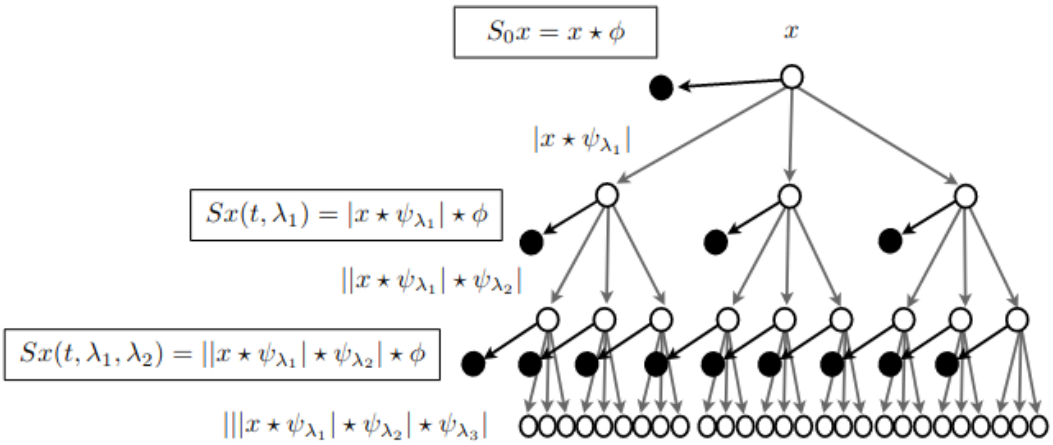

FIGURE 2.4: Sketch of a three-layer scattering network, retrieved from Andén and Mallat (2014)

at the first layer and $\psi_{\lambda_2}(t)$ at the second layer, we can calculate the second-order scattering coefficients:

$$S_2x(t, \lambda_1, \lambda_2) = ||x(t) \star \psi_{\lambda_1}(t)| \star \psi_{\lambda_2}(t)| \star \phi(t) \tag{2.8}$$

By repeating this operation many times, we can retrieve higher-order scattering coefficients which add more and more information. However, Andén and Mallat (2014) already concluded that the information gain beyond second-order scattering coefficients is marginal compared to the increasing computational costs. Therefore, we limit ourselves here to a two layer scattering network recovering first- and second-order scattering coefficients.

Figure 2.4 provides a general sketch of a three-layered scattering network and Figure 2.5 shows a two-layered scattering network applied to continuous seismic data. At each layer of the network we retrieve the scattering coefficients Sx and the scalograms $U(x)$ are forwarded to the next layer. Note that this architecture resembles a convolutional neural network with the difference that each layer produces an output and the neurons are restricted to wavelets. Andén and Mallat (2014) also propose to output a time-averaged descriptor S_0x of the input signal before calculating the first-order scattering coefficients. In this work we disregard S_0x, since S_1x and S_2x already provide a highly redundant representation of the data.

2.2.2 A scattering network with Morlet wavelets

We restrict the wavelets of the scattering network to Morlet wavelets as initially proposed in Bruna and Mallat (2013) and Andén and Mallat (2014). The Morlet wavelet $\psi(t)$ with a center frequency f is a complex exponential multiplied with a Gaussian window:

$$\psi(t) = \exp(-i2\pi ft) \exp(-t^2/a^2) \tag{2.9}$$

While f are the center frequencies defining the modulation of the Morlet wavelet, a defines the exponential drop-off of the waveform. We define a as a function of the bandwidth d and the center frequency f, which in turn depends on the Nyquist frequency f_N of the signal $x(t)$ and the dilation factor λ:

$$a_j = \frac{d}{f} = \frac{d}{\lambda f_N} \tag{2.10}$$

2.2.3 Becoming translation invariant with pooling

We choose $\phi(t)$ to be a pooling operation which ensures a stable and translation invariant representation for each window. The pooling operation retrieves a single value for each scale in the scalogram and, thus, acts as a low pass filter and downsampling operation (Dumoulin and Visin, 2016). There are many different types of pooling operation, filtering different types of information. In Seydoux et al. (2020) the authors applied the scattering network with an average pooling, which averages the scattering coefficients and collapses the time axis within the sliding window. Other possibilities are maximum pooling or median pooling where either the maximum or median value is taken for each scale in the scalogram. In this work, we will consider average, median and maximum pooling as potential filters.

2.2.4 Collecting the scattering coefficients into a data matrix

The scattering network produces scattering coefficients at each layer and they have to be collected and organized before clustering. Clustering is usually applied to a data matrix, where the columns correspond to the observations and the rows correspond to the features describing each observation. In our case of waveform clustering, the data matrix contains the scattering coefficients where the columns correspond to the sliding window in time and the rows contain the scattering coefficient retrieved at each layer. For a visual description for this procedure consider Figure 2.5. The scattering network is applied with a sliding window on one continuous seismogram and we collect the first- and second-order scattering coefficients for each window. Using three component seismograms, we apply this approach for each channel and concatenate the first- and second-order scattering coefficients of all three channel in columns. This way we create a data matrix for continuous three component seismograms which is the basis for all the following tasks (see Figure 2.1).

2.3 Feature extraction for clustering

With the scattering coefficients we finally found a translation-invariant representation stable towards small signal deformations. Unfortunately, this representation comes with a new and unwanted property regarding clustering tasks: it is high-dimensional. In fact, high-dimensional data is very problematic for many machine learning tasks and this phenomena was coined the curse of dimensionality by Bellman (1966). In our specific case, we want to use agglomerative hierarchical clustering which relies on distance measurements where a close neighbor shares more similarities with an object than a far away neighbor. The general notion and intuition we have about distance and neighborhood in a two- or three-dimensional space does not translate easily into higher-dimensional space (Domingos, 2012). For a wide variety of data distributions and distance measurements in a large number of dimensions, the ratio of a distance to the closest and furthest neighbor to an object tends to be almost 1 (Aggarwal, Hinneburg, and Keim, 2001). Thus, a distance measurement in a high-dimensional space is not able to tell us what is close and what

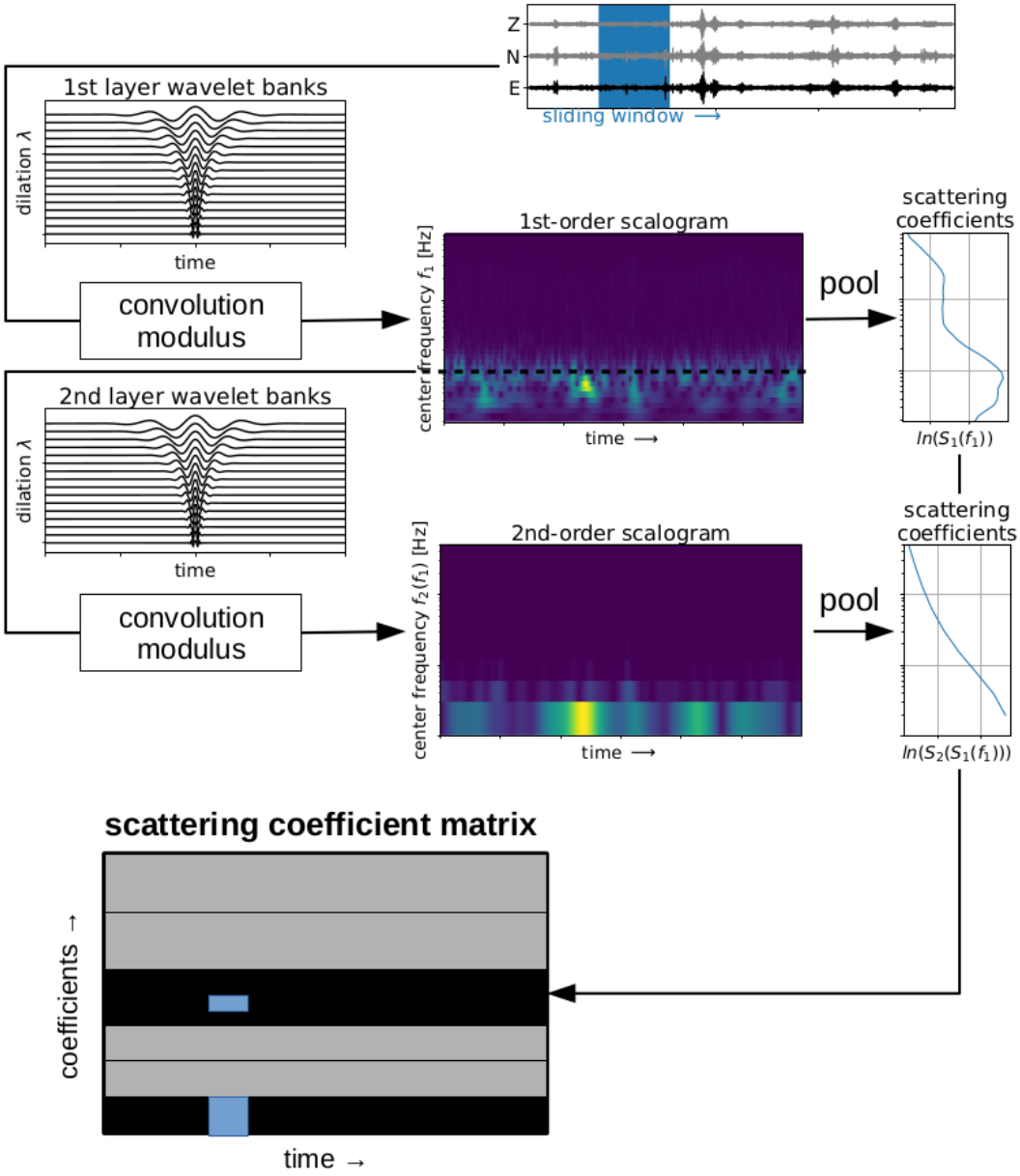

FIGURE 2.5: A detailed view on a two-layered scattering network applied to continuous three-component seismograms with a sliding window. The dashed line in the 1st-order scalogram indicates the data row which is convolved with the 2nd-layer wavelet banks. The blue boxes in the scattering coefficient matrix show schematically where these specific scattering coefficients are stored. The complete strategy for hierarchical waveform clustering is presented in Figure 2.1.

is far and the notion of neighborhood becomes meaningless. Moreover, with the increasing number of dimensions, we also increase the number of neighbors with the same distance. Therefore, clustering algorithms which use distance measurements such as agglomerative hierarchical clustering are sensitive to these problems and can create meaningless results if applied to high-dimensional data.

2.3.1 Reducing the number of dimensions

The solution to the curse of dimensionality is to extract a low-dimensional representation with the most relevant information (features) and perform the clustering task within that subspace (feature space). In a perfect world, this low-dimensional representation corresponds to the data's intrinsic lower dimension, which is the minimum number of variables to describe fully the data. For instance, in our case, the scattering coefficients provide a redundant representation, since the first-order wavelets are densely spaced with overlapping frequencies. However, we often do not know the data's high-dimensional structure and, thus, reducing the dimensions of a dataset means a loss of information. Therefore, dimensionality reduction techniques have to make compromises about what type of information to preserve. This compromise results in a variety of methods focusing on the preservation of different structures in the data. For instance, the methods could be divided into two groups aiming either at preserving the pairwise distance structure between all data points or at preserving local over global distances. Preserving pair-wise distances is important if we utilize clustering as a next step. This is particular true for hierarchical clustering which computes pair-wise distances for measuring dissimilarity. Methods which aim at preserving local structures over global structures are great tools for visualization since they can compress a lot of structure in only two or three dimensions. However, their distortion of global structures makes them less suitable for clustering tasks, since we lose information about inter-cluster relations. In the following we introduce the principal and independent component analysis (PCA/ICA) which perform a matrix factorization and reduce the dimensions linearly with the aim of preserving pair-wise distances.

2.3.2 Matrix factorization

Assume that we collected the scattering coefficients of a seismic time series in a data matrix $\mathbf{X} \in \mathbb{R}^m x \mathbb{R}^n$ with n samples in time and m coefficients per time step (see the scattering coefficient matrix in Figure 2.1 and Figure 2.5). Matrix factorization describes the data matrix $\mathbf{X}$ as a product of two matrices $\mathbf{A} \in \mathbb{R}^m x \mathbb{R}^k$ and $\mathbf{Y} \in \mathbb{R}^k x \mathbb{R}^n$:

$$\mathbf{X} = \mathbf{AY}, \tag{2.11}$$

where $\mathbf{Y}$ describes the k-dimensional feature space and $\mathbf{A}$ is the linear operator providing the mapping between the high-dimensional data and the low-dimensional feature space. In the hierarchical waveform clustering approach we utilize both matrices for exploring the content of the data (see Figure 2.1). Equation 2.11 can be solved in many ways under different conditions, resulting in a large variety of suitable algorithms.

Matrix factorization is also a common framework used in applications for blind source separation, aiming at unmixing a set of signals into its original source signals with little or no information. The original source signals in $\mathbf{Y}$ are multiplied with the *mixing* matrix $\mathbf{A}$, yielding the recorded signals in $\mathbf{X}$. A classic example of blind

source separation is the cocktail party problem where a person at a cocktail party - listening to all the simultaneously speaking people - tries to identify one voice speaking. This problem is relatively simple for humans, however, in digital signal processing it poses a difficult task.

2.3.3 Principal Component Analysis (PCA)

One of the most common matrix factorization is the principal component analysis (PCA). PCA tries to find a set of orthogonal unit vectors (orthonormal basis) which explains most of the data's variance. The first principal component finds the axis in the data space which maximizes its variance without any constrain of orthogonality. The second principal component finds an orthogonal axis to the first one, while also maximizing the variance. This can continue until the retrieved principal components explain the total variance of the data set. Note that the first component explains the greatest variance, the second component the second-greatest variance and so on. Thus, the explained variance ranks the principal components. The principal components are the eigenvectors of the data's covariance matrix and, therefore, the eigendecomposition of the data's covariance matrix or the singluar value decomposition (SVD) of the data matrix deliver the principal components. The SVD states that the data matrix $\mathbf{X}$ can be decomposed as following:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{2.12}$$

with $\mathbf{U} \in \mathbb{R}^m x \mathbb{R}^m$ containing m orthonormal vectors called the left singular vectors of $\mathbf{X}$, $\mathbf{V} \in \mathbb{R}^n x \mathbb{R}^n$ containing n orthonormal vectors called the right singular vectors of $\mathbf{X}$ and the positive diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^m x \mathbb{R}^n$ containing the singular values of $\mathbf{X}$. Regarding equation 2.11, the principal components are the k rows of $\mathbf{Y}$ and the pseudo-inverse of $\mathbf{A}$ contain the right singular vectors of $\mathbf{X}$.

2.3.4 Independent Component Analysis (ICA)

The Independent Component Analysis (ICA) is considered a generalization of PCA, since it relaxes the constrain of orthogonality and it has a stronger definition of independence regarding its components (Comon, 1994). Two variables are considered statistically independent if the realization of one does not affect the probability distribution of the other. Principal components are uncorrelated but not necessarily statistically independent. Independent components are statistically independent and, therefore, also uncorrelated.

Independence and non-Gaussianity

ICA solves equation 2.11 by maximizing the independence of its components. It maximizes the independence by maximizing the non-Gaussianity of each component. To understand better the relation between independence and non-Gaussianity, imagine a two-dimensional point cloud which is Gaussian on both axis. All the set of two axes we would draw through this point cloud would show the same independence and, thus, ICA trying to maximize the independence of two components would fail (Hyvärinen and Oja, 2000). Therefore, the aim of ICA trying to maximize the independence can be understood as the maximization of the non-Gaussianity of the sources. For the same reason, ICA can retrieve at maximum one Gaussian component, since this problem would be unsolvable with more than one Gaussian component. This approach can also be justified with the mindset of blind source

separation and the central limit theorem, which states that the sum of a sufficient number of distributions tend to be Gaussian. Thus, a strongly non-Gaussian signal is unlikely the sum of many signals with different distributions.

How to maximize independence?

In order to maximize the independence of the components, we need a measurement of non-Gaussianity. One type of measurement would be the fourth-order cumulant, also called kurtosis, which is easy to implement and to understand. However, it is not a robust measurement since it is sensitive to outliers (Hyvärinen and Oja, 2000). A more robust measurement is the so-called negentropy, which is based on entropy. Entropy can be interpreted as the degree of information that the observation of a variable reveals. The value of Entropy is large for unpredictable and unstructured variables. Since Gaussian variables have the largest entropy among all random variables of equal variance, the entropy serves as a good estimation of non-Gaussianity in ICA. The negentropy is a modified differential entropy which is always non-negative. It is zero if and only if the variable is Gaussian. In terms of statistical properties, it is an optimal estimator for Gaussianity, however, it is difficult to compute since it needs the probability density function (PDF) of the variable. Therefore, algorithms such as FastICA use an estimation of the PDF, which makes the computation much simpler. For the sake of completeness, we want to mention that also other method exists to estimate the ICA. However, introducing all the different approaches is out of the scope here and in this work, we will use the FastICA approach with the negentropy measurement. As a preprocessing step, FastICA whitens the data and reduces the dimensionality with a PCA. Since the independent components in $\mathbf{Y}$ and the mixing matrix $\mathbf{A}$ are unknown, ICA is not able to provide an ordering or scaling (including the sign) of the components. Any scalar value, we would multiply with the components, could be added as a division to the mixing matrix and ,similarly, we could change freely the order of the components.

Visual comparison of PCA and ICA

Figure 2.6 shows a comparison of PCA and ICA applied to data retrieved from Student-t distributions, which are bell-shaped as Gaussian distributions but with heavier tails. The two variables x and y are observations based on a mixing of two Student-t distributions with a low number of degrees of freedom. While PCA finds orthogonal components, which maximizes the variance, ICA identifies non-orthogonal directions of maximal non-Gaussianity and, thus, unmixes the mixed observations into its underlying Student-t distributions.

2.4 State of the art: seismic waveform clustering

We have outlined and explained the complete strategy shown in Figure 2.1. A scattering network transforms the continuous seismogram with a sliding window into a clustering-adapted representation. PCA or ICA retrieve the most relevant features from the high-dimensional data matrix, which are then used for agglomerative hierarchical clustering. Following the idea of exploring continuous seismograms in a data-driven fashion with waveform clustering, some strategies were proposed and tested in the recent years. To our knowledge, Köhler, Ohrnberger, and Scherbaum (2010) were the first ones who proposed a method which scans and labels continuous seismograms in a data-driven fashion. A sliding window scans through the

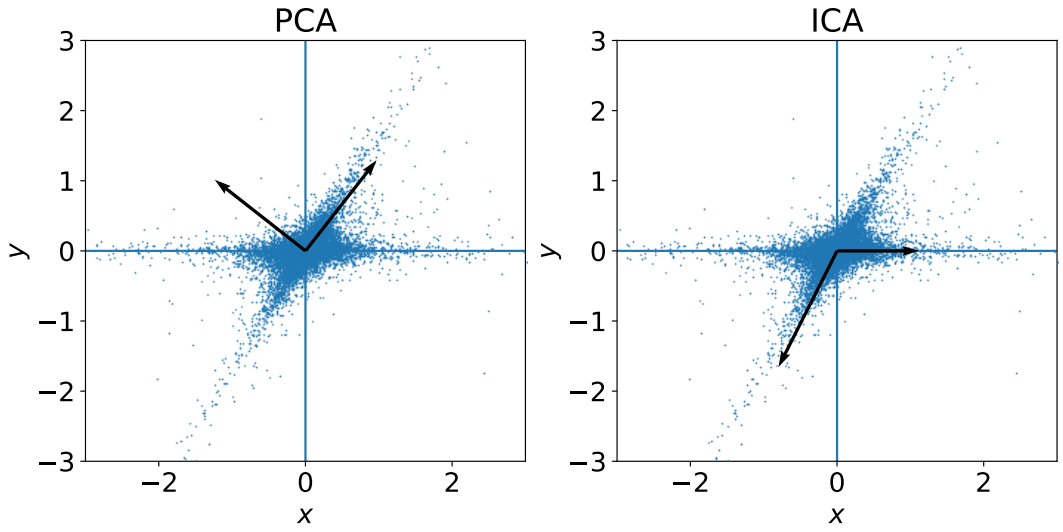

FIGURE 2.6: PCA and ICA applied to a two dimensional dataset consisting of the variable x and y, which are mixtures of two Student-t distributions.

seismogram and calculates a set of hand-designed features such as amplitude ratios for each window. Then, self organizing maps pick the most relevant features and find groups of similar signals based on a similarity measurement with the selected features. With that approach they are able to identify long-term changes in the ambient seismic wavefield and different types of short-term events related to volcanic activity such as volcanic-tectonic earthquakes and rockfalls. Johnson et al. (2020) analyzed continuous seismic time series recorded by a dense array and identified different classes of weak ground motion with methods of unsupervised learning. For a sliding window of 1 s they calculate the spectrograms, retrieve the most relevant features with a principal component analysis (PCA) and group the windows based on a k-means clustering. With that method, they find 5 classes of weak ground motion and study their temporal and spatial occurrence. Snover et al. (2020) analyzed continuous seismic data recorded by a dense seismic array in Long Beach, California, to find signal classes related to urban activity. As in Johnson et al. (2020), they calculate spectrograms for a sliding window. However, instead of using a linear transformation such as PCA to retrieve the most relevant features, they train an auto-encoder to reduce the number of dimensions in a non-linear way. The data points in the learned low-dimensional representation are then clustered with k-means. In a procedure called deep embedded clustering, they learn the auto-encoder and clustering simultaneously by optimizing the clustering and reconstructions loss. Seydoux et al. (2020) clusters continuous single station seismograms recorded in the vicinity of a landslide event in Greenland. Their approach is able to detect blindly precursory events before the landslide occurs. Instead of spectrograms, they utilize scattering coefficients calculated by a learnable deep scattering network. The mentioned studies focused on partitioning clustering and few studies applied hierarchical clustering approaches to seismic data. Unglert, Radić, and Jellinek (2016) tested agglomerative hierarchical clustering on self organizing maps and principal components retrieved from spectrograms in a seismo-volcanic context. They conclude that the hierarchical clustering in combination with PCA is the better choice for seismic signal clustering and data exploration. The above mentioned studies used and explored the whole seismic time series without any pre-selection criteria. However, note that there are

also studies which pre-select certain parts of the time series and apply a feature generation and clustering algorithm only on the selection. This pre-selection can be based on an earthquake catalog (see e.g. Sick, Guggenmos, and Joswig, 2015; Holtzman et al., 2018) or STA/LTA trigger (see e.g. Jenkins et al., 2021). The pre-selection is useful if the signal of interest is known and the pattern recognition tool is supposed to find patterns within that certain class of signals.

How does the presented thesis relate to the state of the art?

The goal of the presented thesis is the data-driven exploration of continuous seismograms but provides an alternative approach with three major differences compared to the mentioned studies. Firstly, we utilize the scattering network for generating a novel representation for seismic time series. Most studies have either utilized hand-designed features or spectrograms as a data representation for waveform clustering. As mentioned earlier, spectrograms are not stable to small signal deformations and hand-designed features are less suited for exploring unknown or poorly defined signals, since they are based on expert and domain knowledge. Therefore, we propose the scattering coefficients, providing a broader description of the signal than the hand-designed features and more stability for small signal deformation than spectrograms.

Secondly, we utilize hierarchical clustering as an interactive exploration tool. Other studies have mainly focused on end-to-end approaches which deliver one cluster solution for interpretation (Snover et al., 2020; Seydoux et al., 2020). Unglert, Radić, and Jellinek (2016) have utilized hierarhical clustering but in our opinion they did not harness its full potential in retrieving multiple clustering solutions with the dendrogram.

Thirdly, we try to interpret the feature space given by PCA or ICA as an additional source for data exploration analysis. Mostly, PCA was applied to retrieve a low dimensional representation for clustering, but its principal components were rarely used for data exploration. To our knowledge, ICA has been utilized for unmixing directly seismic time series (e.g. Ciaramella et al., 2004) but not in the context of feature extraction for seismic data. In general, the proposed strategy introduces concepts new to Seismology and helps exploring the data from different angles without providing the one and only cluster solution.

2.5 Manifold learning for visualization

Both PCA and ICA perform a linear mapping from the high-dimensional input data to the lower-dimensional feature space, while aiming at preserving the pair-wise distances between all data points. Preserving pair-wise distances results in keeping dissimilar data points far apart but it can miss important local structures where similar data points are close to each other, especially, if the high-dimensional data is distributed near or on a complex non-linear manifold. Real world data such as seismograms often describe complex processes with non linear effects and, thus, we can assume that the scattering coefficients are distributed non-linearly on or near a manifold which we do not know. Non-linear methods such as manifold learning can account for this complexity and keep similar points close to each other. There is a large variety of non-linear dimensionality reduction techniques such as self-organizing maps, kernel PCA, t-distributed stochastic neighbor embedding (t-SNE), Laplacian Eigenmaps and many more. A more recent manifold learning technique

is uniform manifold approximation and projection (UMAP), which seems to be very promising tool in visualizing and clustering high-dimensional data.

As a visual introduction to manifold learning and how they compare to PCA, consider Figure 2.7 retrieved from the original paper of UMAP (McInnes, Healy, and Melville, 2018). PCA, t-SNE, UMAP and other non-linear techniques have been applied to various well-known datasets. For example, MNIST is a database containing 70.000 images of handwritten digits from 0 to 9, commonly used for image processing and recognition tasks. These images are vectorized and then fed into one of the methods to retrieve a two-dimensional representation. Data points representing the same number are shown the same color. PCA is able to assign the handwritten digits to different areas but the boundaries are fuzzy and the clusters are not well separated. t-SNE and UMAP do a much better job in clustering the different digits. The same holds true for the other dataset such as Fashion MNIST. This comparison highlights the potential of manifold learning techniques. Without going into further details, the inner workings of UMAP is based on topological data analysis and Riemannian Geometry, providing a complex but safe and sound mathematical background (see the original paper for more details McInnes, Healy, and Melville, 2018). It shares similarities with t-SNE, which has been used extensively for visualizations since its appearance in the 2000s (Maaten and Hinton, 2008). However, compared to UMAP, t-SNE performs poorly in preserving global structures and its computation time is much slower (Becht et al., 2019). We mention these techniques more as an outlook and will show an application of UMAP later in Chapter 5.

2.5.1 Hyperparameters of UMAP

Manifold learning techniques such as UMAP come with a set of hyperparameters to tune, which impact the found embedding. As with PCA and ICA one hyperparameter is the number of dimensions or components. The more dimensions we keep, the more information we retrieve. The other hyperparameters are the number of neighbors and the minimum distance, which draw the focus either towards preserving local or global structures. In the following we provide a very short intuitive description of what these parameters control.

Number of neighbors

This parameter limits the number of neighboring points when UMAP learns the local manifold structure. A low number draws the focus to local structure while losing the bigger picture. A large number draws the focus on the global structure while losing finer details.

Minimum distance

This parameter controls how closely UMAP is allowed to bring data points together. A low number results in a more dense and clumpier representation and preserves better the local structure of the data. A large number avoids putting points close to each other and draws a broader picture of the data.

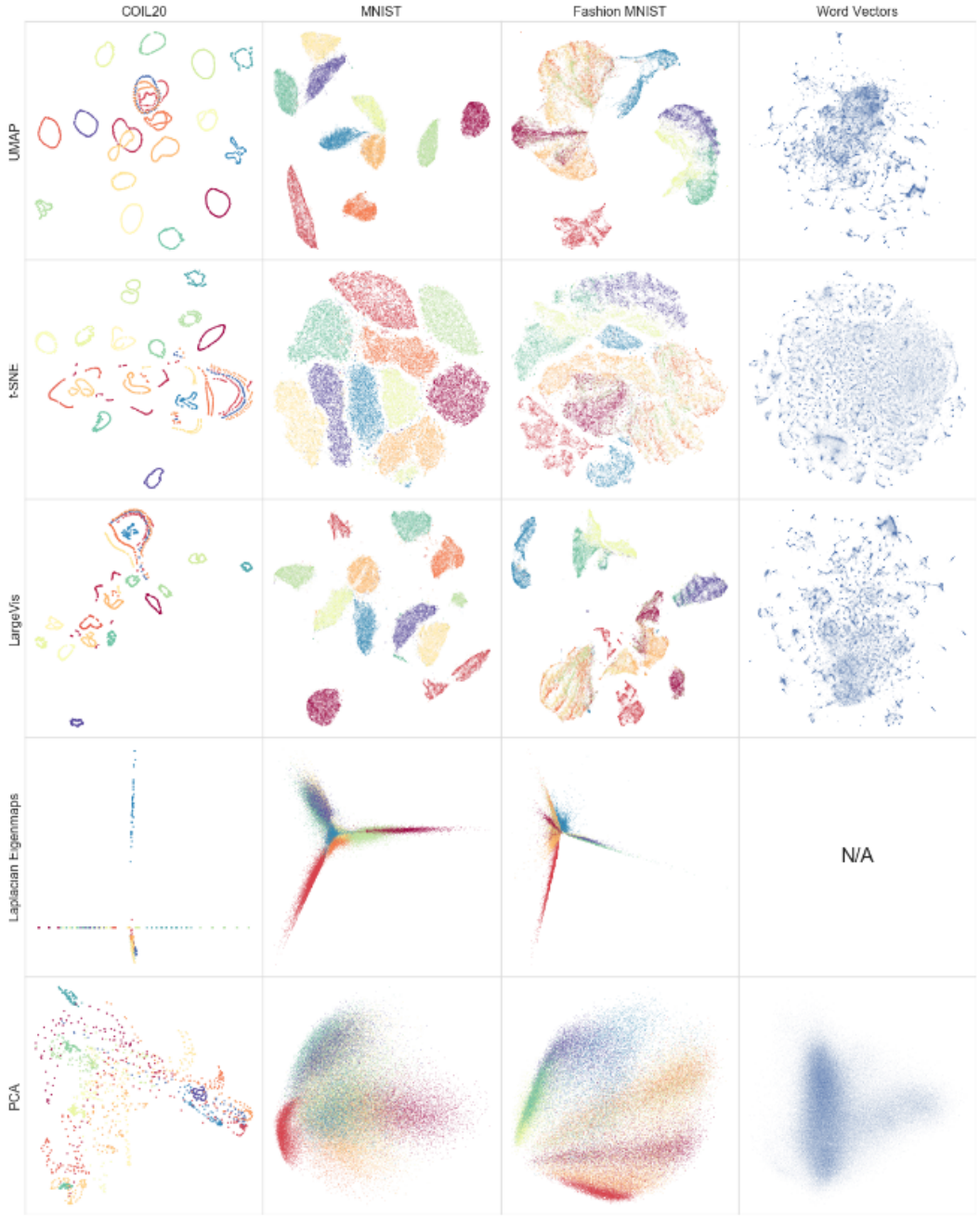

FIGURE 2.7: Different linear and non-linear techniques for dimensionality reduction applied to various databases, retrieved from McInnes, Healy, and Melville (2018)

Chapter 3

Hierarchical exploration of continuous seismograms with unsupervised learning

René Steinmann, Léonard Seydoux, Michel Campillo
Article published in JGR: Solid Earth

The following chapter covers the first application of hierarchical waveform clustering to single station data recorded in the vicinity of the North Anatolian Fault, Turkey. This work marks the beginning of my PhD which started with the learnable scattering network presented in Seydoux et al. (2020). We realized that non-learnable Gabor wavelets preserve already most of the interesting structure in the seismic data without the need for learning the wavelet. Therefore, we moved away from the end-to-end approach and turned towards a more exploratory analysis with hierarchical clustering. We chose this dataset since it contains a burst of similar small magnitude earthquakes (a repeating pattern with low SNR) and different types of seismic signals with anthropogenic origin. At this stage, we did not yet focus much on the temporal patterns revealed by ICA, but we already noted that these independent components contain interesting information and could be used for data exploration. Time moved on since then and personally, I would rather disagree with some statements given in this article. For example, we describe the Ward's method as adapted to the seismic signal class distribution, which I would not totally agree with today. Ward's method definitely delivers interesting results, but its mathematical definition favors even-sized Gaussian-distributed clusters. In particular the feature space retrieved with ICA does not provide cluster shapes which are Gaussian.

3.1 Abstract

Continuous seismograms contain a wealth of information with a large variety of signals with different origin. Identifying these signals is a crucial step in understanding physical geological objects. We propose a strategy to identify classes of signals 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. Based on the extracted features, agglomerative clustering then groups these waveforms in a hierarchical fashion and reveals the process of clustering in a dendrogram. We use the dendrogram to explore the seismic data and identify different classes of signals. To test our strategy, we investigate a two-day-long seismogram collected in the vicinity

of the North Anatolian Fault, Turkey. We analyze the automatically inferred clusters' 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 burst that includes around 200 events with similar waveforms and high-frequent signals with correlating envelopes and an anthropogenic origin. The application shows that the cluster hierarchy helps to identify particular families of signals and to extract subclusters for further analysis. This is valuable when certain types of signals, such as earthquakes, are under-represented in the data. The proposed method may also successfully discover new types of signals since it is entirely data-driven.

3.2 Introduction

Continuous seismograms contain a rich amount of information as a large variety of signals can be observed therein. Determining the origin of these different signals is crucial in understanding the physical geological objects. For example, faults and plate boundaries accommodate the tectonic loading by releasing energy in different fashions (Ide et al., 2007), the most known and well-understood signals being earthquakes, radiating seismic waves visible in most seismograms. Based on their signal characteristics, seismologists developed many tools to detect earthquakes in seismograms such as the short time average to long term average STA/LTA (e.g. Allen, 1978). Only 20 years ago, a new signal with tectonic origin has been discovered and designated as a non-volcanic tremor because of the similarities with volcanic tremors (Obara, 2002). However, non-volcanic tremors are often of weak amplitude with poorly defined signal characteristics; their detection is a more challenging task than detecting earthquakes. Other than signals with tectonic origin seismometers also record the oceanic microseisms (for a recent review see e.g. Ebeling, 2012), rockfalls and other mass movements (e.g. Lacroix and Helmstetter, 2011; Deparis et al., 2008), ground and air traffic (e.g. Riahi and Gerstoft, 2015; Meng and 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. Since the 1970s seismology benefits from artificial intelligence developments, bringing machine-learning-based solutions for exploring seismic data and recognizing patterns (e.g. Allen, 1978). More recently an unsupervised learning strategy called clustering was utilized to explore seismic data and find families of similar signals (Köhler, Ohrnberger, and Scherbaum, 2010; Holtzman et al., 2018; Mousavi et al., 2019; Seydoux et al., 2020; Johnson et al., 2020; Snover et al., 2020; Jenkins et al., 2021). In contrast to supervised learning strategies, clustering does not rely on a labeled training set and human expert knowledge (Goodfellow, Bengio, and Courville, 2016). Thus, clustering seismograms can help identifying families of signals which are not yet discovered or are poorly defined such as non-volcanic tremors.

In the present paper, we introduce a new strategy to use clustering as an exploration tool for continuous seismograms. Our strategy follows the idea that seismic signals are grouped in a hierarchy of classes following a specific similarity measurement, as schematized in Figure 3.1. Note that this illustration aims at sketching

the concept rather than being complete or accurate. We consider the similarity between classes of signals to be measured on a set of signal characteristics that can be human-defined (such as mean frequency and signal duration) or learned with machine-learning tools, as we propose 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 3.1). 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 signal characteristics may be relevant at each node of the tree.

The sketch presented in Figure 3.1 also illustrates the problems of designing a class 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 is hard to design by hand. Hierarchical clustering produces precisely this kind of tree, called a dendrogram, based on the exploration of the similarity of signals present in the input data. Therefore, we propose to represent continuous seismograms as a dendrogram and utilize it to explore the content of the data and identify different types of seismic signals. We want to emphasize that clustering identifies groups of similar objects, but it does not provide any meaningful labels, such as the labels in Figure 3.1. In an extra step, the found clusters can be labelled by analyzing the inherent properties of the clusters.

In the following section, we present the workflow to build a dendrogram from continuous single-station data. We introduce the concept of hierarchical clustering and how we transform continuous seismograms to a meaningful input (features) for the hierarchical clustering. In section 3, we introduce a data set to apply and test the proposed 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.

3.3 Method

A sketch of the hierarchical clustering workflow is depicted in Figure 3.2. In the following lines, we start with the concept of clustering in general and hierarchical clustering in particular. Then, we explain how we transform seismograms into a meaningful input for the cluster analysis.

3.3.1 Hierarchical clustering

In general, cluster analysis groups objects based on their similarity to each other (e.g. Xu and Wunsch, 2008). Objects in the same cluster are more similar to each other than objects in separated clusters. The similarity between objects is measured on a set of certain characteristics called features. Finding the most relevant features for this task will be discussed later.

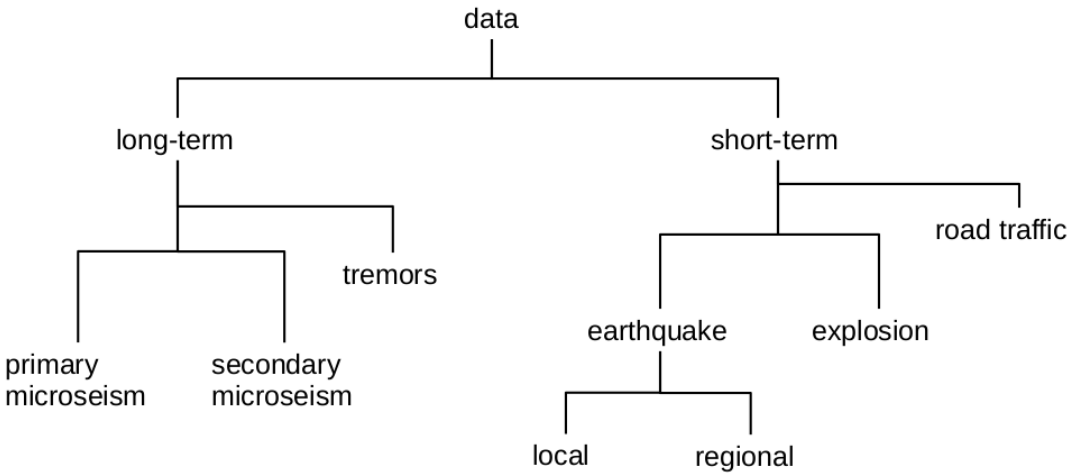

FIGURE 3.1: Illustration of a possible hierarchy of seismic signals found in seismograms. The different branches represent how a signal class splits into different sub-classes depending on a given similarity measure. Here the different classes of signals are thought in a hierarchical way, based on arbitrary 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.

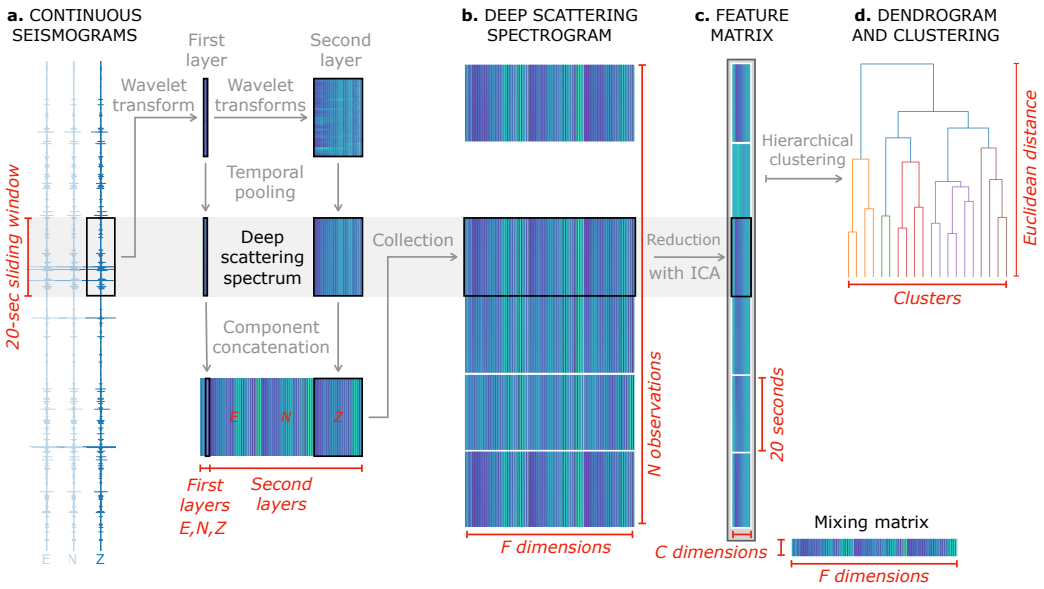

FIGURE 3.2: **Proposed workflow for hierarchically exploring continuous seismograms**. **(a)** Input continuous three-component seismograms, as detailed in Section 3.4. **(b)** Deep scattering spectrogram of the seismograms, with a temporal resolution of about 20 s and a high number of dimensions, detailed in Section 3.3.2. **(c)** The feature matrix extracted from the deep scattering spectrogram with independent component analysis, following the description in Section 3.3.3. **(d)** Dendrogram calculated from a similarity measurement in the feature space, as explained in Section 3.3.1.

Various algorithms exist to find groups of objects in a data set. This study utilizes hierarchical clustering with a bottom up approach, namely agglomerative clustering. Hierarchical clustering relies on a similarity matrix, which defines the similarity (e.g., a specific distance in the feature space) between all objects in a data set (Johnson, 1967). With a bottom-up approach, all objects start in a singleton cluster. The clusters start merging based on the similarity matrix until all objects 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 as depicted in Figure 3.1.

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 objects x_i into clusters such that the within-cluster variance remains minimal after merging different clusters. The within-cluster variance σ quantifies the spread of each cluster in the feature space (for more details see Appendix 3.A of this chapter). By minimizing the overall variance, $\sum_{c=1}^{K} \sigma_c$ with K being the number of clusters, the Ward's method allows for clusters of variable population sizes and variances. Thus, it 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, where often ambient seismic noise outweighs signals with a tectonic origin.

It is worth mentioning that hierarchical clustering especially with the Ward's method can be computationally expensive. However, algorithms have been improved over time and became more efficient. In this study we utilize the python package **fastcluster**, which has a time complexity of $\mathcal{O}(N^2 D)$ with N elements in $\mathbb{R}^D$ and a memory complexity of $\mathcal{O}(ND)$ (Müllner, 2013). More recently, the use of hyperbolic embeddings for preserving the hierarchical structure of the data seems to be a promising way to reduce even further the computational costs (Chami et al., 2020).

3.3.2 Finding an appropriate representation of seismograms: the deep scattering spectrum

In order to detect and identify classes of signals in continuous seismograms with hierarchical clustering, the seismograms have to be transformed into a meaningful input for the cluster analysis. For that purpose, we calculate features for fixed windows of the seismogram. Thus, each window will be assigned a cluster based on the features for this window. Note that this process simplifies the complexity of seismic data, since multiple types of signals can occur simultaneously. Common cluster analysis such as hierarchical clustering neglect this fact and can only assign a single cluster to an object. Besides the choice of the applied metric within hierarchical clustering, the choice of features is another important factor, which determines the outcome of the cluster analysis. 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, Bengio, and Hinton, 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.

A time-frequency representation such as the spectrogram is one way to create a set of features for classifying seismic signals (Johnson et al., 2020; Snover et al., 2020; Jenkins et al., 2021). 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 (see Figure 3.2a and b). 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. The authors of 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 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 (i.e. wavelet transforms):

$$S^{(\ell)}\left(t, f_{n_1}^{(1)}, f_{n_2}^{(2)}, \ldots, f_{n_\ell}^{(\ell)}\right) = \max_{[t,t+dt]} \left|\phi^{(\ell)}\left(f_{n_\ell}^{(\ell)}\right) \star \left|\ldots \star \left|\phi^{(2)}\left(f_{n_2}^{(2)}\right) \star \left|\phi^{(1)}\left(f_{n_1}^{(1)}\right) \star x\right|\right|\right|\right|, \tag{3.1}$$

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. In contrast to the Fourier transform, the center frequencies of the wavelets are placed logarithmically. In this study, we only consider a scattering network with 2 layers (as depicted in Figure 3.2) since Andén and Mallat (2014) argued that more layers do not necessarily introduce new valuable information. The first layer in the network creates N_1 scalograms per channel of the seismic station. In the second layer another wavelet transform is applied to each scalogram of the first layer. Thus, the second layer contains $N_1 * N_2$ scalograms. A maximum pooling operation is then applied over each scalogram to retrieve the scattering spectrum. The entries of the scattering spectrum pooled from the first layer scalogram are called first-order scattering coefficients. The entries of the scattering spectrum pooled from the second layer scalogram are called second-order scattering coefficients, which contain important information about the attack and modulation of a signal. While the scalograms still

have the same sampling rate in time as the input data, the temporal pooling collapses the time dimension of the scalogram and produces a scattering spectrum for each input window. Note that each input channel from the seismic station is treated separately and their deep scattering spectrum are concatenated into a deep scattering spectrum vector with the size $3 * N_1 + 3 * N_1 * N_2$ for a three-component seismogram. For each input window, a deep scattering spectrum vector is created, which are then merged into the deep scattering spectrogram. The final sampling rate of the deep scattering spectrogram is defined by the size of the input window. In Seydoux et al. (2020), the authors initialize Gabor wavelets with amplitudes and derivatives on a certain sets of knots and interpolate then with Hermite cubic splines. With respect to the clustering loss, they learn the parameters on these knots governing the shape of the wavelets. In this study, we directly use the initialized Gabor wavelets with zero phase shift and do not apply any learning of the wavelets. 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. For the interested reader we refer to Andén and Mallat (2014) and Seydoux et al. (2020).

3.3.3 Features extraction from deep scattering spectrogram

The deep scattering spectrum can have more than 1,000 dimensions and, thus, the conditions for clustering are not favorable (Kriegel, Kröger, and Zimek, 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 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:

$$\mathbf{x} = \mathbf{s}\mathbf{A}, \tag{3.2}$$

where $\mathbf{x} \in \mathbb{R}^{N \times F}$ are the N observations of dimension F, $\mathbf{A} \in \mathbb{R}^{F \times C}$ is the mixing matrix, and $\mathbf{s} \in \mathbb{R}^{C \times N}$ are the unmixed sources (namely, the C unmixed sources obtained from ICA). The observations $\mathbf{x}$ are therefore a linear combination of the independent sources $\mathbf{s}$, with the mixing weights gathered in $\mathbf{A}$. A test of statistical independence is required to solve Equation 3.2 while ensuring the sources $\mathbf{s}$ to be independent. This concept is illustrated in Figure 3.2, where the unmixed sources are considered as features in our workflow (therein called feature matrix). These sources are obtained from applying the unmixing matrix, the pseudo inverse of the mixing matrix $\mathbf{A}$, to the deep scattering spectrogram. 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 and 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.

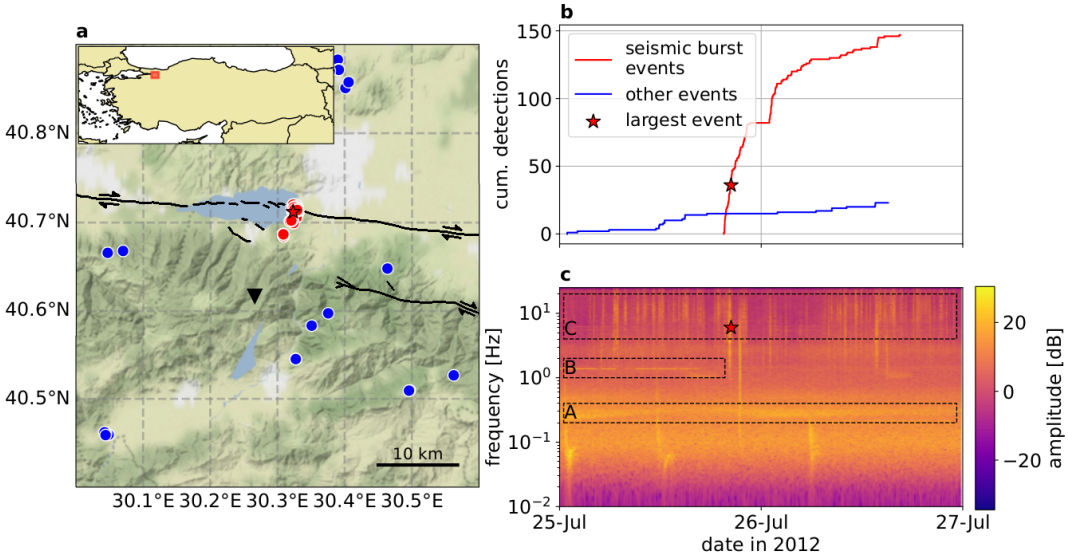

FIGURE 3.3: Geological context and seismic data used in the present study. **(a)** Map of the North Anatolian fault zone showing station DC06 (black triangle), the seismic burst (red dots) including the largest event (red star) and other seismic activity (blue dots); all detected with a template matching strategy. The geological faults that ruptured after 1900 (black lines) are adapted from Emre et al. (2011). **(b)** Cumulative detections of the seismic burst (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.

3.4 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 3.3a). Originally, the experiment was conducted to investigate the crustal structure beneath the western segment of the North Anatolian Fault. 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).

We choose to analyze the seismic data from the 25th to the 27th of July 2012. During the period of these two days, a high rate of localized seismicity with 148 cataloged events occurred on and around the northern strand of the North Anatolian fault (see Figure 3.3a and b). In this study, we refer to this high rate of seismicity as a seismic burst. The catalog explains the series of events with 17 templates having their hypocenters close to each other (Figure 3.3a, red dots). Since the seismic burst causes a repeating pattern in the seismogram with short time-warping deformations due to slight changes of the hypocenters, it is an interesting study case for our proposed method. Station DC06 is close to the seismic burst and records the time period of interest without data gaps. Thus, we choose the three-component seismograms of this station. The sampling rate of the data is 50 Hz.

The spectrogram of the east component of station DC06 is presented in Figure 3.3c. 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 seismic sources. The event with the largest magnitude of the burst is also easy to spot during the evening of the 25th in the spectrogram.

3.5 Results

3.5.1 Feature space

Firstly, we use the continuous three-component seismograms to calculate the deep scattering spectrogram with a two-layered scattering network (as detailed in Equation 3.1). The network parameters are physics-driven and can be adjusted according to the goal. In this study, the first layer contains 24 Gabor wavelets with center frequencies between the Nyquist frequency of the seismogram (25 Hz) and 0.78 Hz with a spacing of 4 wavelets per octave. The second layer contains 14 Gabor wavelets with center frequencies between 25 Hz and 0.19 Hz with a spacing of 2 wavelets per octave. This setup results in 24 wavelet transforms per channel in the first layer and 336 ($24 * 14$) wavelet transforms per channel in the second layer. Because the deep scattering spectrum is a concatenation of the first- and second-order scattering coefficient of each input channel, the total number of scattering coefficients is 1080 (dimension F in Figure 3.2). For the temporal pooling operation, we apply maximum pooling, since we are interested in detecting and classifying non-stationary events such as the seismic burst. If the focus of classification is the background noise, average pooling might be the better choice (as suggested in Seydoux et al., 2020). The moving pooling window is 20.48 s large and does not overlap. Hence, the time resolution of the deep scattering spectrogram is also 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 3.B of this chapter). 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 3.B and Figure 3.B.1 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 unmixed sources building the feature space is depicted in Figure 3.B.2 in Appendix 3.B of this chapter.

3.5.2 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. The distance threshold sets the boundaries for the possible distances between points within a cluster. While a larger distance threshold allows larger and fewer clusters to form, a smaller distance threshold extracts smaller but many clusters. Note

that the distance threshold is only used to extract different cluster solutions based on the similarity matrix; it is not a hyperparameter affecting the similarity matrix. In Figure 3.4a 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 3.4b shows the averaged first-order scattering coefficients of these four clusters. These first-order scattering coefficients describe the frequency characteristics of each cluster. Figure 3.4c presents the normalized cumulative detection rate of each cluster, with the seismic burst 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 3.4d. 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 3.4d) 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 3.4a). Compared to the other clusters, its scattering coefficients for all frequencies are relatively low except for a local maximum around 1.5 Hz (Figure 3.4b). Looking at the corresponding cumulative detection curve (Figure 3.4c), 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 3.3c).

Cluster B contains about 19 % of the data samples (Figure 3.4d) and has relatively large scattering coefficients for frequencies above 10 Hz (Figure 3.4b). 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 3.4c). 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 3.4d). Compared to the other clusters, it also has the lowest scattering coefficients at all frequencies (Figure 3.4b). Looking at the cumulative detection curve (Figure 3.4c), 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 contains mostly ambient seismic noise data.

Finally, cluster D contains about 4 % of data set (Figure 3.4d) and is the smallest of the four clusters (Figure 3.4d). The corresponding first-order scattering coefficients show a local maximum around 5 Hz (Figure 3.4b). Its cumulative detection curve correlates well with the detections of the seismic burst (Figure 3.4c), with additional detections before the seismic burst starts. All these observations indicate that cluster D is probably related to nearby seismic activity in general.

3.6 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 burst 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. To underpin the statement that the deep scattering spectrum is a superior representation for the task at hand than the Fourier-transform spectrum, we also create and interpret a dendrogram based on the Fourier-transform of the same data set (see Appendix 3.D of this chapter).

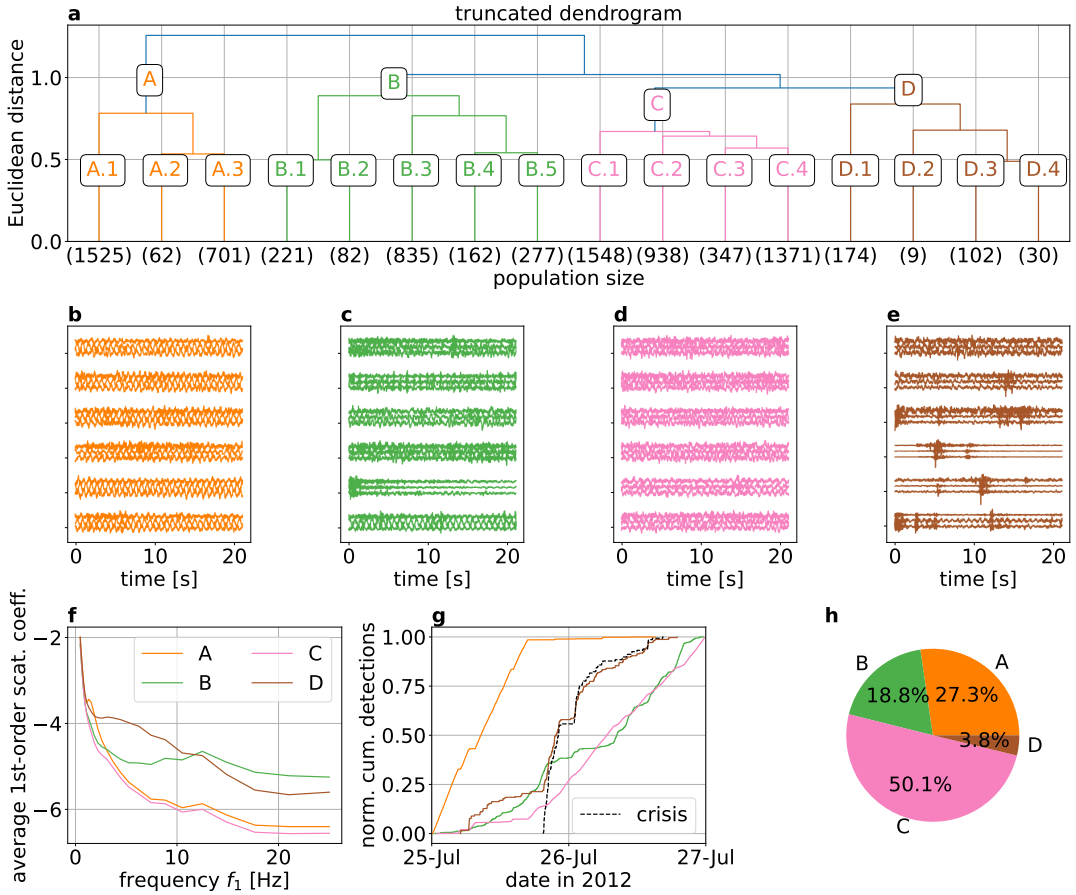

FIGURE 3.4: Dendrogram analysis and statistical characteristics of the different clusters. **(a)** Dendrogram calculated in the feature space (see Sec. 3.3.1 for explanations). The dendrogram is here truncated in order to form 16 clusters. The clusters marked with a letter are considered the main clusters, and the subclusters are indicated with numbers. The numbers in the parenthesis indicate the number of samples in each cluster. **(b, c, d and e)** depict random examples of waveforms for the four main cluster A,B,C and D, respectively. **(f)** Averaged first-order scattering coefficients for main clusters A, B, C and D. **(g)** Normalized cumulative detections of main clusters A, B, C and D, and of the seismic burst obtained from the multi-station template-matching catalog. **(h)** Relative size of the main clusters compared to the size of the entire data set.

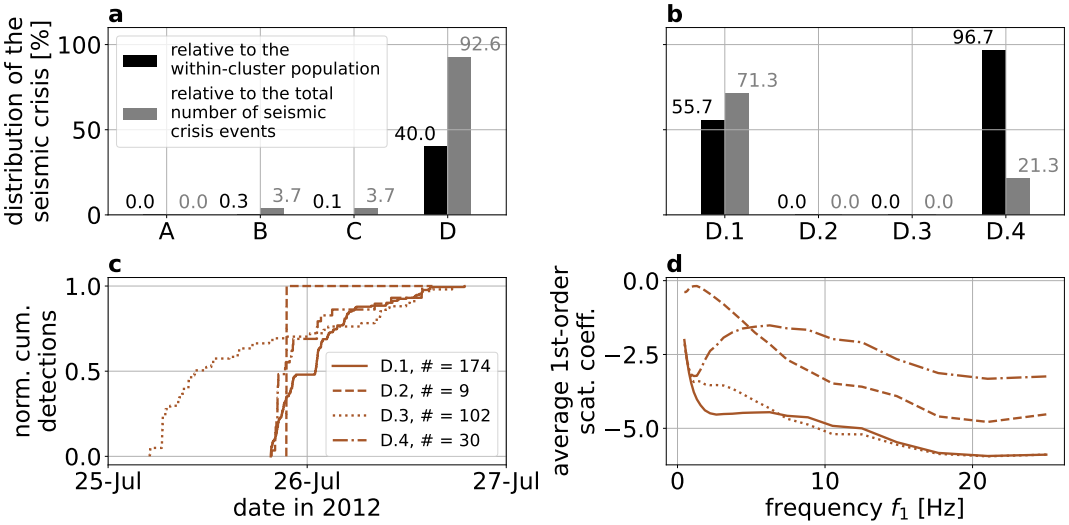

FIGURE 3.5: Identification of the seismic burst within the main and subclusters. **(a)** The distribution of the seismic burst across the four main clusters. **(b)** The distribution of the seismic burst across the four subclusters in the main cluster D. **(c)** Normalized cumulative detection curves for the subclusters in the main cluster D. **(d)** Averaged first-order scattering coefficients for the subclusters in the main cluster D.

3.6.1 Identification of the seismic burst within the dendrogram

Firstly, we identify all time segments containing onsets of the events of the seismic burst and observe which clusters those time segments belong to. The template matching catalog contains 148 detections related to this seismic burst. However, we only associate 136 samples in the feature space with the seismic burst, since one sample represents about 20 s of waveform data and, thus, can contain multiple events. Figure 3.5a shows that a large majority of the samples, which contain arrivals of the seismic burst, fall into cluster D (92.6 %). On the other hand, only 40 % of cluster D is related to the seismic burst, underpinning the statement that this cluster is related to general seismic activity. Cluster B and C share the remaining 7.4 % of the burst. Compared to the large population sizes of clusters B and C, the contribution of the burst almost vanishes (0.3 and 0.1 %). Cluster A contains no detections of the burst. While cluster D contains the majority of the seismic burst, 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 3.4a 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 burst across the four subclusters in main cluster D. From Figure 3.5a, we know that more than 92 % of the burst was found in cluster D. We observe in Figure 3.5b 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 burst and will be discussed later. If we look at the cumulative detection curve of each subcluster in D (Figure 3.5c), we see that cluster D.1 and D.4 share a very similar temporal pattern. The corresponding averaged first-order scattering coefficients (Figure 3.5d) explain why the burst 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 magnitude of the events seems to be the characteristic that separates the burst into two clusters. Besides, we observe that 56 % of D.1 and 97 % of D.4 can be explained by the cataloged burst. This observation raises the question: what are the samples in D.1 and D.4 that cannot be related to the seismic burst 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 3.6a, b and c show the corresponding waveforms of all 204 data points of the two subclusters D.1 and D.4. For presentation purposes we align the waveforms accordingly to their maximum correlation with a template waveform from the subcluster. For all waveforms we observe the *P* and *S* seismic phase 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 burst is split into two subclusters due to magnitude differences. The magnitude estimations of the template matching catalog confirms this assumption (see Figure 3.6d). 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 burst. 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 3.C of this chapter.

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 136 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 3.7a). In magenta, we highlight the distance between the sample and its respective subcluster. In cyan, we highlight the distance between the sample and subcluster D.1 containing the low magnitude events of the burst. 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 3.7b 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

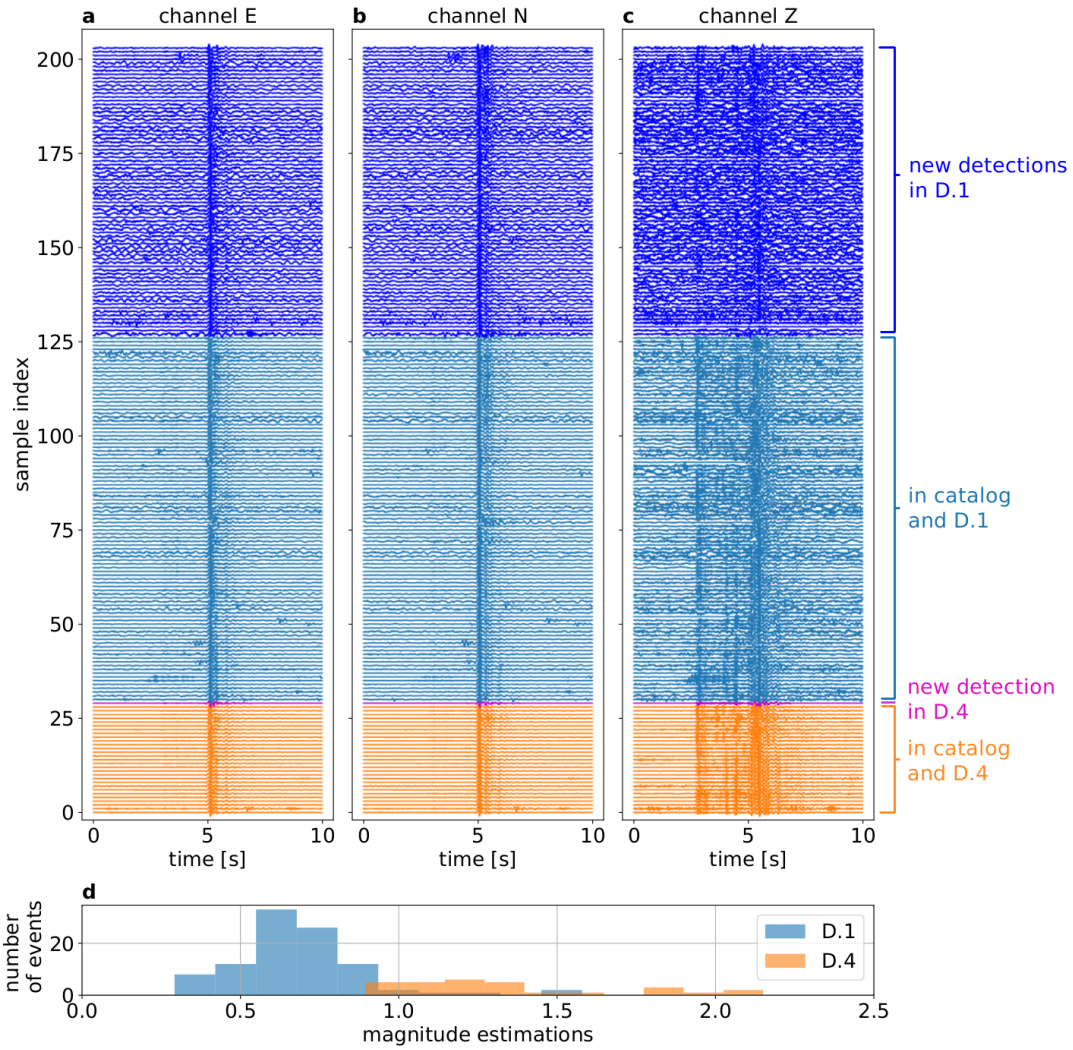

FIGURE 3.6: **(a,b,c)** Waveform data from subcluster D.1 and D.4. The color code indicates the according subcluster and if the event is mentioned by the catalog. **(d)** Magnitude estimations of the cataloged events of the seismic burst found in subcluster D.1 and D.4.

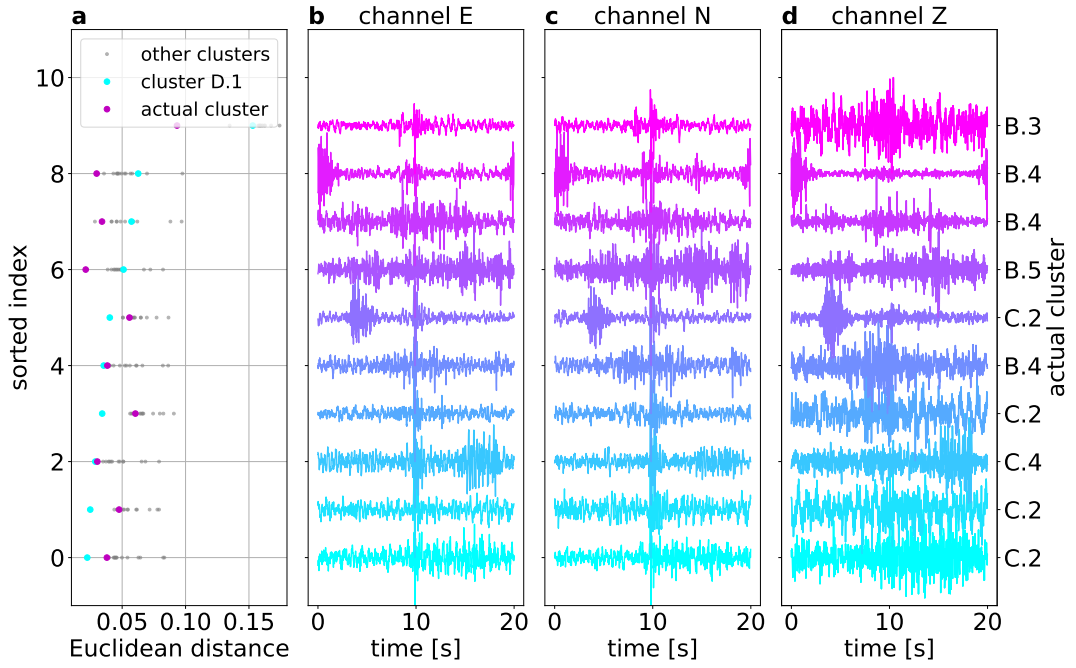

FIGURE 3.7: Analysis of the misidentified earthquake waveforms. **(a)** Distances between misidentified data points containing an event from the catalog and the centroids of all clusters. The magenta points show the distance between the data point and the centroid of its own respective subcluster. The 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.

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 assigning a single cluster to 20 sof waveform data, which can contain multiple signals.

3.6.2 Neighboring clusters of the seismic burst in the feature space

Having identified most of the seismic burst 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 burst.

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 3.8. Figure 3.8a indicates the samples which can be explained by arrivals of a regional or local event, and Figure 3.8b shows the samples which can not be explained by arrivals of a regional or local event. Note that one sample in the feature space represents ca. 20 s of waveform data and each horizontal waveform displayed in Figure 3.8 contains multiple consecutive 20 s windows. Subcluster D.2 contains only nine samples corresponding to two seismic events indicated in blue in Figure 3.8a. 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 3.5d), 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 3.A.1 in Appendix 3.A of this chapter. Moreover, the within-cluster variance σ_c in the top panel of Figure 3.A.1 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 feature space due to their high amplitudes at lower frequencies.

Moreover, we observe that the catalog can explain 67 % of all samples of D.3 (a random selection of waveforms are shown in Figure 3.8a). The other 33 % are shown in Figure 3.8b, 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 3.8b 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 3.5d). 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 burst, other subclusters containing seismic activity can be found.

3.6.3 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 and Mallat, 2014). Consequently, we should find clusters with weakly correlating waveforms but strongly correlating envelopes.

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 analytic signal, which is a complex-valued representation of the waveform disregarding the negative frequencies from the Fourier transform. A median-filter smoothens the envelope. The averaged results are depicted in Figure 3.9a. We firstly observe that CC_E is more

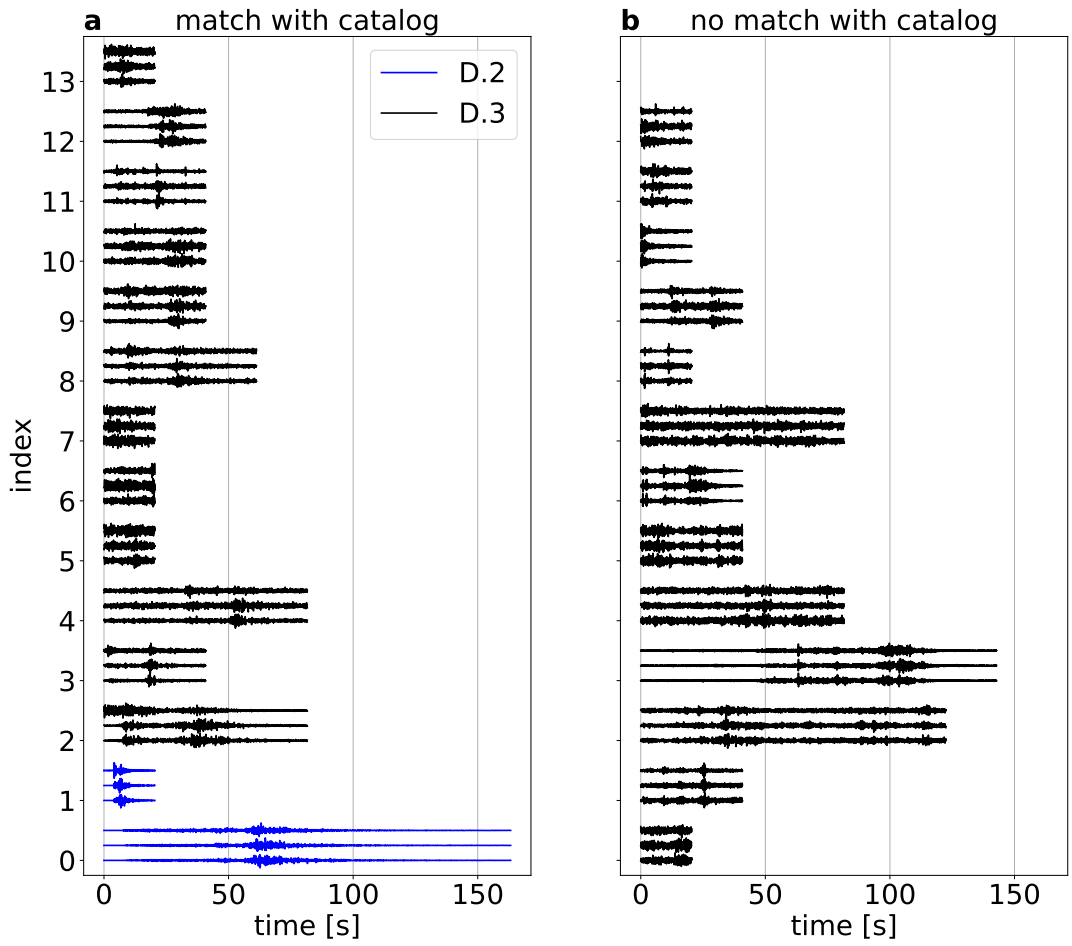

FIGURE 3.8: Seismic waveforms identified in subclusters D.2 and D.3. **(a)** waveform data of D.2 and D.3 where the phase arrivals match the merged catalog. **(b)** waveform data of D.3 which do not correspond to phase arrivals from the merged catalog.

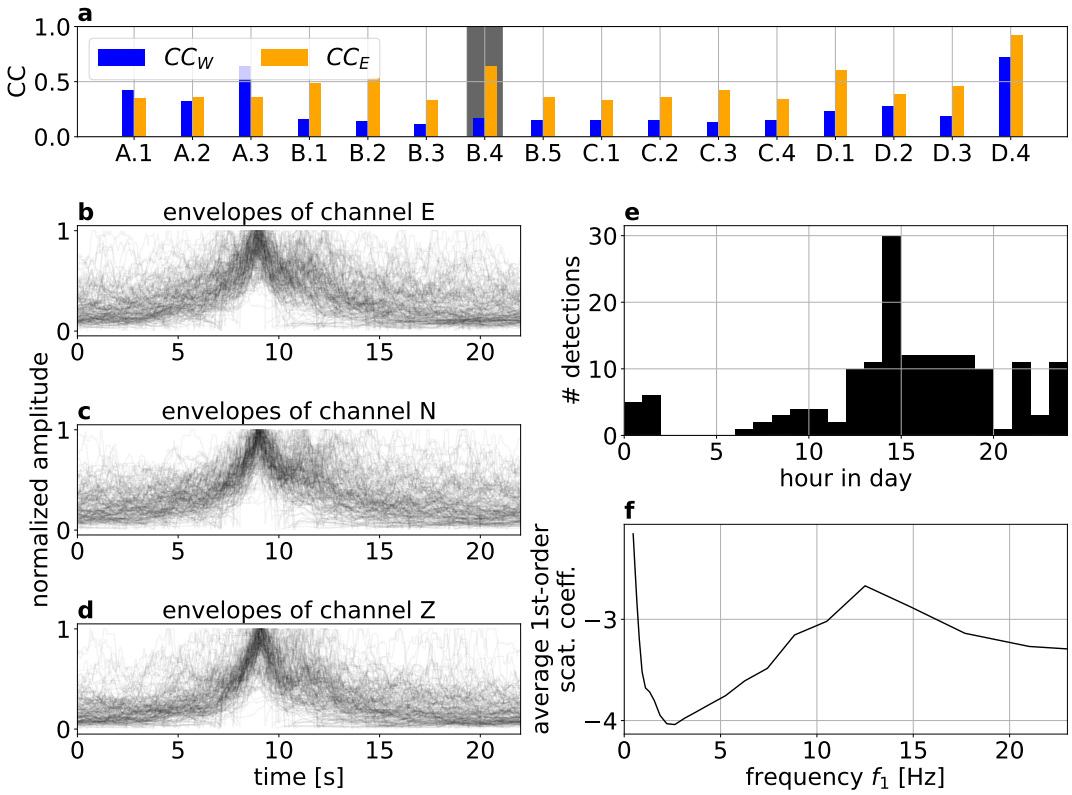

FIGURE 3.9: Interpretation of subcluster B.4. **(a)** Averaged correlation coefficient for the waveforms CC_W and for the envelopes CC_E for all 16 subclusters. **(b,c,d)** Aligned envelopes for the three channels for subcluster B.4. **(e)** Number of detections per hour for subcluster B.4. **(f)** Averaged first-order scattering coefficients for subcluster B.4.

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 3.9b 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 3.9e 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 3.9f 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.

3.6.4 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 3.9). 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 3.4b 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 3.A.1 in Appendix 3.A of this chapter). This indicates that they contain similar signals. Combining these facts,

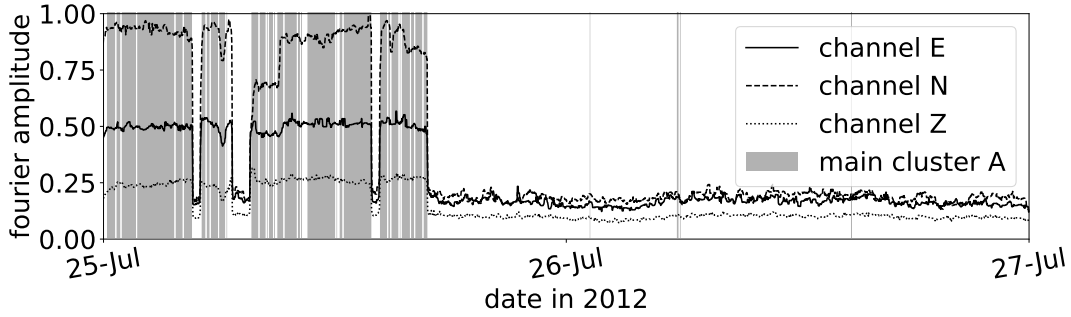

FIGURE 3.10: Fourier amplitude of all three channels calculated over 10 min windows in the frequency range of 1.4 to 1.6 Hz together with the activation of the main cluster A

we conclude that this cluster contains ambient noise without any significant activity of transient signals.

Cluster A seems to correlate with the monochromatic noise source around 1.5 Hz (see Figure 3.3c and 3.4c). 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 3.10. 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. As with the misidentified events in Figure 3.7, the problem is related to assigning a single cluster to 20 s of waveform data containing multiple types of signals. It is also interesting to note that subcluster A.1 and A.3 show larger correlation coefficients for the waveforms than for the envelopes (Figure 3.9a). 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.

3.7 Conclusion

In this study, we proposed a new way of exploring the content of continuous seismograms and identifying different types of signals present in the data. Our approach is based on hierarchical clustering, which offers many cluster solutions with the dendrogram and, thus, delivers a tool for exploring the data. The hierarchical clustering is applied to a low-dimensional feature space extracted from the deep scattering spectrogram of the continuous seismogram. A primary advantage of the workflow compared to other machine learning algorithms for classifying continuous seismic data is the interpretability at each step and the deep scattering spectrum, which seems to be a promising representation of seismic data for classification purposes.

For an application in this study, we chose a 2-day long three-component seismogram containing a nearby seismic burst with 148 cataloged events with similar waveforms. These labels served as a sanity check for the algorithm. Firstly, we extracted a cluster solution with four main clusters to get a rough overview of the data. With the cluster size, the temporal detection, and averaged first-order scattering coefficients, we delivered an interpretation of each cluster and could identify a cluster containing mostly waveforms related to earthquakes. Inside this specific

cluster, we found two subclusters containing almost all cataloged events of the seismic burst. While the events of the seismic burst split into two subclusters due to magnitude differences, 77 uncataloged events with similar waveforms were found. The case of the seismic burst shows that we can identify a repeating pattern with slight variations of the waveforms and low SNR in an unbalanced data set. The few misidentified events highlight the multi-label characteristics of seismograms. Multiple signals can arrive simultaneously and, thus, assigning a single label to a part of the seismogram does not reflect the whole truth. Integrating this issue into clustering seismograms is an interesting aspect for future work. Besides the seismic burst, we also identified signal families with anthropogenic origin and a large cluster containing ambient seismic noise. The different types of signals show that the strategy is able to group signals with correlating waveforms, envelopes or similar frequency characteristics.

We want to emphasize here that hierarchical clustering and the dendrogram itself does not deliver meaningful labels for the clusters. Interpreting the different cluster solutions with certain characteristics such as the temporal detection curve is a crucial step towards understanding and revealing the content of the data. Until the point of hierarchical clustering, the proposed workflow is an unsupervised and data-driven strategy to find groups of similar seismic signals. After that point, we use the output of that strategy to do an interpretation and assign meaningful labels to the retrieved clusters.

As most machine learning algorithms, the proposed strategy relies on a few parameters to tune. The hyperparameters of the deep scattering network are mainly physics-driven and depend on the pre-defined task. As with Fourier spectrograms, we can control the window size and frequencies of interest. For example, low frequent first-order wavelet filters might not be necessary for finding groups of anthropogenic signals. Maximum pooling is more interesting than average pooling if the signals of interest have a transient character such as earthquakes. After designing the deep scattering network, the number of components in the independent component analysis is an exploratory task. It is a trade-off between keeping crucial information and producing a low-dimensional representation to avoid the curse of dimensionality.

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 temporal detection curves of the clusters can be easily correlated with other time series such as GPS displacement or environmental parameters to search for signal classes related to certain physical processes. A specific interesting application would be the North Anatolian Fault, where seismologists assume the presence of non-volcanic tremors but conventional methods did only deliver null results so far (Pfohl et al., 2015; Bocchini et al., 2021). In contrast to conventional tremor detection algorithm, our approach could identify signals related to tectonic processes without assuming any signal characteristics. In the same sense, 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 method can be helpful to extract particular types of noise for performing ambient noise cross-correlation potentially enhancing the signal quality.

3.8 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://doi.org/10.5281/zenodo.5518136`. The python packages `ObsPy`, `SciPy` and `Scikit-learn` were heavily used for processing the data (Beyreuther et al., 2010; Virtanen et al., 2020; Pedregosa et al., 2011b). 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. Moreover, the authors want to thank the editors, Andrew Valentine and an anonymous reviewer for their time and effort to improve the study in many ways.

Appendices

3.A 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:

$$d_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2, \tag{3.3}$$

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:

$$\sigma_i = \frac{1}{N_i} \sum_{n \in i}^{N_i} \|\hat{\mathbf{y}}_n - \boldsymbol{\mu}_i\|_2^2. \tag{3.4}$$

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 3.A.1.

3.B Number of relevant independant components

Setting the number of dimensions for a dimensionality reduction technique such as the ICA 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 $\hat{\mathbf{x}}^{(C)}$, obtained from Equation 3.2 with C independent components, as

$$\epsilon(C) = \frac{\sum_{i=0}^{N} |x_i - \hat{x}_i^{(C)}|}{N}. \tag{3.5}$$

Figure 3.B.1 depicts the reconstruction loss $\epsilon(C)$ for an increasing number of independent components (sources) C. 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 components 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 unmixed sources calculated from the data set are shown in Figure 3.B.2. To see if a single source already shows a clear distinction between the seismic burst and the rest of the data, we marked in blue the samples containing at least one earthquake from the burst. It appears that the ninth unmixed source seems to separate the seismic burst from the rest of the data. This observation raises the question if other trends, such as the background noise, can be correlated with specific unmixed sources.

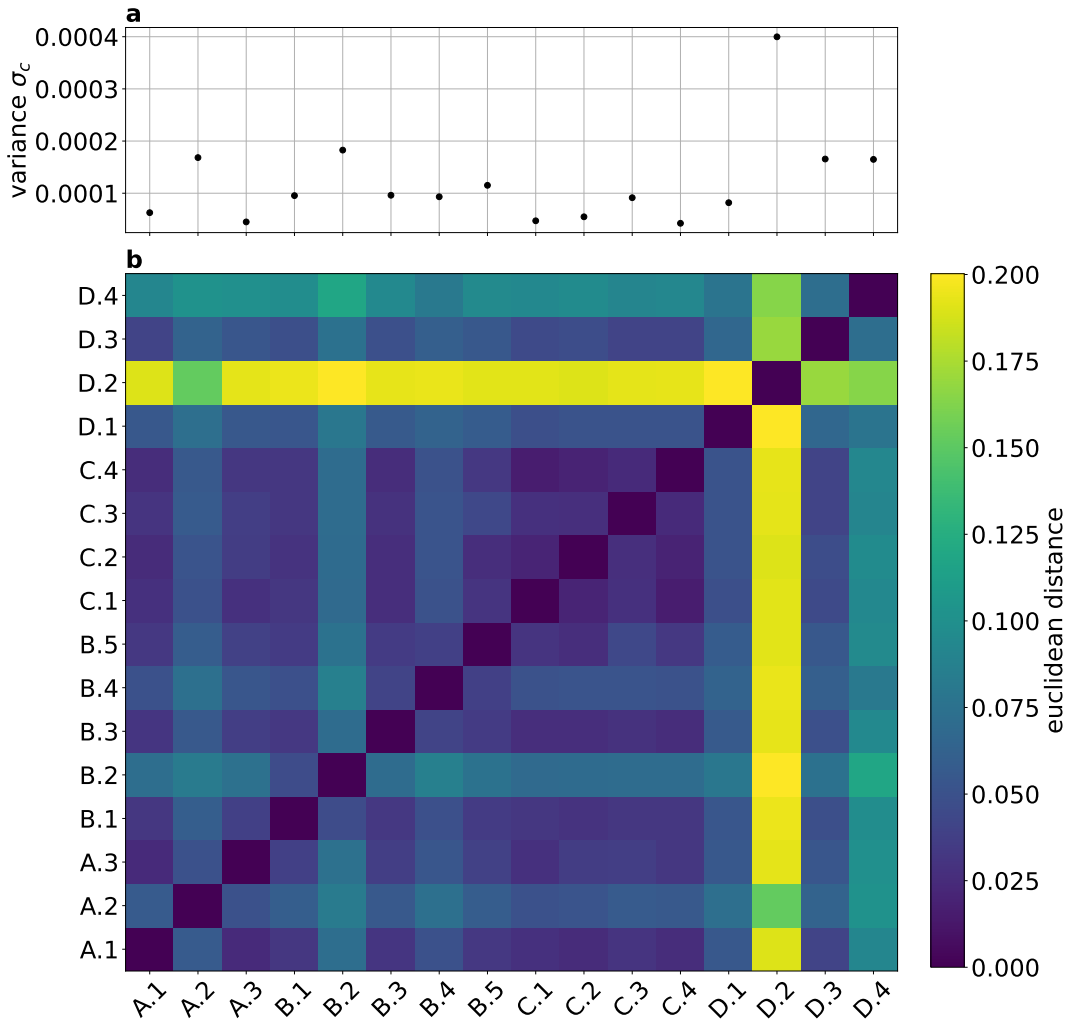

FIGURE 3.A.1: Inter-cluster distances and within-cluster variances. **(a)** Within-cluster variance according to equation 3.4 for all 16 subclusters. **(b)** Inter-cluster distance according to equation 3.3 between all 16 subclusters.

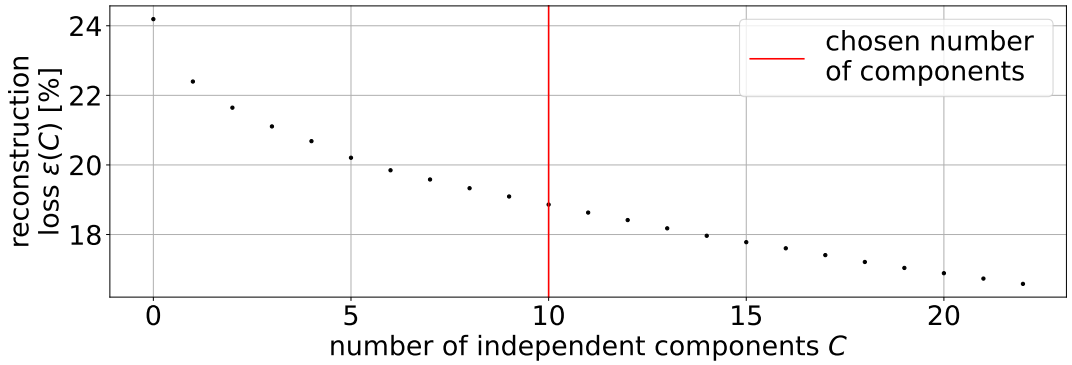

FIGURE 3.B.1: Reconstruction loss with independent component analysis from the deep scattering spectrogram. The reconstruction loss $\epsilon(n)$ is calculated from Equation 3.5 as a function of the number of independent components n.

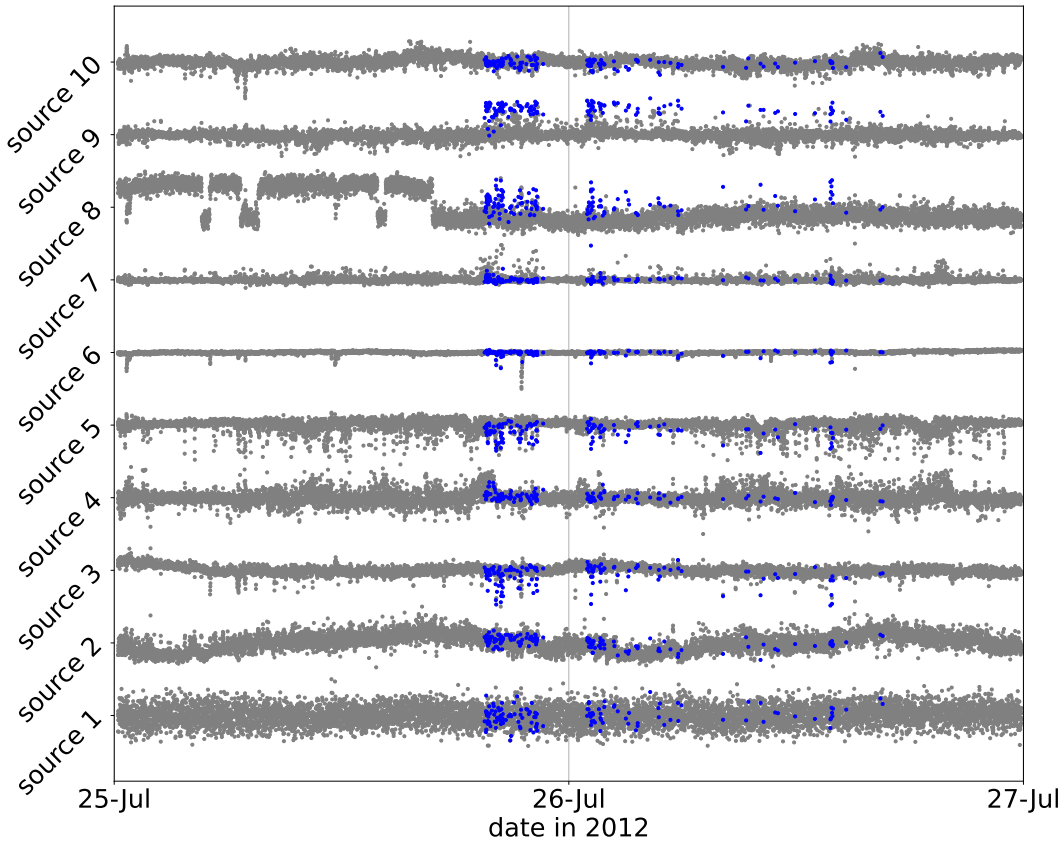

FIGURE 3.B.2: Time series of the ten unmixed sources of the deep scattering spectrogram. The samples containing one or more arrivals of the earthquake from the nearby seismic burst are highlighted with blue dots.

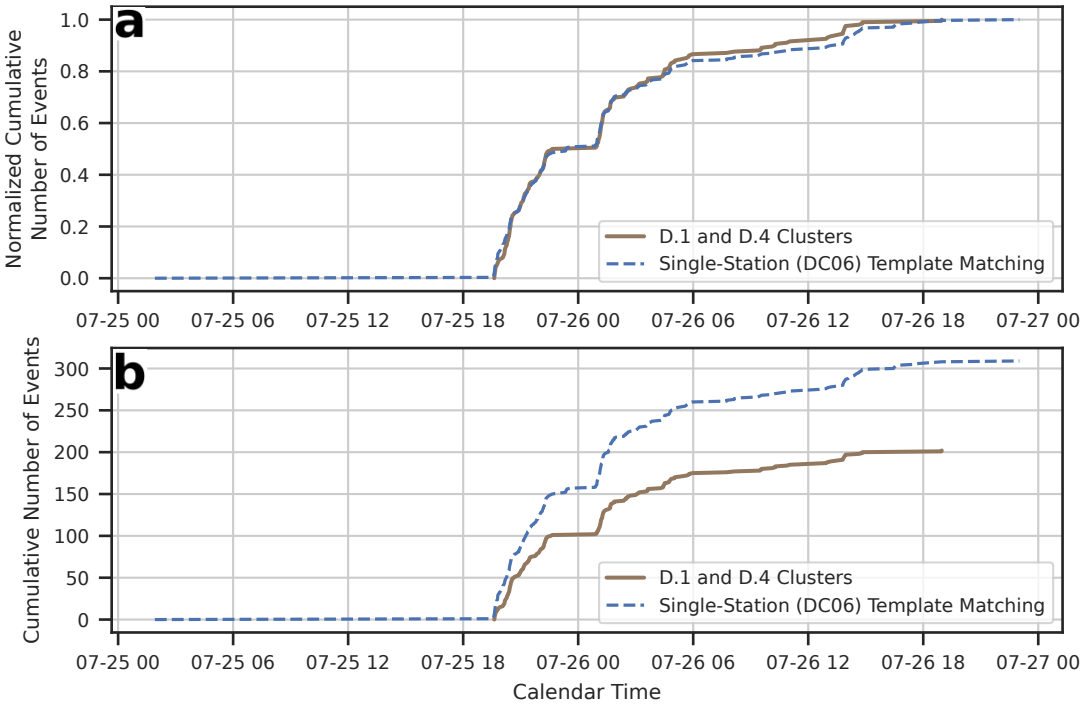

FIGURE 3.C.1: Comparison between the earthquake catalog from clusters D.1 and D.4 (thick brown line), and the single-station (DC06) template matching catalog (dashed blue line). **(a)** Normalized cumulative number of events. **(b)** Cumulative number of events. The single-station template matching catalog documents about 50% more events.

If we compare with the spectrogram of Figure 3.3c we see that the second unmixed source seems to correlate with the variations around 0.2 Hz and the eighth unmixed source seems to correlate with the monochromatic noise source around 1.5 Hz. This quick visual inspection shows us that the feature space can already be physically interpreted, and the ICA separates different signals on its different unmixed sources, which is favorable for further analysis by clustering algorithms.

3.C Comparison with Single-station Template Matching

Station DC06 recorded higher signal-to-noise ratio S-waves from the seismicity burst 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 3.C.1). 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.

3.D Qualitative Comparison with hierarchical clustering based on spectrograms

In our study, we use a deep scattering spectrum instead of a Fourier-transform spectrum, since it is more suitable for classification purposes (Andén and Mallat, 2014). In the following lines, we create and interpret a dendrogram based on Fourier-transform spectral features to verify this claim for seismograms. For the sake of comparison, the window size of the Fourier-transform equals the pooling window of the scattering network, which is 20.48 s. Moreover, the considered frequency range of the Fourier-transform is adapted to the frequency range of the first order scattering coefficients. The three-component spectrogram with 1440 spectral coefficients per time step is then used to calculate ten independent components, which resemble the feature space for the dendrogram. Thus, we only replaced the scattering coefficients with spectral coefficients of comparable time and frequency properties.
To compare the clustering outcome, we retrieve 16 subclusters, which can be grouped into the three main clusters A',B' and C' (see Figure 3.D.1a). The time evolution curves and the cluster sizes in Figure 3.D.1b and c show if the retrieved main clusters are the same as in Figure 3.4. Cluster A' matches very well with cluster A in terms of cluster size and temporal detection curve. Thus, Cluster A' is also related to the monochromatic signal. Cluster B' matches with the detection curve of Cluster C, however, Cluster B' contains more data than Cluster C. Thus, Cluster B' is also related to ambient signals but possibly contains also additional types of signals. The normalized detection curve of Cluster C' matches with Cluster B, however, Cluster C' is not even half of the size of cluster B. Hence, Cluster C' is probably related to high-frequent urban signals. Cluster D, which is related to general seismicity, does not appear within the main clusters based on spectral coefficients. In fact, most of the seismic burst is within cluster B', which is mainly related to ambient signals (see Figure 3.D.1d). Hence, we can assume that Cluster C and D are unified here in Cluster B'. Retrieving subclusters at a lower distance threshold than the three main clusters could possibly reveal a few subclusters related to the seismic burst. However, 12 out of 16 subclusters contain events from the seismic burst (see Figure 3.D.1e). It is not possible to identify a few clusters which are purely related to the seismic burst. Subcluster B'.1 and B'.2 contain ca. 20 % of the cataloged seismic burst respectively, however, most of the subcluster (>96 %) is not related to the cataloged seismic burst. This example shows that a deep scattering spectrum delivers a better representation for classification purposes than a Fourier transform spectrum. This is particularly true for classifying reoccurring transient signals in a relative large data set such as the events of the seismic burst within the continuous seismogram.

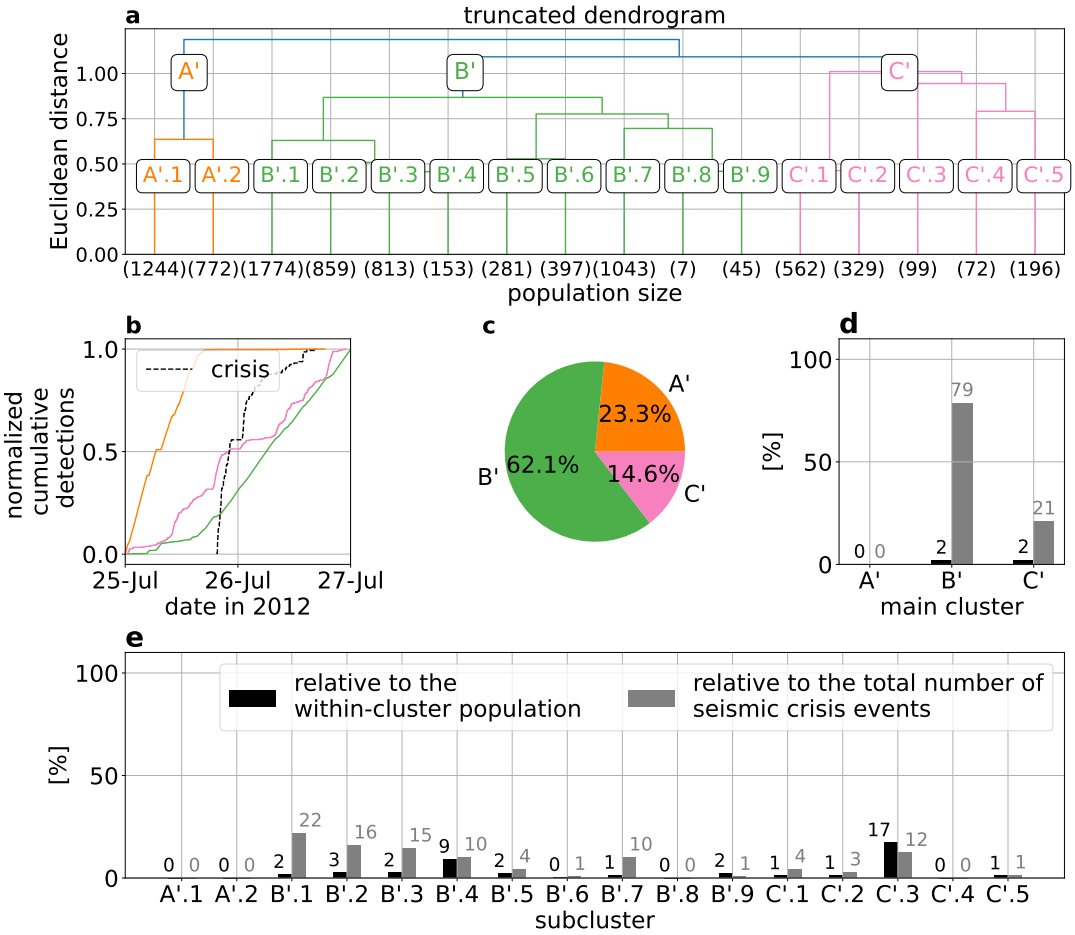

FIGURE 3.D.1: Dendrogram analysis based on spectrogram features and statistical characteristics of the different clusters. **(a)** Dendrogram calculated in the feature space. The dendrogram is here truncated in order to form 16 clusters. The clusters marked with a letter are considered the main clusters, and the subclusters are indicated with numbers. The numbers in the parenthesis indicate the number of samples in each cluster. **(b)** Averaged first-order scattering coefficients of main clusters A, B and C. **(c)** Normalized cumulative detections of main clusters A, B and C, and of the seismic burst obtained from the multi-station template-matching catalog. **(d)** The distribution of the seismic burst across the three main clusters. **(e)** The distribution of the seismic burst all subsclusters.

Chapter 4

AI-based unmixing of medium and source signatures from seismograms: ground freezing patterns

René Steinmann, Léonard Seydoux, Michel Campillo
Article published in Geophysical Research Letters

This chapter covers the second article of my PhD thesis, following the work and ideas presented in Chapter 3. It applies the hierarchical waveform clustering to a dataset recorded in the city of Hamburg, Germany. During the recording time of the dataset, the first centimeters of the surface changes constantly due to freezing and thawing, while many non-stationary seismic signals with anthropogenic origin occur. Out of curiosity, we wondered if we would be able to identify a cluster within the dendrogram corresponding to the superficial freezing process. This was the starting point of this study and it lead us to the conclusion that different processes are encoded onto the different components found by ICA, isolating the continuous freezing and thawing process on a single independent component. This result motivated us to include the feature space more into the data exploration analysis as it is now depicted in Figure 2.1.

4.1 Abstract

Seismograms always result from mixing many sources and medium changes that are complex to disentangle, witnessing many physical phenomena within the Earth. With artificial intelligence (AI), we isolate the signature of surface freezing and thawing in continuous seismograms recorded in a noisy urban environment. We perform a hierarchical clustering of the seismograms and identify a pattern that correlates with ground frost periods. We further investigate the fingerprint of this pattern and use it to track the continuous medium change with high accuracy and resolution in time. Our method isolates the effect of the ground frost and describes how it affects the horizontal wavefield. Our findings show how AI-based strategies can help to identify and understand hidden patterns within seismic data caused either by medium or source changes.

4.2 Introduction

Continuous seismograms are time series of the ground motion recorded at a single location and provide a vast amount of information about processes occurring at the Earth's surface and interior. The recorded ground motion at a given location results from the convolution of the medium's impulse response — expressed as the Green's function — and the seismic waves emitted by various sources, often simultaneously. Thus, continuous seismograms are goldmines to study the medium's properties or sources in time. However, unmixing source or medium changes is often not easy, especially if source and medium changes coincide. For instance, seismic recordings in the vicinity of volcanoes, where many different source and medium effects occur, are challenging and complex datasets to analyze.

To better explore continuous seismic data, seismologists developed many data processing tools to extract valuable information for the task at hand. For example, the Short-Term-Average to Long-Term-Average energy ratio (STA/LTA) scans the continuous recordings for impulsive signals (Allen, 1978). On the other hand, passive image interferometry can interrogate the medium regularly by exploiting the ambient seismic signals of a dataset (Sens-Schönfelder and Wegler, 2006). Undoubtedly, these tools delivered many new insights into the processes happening at and inside the Earth. However, it is important to note that the design of the tools and the related preprocessing favors certain processes in the seismic data. This can be a problem if the source or medium processes encoded in the seismic data are poorly understood. For example, non-volcanic tremors were detected about twenty years ago (Obara, 2002), and still today, the physical mechanism and signal properties of such events are not well apprehended. Therefore, it remains unclear if these signals do not exist in specific environments or if the detection tools are not adapted to the task (Pfohl et al., 2015; Bocchini et al., 2021).

Artificial intelligence (AI) can help overcome those blind spots and discover new signals or hidden patterns within the data. Recently, clustering gained attention as a method to identify families of signals in the continuous seismograms (Köhler, Ohrnberger, and Scherbaum, 2010; Holtzman et al., 2018; Mousavi et al., 2019; Seydoux et al., 2020; Johnson et al., 2020; Snover et al., 2020; Jenkins et al., 2021; Steinmann et al., 2022). In the most common approach, characteristics — often called features — are calculated for a sliding window. Then, clustering algorithms perform a similarity measurement within the set of characteristics and assign a cluster to each window. Until now, the applications showed that this approach mainly identifies families of signals related to source processes such as geothermal activity (Holtzman et al., 2018), different types of anthropogenic activity (Snover et al., 2020), seismic background activity (Johnson et al., 2020) or precursory signals of a landslide (Seydoux et al., 2020). To our knowledge, medium changes have been disregarded so far in this task.

In the present study, we make the first attempts towards inferring not only source processes but also medium changes from continuous single station seismograms in a data-driven fashion.

4.3 A thin ground frost layer visible in temperature data and seismic velocity variations

The study site is located in the city of Hamburg, Germany (Figure 4.1a). Besides the three broadband sensors WM01, WM02, and WM03, the site includes various

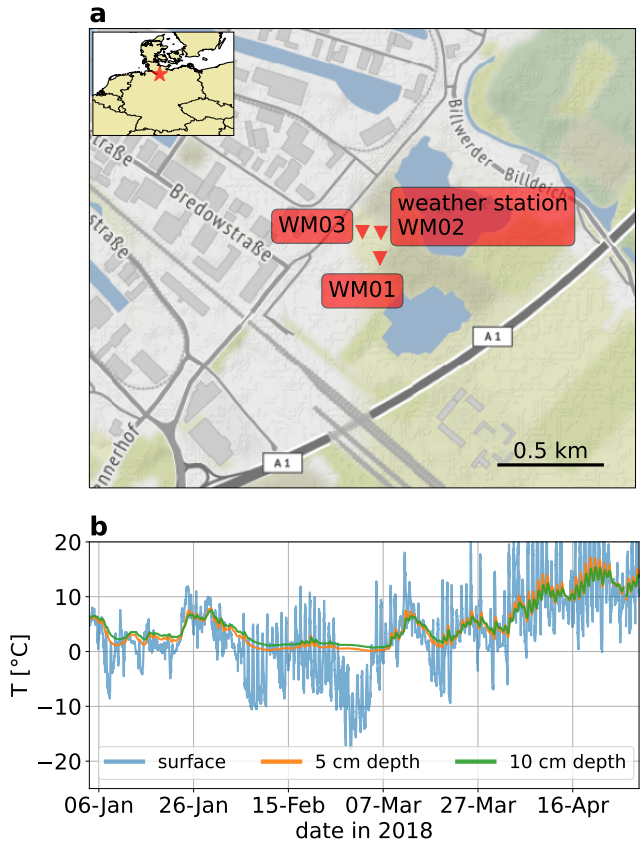

FIGURE 4.1: **Temperature data and location of seismic stations.** (**a**) Map of the measuring site in Hamburg, Germany, with the three broadband and three-component seismic sensors WM01, WM02, and WM03. (**b**) Temperature time series measured at the surface, 5 cm and 10 cm depth close to station WM02 with a sampling period of 10 min.

meteorological sensors near station WM02. At 5 cm, 10 cm, 80 cm, and 120 cm depth and at the surface, temperature sensors deliver a measurement every 10 min. Figure 4.1b depicts the temperature time series at the surface, 5 cm, and 10 cm depth from January 4 to April 30 in 2018. Until the end of March, the air temperature ranges between −20 °C and 20 °C indicating a continuous freezing and thawing of the near-surface. In particular, the end of February is a cold period with freezing air temperature during daytime and nighttime. However, at 5 cm and 10 cm depth, the sensors do not reach below 0 °C and do not follow the air temperature as they do later in March. This is known as the zero-curtain effect: the phase change from water to ice in the soil releases latent heat, which causes the freezing process to slow down (Outcalt, Nelson, and Hinkel, 1990). This implies that the ground frost is not deeper than 5 cm during the coldest period.

The freezing and thawing process on a centimeter scale was well tracked with seismic velocity variations retrieved from passive image interferometry applied to the data from the three broadband stations WM01, WM02 and WM03 (Steinmann, Hadziioannou, and Larose, 2021). Freezing periods caused a velocity increase and thawing periods caused a velocity decrease. The local seismic wavefield comprises many non-stationary seismic sources related to the anthropogenic activity, such as commuter and freight trains in the south, a highway passing in the southeast (labeled A1 on Figure 4.1a), a close gravel pit (marked by the two nearby lakes on

Figure 4.1a) and an industrial neighborhood in the northwest. The combination of the continuously changing medium due to the freezing and thawing and many non-stationary seismic sources makes it an interesting study case for our approach to disentangle the medium from the source effects blindly.

4.4 Seismic pattern detection with hierarchical waveform clustering

We search for the imprint of the ground frost within the continuous three-component seismograms recorded by a single station with the hierarchical waveform clustering approach introduced in Steinmann et al. (2022). Hierarchical clustering observes how a dataset merges into clusters based on some similarity criterion (Estivill-Castro, 2002). In our case, we calculate the similarity between waveforms from a set of features derived from a deep scattering spectrogram, as depicted in Figure 4.2. Firstly, we calculate the deep scattering spectrogram of the continuous three-component seismograms with a deep scattering network, as introduced in Andén and Mallat (2014) and adapted to seismology in Seydoux et al. (2020). A deep scattering network is a deep convolutional neural network, where the convolutional filters are restricted to wavelets and the activations to modulus operation. We choose Gabor wavelets as originally proposed in Andén and Mallat (2014) and do not learn the wavelets as the authors did in Seydoux et al. (2020). The output of such a network at each layer allows building the deep scattering spectrogram representation of a continuous multichannel seismogram. This representation of time series is relevant for classification purposes since it preserves signal phenomena such as attack and amplitude modulation. Moreover, a deep scattering spectrogram is locally translation invariant and stable towards small-amplitude time warping deformations (Andén and Mallat, 2014). Indeed, Steinmann et al. (2022) showed that hierarchical waveform clustering performs poorer if the deep scattering spectrogram is replaced by a Fourier-based spectrogram. We depict a two-layer scattering network in Figure 4.2, where we apply a sliding window on a single-component seismogram and calculate the first-order scalogram with the wavelet transform. A second wavelet transform is applied to the first-order scalogram creating the second-order scalogram. A pooling operation collapses the time axis of the scalograms and recovers the first- and second-order scattering coefficients. For each component of the ground motion record, we calculate the scattering coefficients and concatenate them. We repeat this for each window and retrieve the deep scattering spectrogram. The design of the scattering network (number of wavelets, type of pooling, etc.) can be adapted to the task at hand and is explained more in detail in Appendix 4.B of this chapter.

Deep scattering spectrograms are redundant and high-dimensional representations, not directly suited for clustering due to the curse of dimensionality (Bellman, 1966). Therefore, we extract the most relevant characteristics — or features — and reduce the number of dimensions with an ICA, a linear operator for feature extraction, and blind source separation (Comon, 1994). Before applying the ICA, we whiten the deep scattering spectrogram by equalizing its covariance matrix eigenvalues, allowing us to disregard patterns' relative amplitudes as much as possible. The number of most relevant features (or independent components) is often unknown and should be inferred, which is explained more in detail in Appendix 4.C of this chapter.

Lastly, we perform hierarchical clustering in the low-dimensional feature space built by the independent components. Clustering aims at grouping objects — here

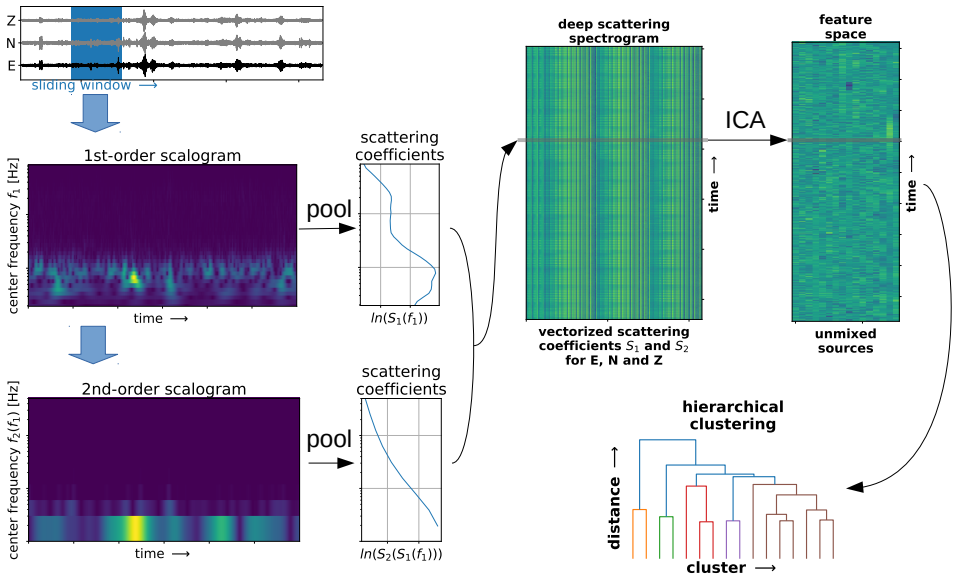

FIGURE 4.2: **Sketch of the hierarchical waveform clustering approach.** A two-layer scattering network with wavelet transforms, modulus and pooling operations calculates the deep scattering spectrogram. An independent component analysis (ICA) extracts the most relevant features, which are used for hierarchical clustering.

defined as data points in a given feature space — based on a similarity or dissimilarity measurement. With a bottom-up approach of hierarchical clustering, also called agglomerative clustering, all objects start in a singleton cluster and merge to larger clusters until all objects unify in a single cluster (Johnson, 1967). A dendrogram depicts this process, representing the inter-cluster similarity in a cluster-distance diagram. The similarity measurement, which drives the cluster merging, is often a distance in the feature space between the objects. Thus, the type of distance is the only choice to be made here and determines the structure of the dendrogram. We use Ward's method as a criterion to merge clusters in hierarchical clustering and produce the dendrogram. Clusters are merged with the objective to keep the increase of the total within-cluster variance minimal (Ward Jr, 1963). This allows to find cluster of various size, which fits the nature of seismic data, where ambient seismic activity often outweighs transient signals. Finally, depending on the truncation distance explored in the dendrogram, one can obtain a different number of clusters. This allows exploring the dataset's structure and searching for a cluster of seismic signals related to the ground frost. The dendrogram is unique to hierarchical clustering and the main reason why we choose this clustering algorithm instead of others.

4.5 Cluster of signals occurs during ground frost

We show a truncated dendrogram of the continuous three-component seismogram recorded at station WM01 from January to April 2018 in Figure 4.3a, using a truncation distance to end up with 16 clusters in this case. A data point in the feature space represents 10 min of continuous waveform data without overlap. Moreover, the feature space contains 16 independent components, as a trade-off between keeping enough information and low dimensionality (see Appendix 4.C of this chapter and Figure 4.C.1). Note that finding a cluster related to ground frost effects is an

exploratory task where we do not know where such a cluster would appear in the dendrogram nor if it even exists. As suggested in Steinmann et al. (2022), we extract a few large clusters at a high distance threshold to overview the whole dataset. We can then focus on certain branches in the dendrogram and extract subclusters hierarchically to get a more detailed cluster analysis if needed. In our case, we extract five clusters (hereafter denoted A, B, C, D, and E) at a distance threshold of 0.9 (Figure 4.3a). In the following lines, we will interpret the clusters and assign meaningful labels with certain inherent clusters properties such as the normalized cumulative detections in time (Figure 4.3b–f), the number of detections per hour during the day (Figure 4.3g–k), the number of detections per weekday (Figure 4.3l–p), and the first-order scattering coefficients averaged for each input channel (Figure 4.3q–u). In particular, the normalized cumulative detections in time can help identify a cluster related to the presence of ground frost since the temperature time series indicate the periods of freezing air temperature. Note that a detection refers to a 10 min window of seismic data which is assigned to one of the five clusters.

Cluster A seems to detect in a linear-piecewise way, with no relation to the temperature time series or occurrence of ground frost (Figure 4.3b). This cluster detects only between 05:00 and 18:00 local time from Monday to Friday (Figure 4.3g and i). Note that around 09:00 and 12:00, the detections reach a minimum, coinciding with the typical breakfast and lunch break during workdays. Compared to the other clusters, the averaged first-order scattering coefficients show larger values for frequencies above 1 Hz with a local maximum around 8 Hz on the vertical component (Figure 4.3q). The analysis of these parameters indicates that this cluster contains seismic signals related to anthropogenic sources, mainly active during classical labor hours. The gravel pit with trucks in the direct neighborhood of this measuring site could be a possible source (Figure 4.1a).

Cluster B seems to detect more continuously than cluster A (Figure 4.3c). It is active during the daytime, with a few detections during the nighttime (Figure 4.3h). Interestingly, this cluster peaks at 09:00 and 12:00 when cluster A reaches a minimum of detections. The weekdays show clearly more detections than the weekends, with a peak of detection on Fridays when cluster A shows a minimum of detection during the week (Figure 4.3l and m). The averaged first-order scattering coefficients show similar frequency characteristics as cluster A. However, cluster B indicates no bumps around 8 Hz (Figure 4.3r). The analysis of cluster B suggests that this cluster also relates to anthropogenic activity. Since it shows elevated activity when cluster A reduces its activity (Fridays and 09:00 and 12:00 local time), it is probably related to a different anthropogenic seismic source. Because cluster B also contains some detections during the nighttime and weekends, it possibly contains seismic signals related to nearby road traffic.

Cluster C is the second-largest cluster of the whole dataset (Figure 4.3a). It detects irregularly at all hours and all days (Figure 4.3d, i and n). During the morning and afternoon its detection rate decreases (Figure 4.3i). Moreover, the averaged first-order scattering coefficients show no particular pattern (Figure 4.3s). It is unclear what type of seismic signals cluster C contains. We can only note that it is not related to ground frost since its detections rate does not correlate with freezing temperatures.

Cluster D activates mainly during two periods (Figure 4.3e). At the beginning of February, it accumulates 25 % of its size followed by a slight pause. Then, at the end of February and beginning of March it detects the remaining 75 % of its total size. The detection periods occur during the coldest temperatures recorded at 5 cm depth. Therefore, cluster D most likely groups seismic signals related to ground

frost. Cluster D detects during all hours and all days. However, slightly more detections appear during the weekend and nighttime (Figure 4.3j and o). There are probably two effects that explain this behavior. Firstly, due to colder temperatures, ground frost occurs predominantly at night and so do the associated seismic signals (Figure 4.1b). Secondly, due to anthropogenic activity, the seismic wavefield in an urban environment changes significantly between day and night and weekdays and weekends. Thus, the changing wavefield modulates the signature of the ground frost recorded by continuous seismograms. For instance, a seismogram containing seismic signals generated by road traffic during ground frost could be found in cluster B or D. Indeed, inside cluster B, we can identify subcluster B.1 as anthropogenic seismic signals effected by the ground frost (see Figure 4.3a and Figure 4.C.2). This points out a limitation of clustering: a seismogram containing multiple types of signals is assigned to a single cluster, which oversimplifies the nature of the data and has been already noted by Steinmann et al. (2022). The averaged first-order scattering coefficients show no clear and distinct pattern (Figure 4.3t). Cluster D seems different from Cluster A and B due to lower scattering coefficients for higher frequencies. However, it is unclear how cluster D differs from clusters C and E. We can note that the averaged first-order scattering coefficients do not deliver a unique signature related to these signals.

Cluster E is the largest cluster of the whole dataset (Figure 4.3a). It detects continuously with a decreased detection rate during February when ground frost occurs, with more detections during night and weekends (Figure 4.3f, k, and p). Moreover, the cluster shows lower averaged first-order scattering coefficients at higher frequencies (Figure 4.3u), distinguishing them from clusters A and B but D. The analysis of cluster E indicates that it groups ambient seismic noise without particular transients and ground frost. In fact, it appears that cluster D and E summarize the stationary ambient wave field separated only due to the occurrence of ground frost. Indeed, the combined clusters seems to detect almost continuously during weekends and nights (see Figure 4.C.2).

Summarized, the dendrogram delivers a data-driven overview about the content of the data containing both source and medium effects. We can clearly identify cluster A and B with anthropogenic seismic sources. Inside cluster B we identified a small subcluster containing anthropogenic signals effected by the ground frost. We have reasons to assume that a more detailed cluster solution would reveal a similar subcluster in A. We can not find a meaningful label for cluster C. The largest part of the data is located within cluster E: ambient seismic noise, which is not effected by ground frost. Cluster D seems to be the only cluster related to the freezing of the surface without particular transient signals from anthropogenic activity. The hierarchical clustering approach, together with an interpretation of a cluster solution at a high distance threshold, allowed us to give a detailed analysis of the content of the seismic data. In particular, the cumulative detection curve identifies cluster D as of interest in our study because it relates purely to ground frost. Hence, we do not need to extract a more detailed cluster solution. In the following lines, we analyze how the freezing and thawing process is encoded in the data.

4.6 Disentangling the ground-frost from the urban imprint

Hierarchical clustering built the dendrogram within the feature space extracted by an ICA from the deep scattering spectrogram (Figure 4.2). The features likely reveal insights about the signature of cluster D and, thus, about the ground frost signature.

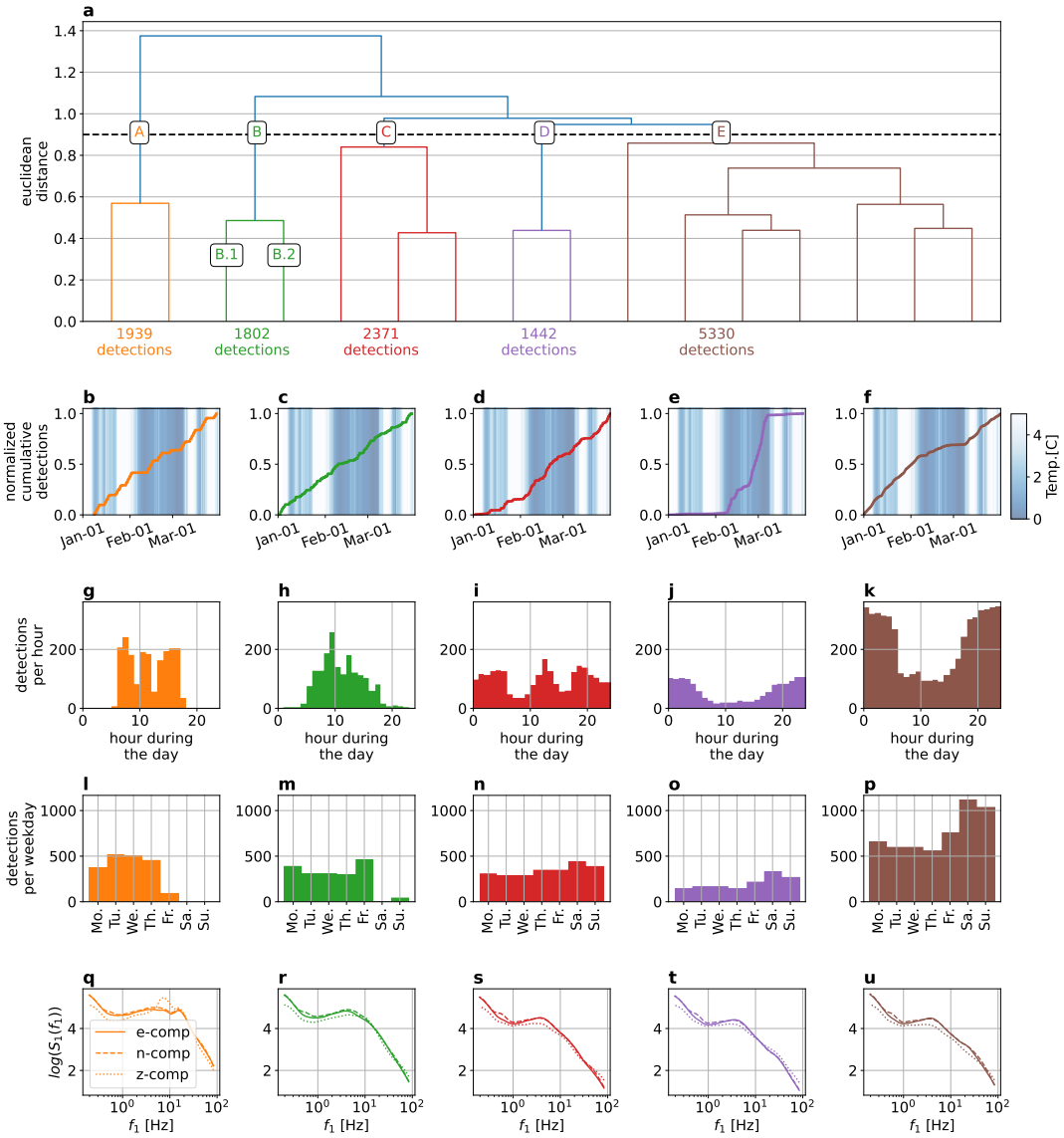

FIGURE 4.3: **Results of seismic data clustering from the three-component broadband station WM01 between 1 January to 1 April 2018.** (**a**) dendrogram with a truncation distance set to obtain 16 clusters. (**b–f**) normalized cumulative detection. (**g–k**) daily occurrence. (**l–p**) weekly occurrence. (**q–u**) averaged first-order scattering coefficients.

Steinmann et al. (2022) already showed that single features retrieved from the scattering coefficients with an ICA could reveal interesting patterns in the seismogram. Therefore, we can likely identify a single feature in our dataset that encodes the seismic signature of the ground frost. The geometric center of a cluster in the feature space, also called centroid, can tell us if one feature is more important than other features. In our case, we define the geometric center of a cluster as the mean of its data points in the 16-dimensional feature space. We note that if all features are equally important in defining a cluster, they should contribute equally to the centroid coordinates. If a few or single features are more important than others, the centroid should have a stronger contribution from them. We calculate the centroid of cluster D and take the modulus, since we are only interested in the amplitude information (Figure 4.4a). We observe that the centroid of cluster D shows a substantial value for feature 15 (Figure 4.4a) regarding the other features. This suggests that cluster D is active when large absolute values on feature 15 occur.

We can also observe how feature 15 evolves in time (Figure 4.4b). Feature 15 shows a significant amplitude decrease at the end of February and the beginning of March. During that time, it seems to mimic the low-frequent trend of the air temperature with a slight offset in time. The beginning of February and mid-March show smaller amplitude decreases after a few consecutive nights of freezing air temperature. Unfortunately, we have no ground truth about the occurrence of ground frost. However, we know that the occurrence of ground frost depends on the amount of time and the amplitude of freezing air temperature. Moreover, thawing air temperatures during the day counteract the nightly built-up of ground frost. A more extended and continuous period of freezing air temperature (like the one at the end of February) results in a thicker layer of ground frost. A colder air temperature can also decrease the temperature inside the layer of ground frost and, thus, increase its stiffness and shear wave velocity (Zimmerman and King, 1986; Miao et al., 2019). These facts, combined with the observation of feature 15 and the air temperature, suggest that this feature tracks the freezing and thawing process of the surface at a high-resolution timescale of 10 min. We emphasize that feature 15 is an entirely data-driven product from a three-component seismogram with minimal processing. In comparison, Steinmann, Hadziioannou, and Larose (2021) tracked the same freezing and thawing process with data from two seismic stations, heavier preprocessing, and a time resolution of 2 days.

Since ICA is a linear operator, we can use only feature 15 to reconstruct the scattering coefficients out of the mixing matrix, defined as the pseudo-inverse of the unmixing matrix (Comon, 1994). This procedure acts as a filter process since we zero all features except feature 15. Due to the large size of first- and second-order scattering coefficients, Figure 4.4c–h show only the first-order original and reconstructed scattering coefficients for all three components. The original coefficients show clearly the urban imprint in the seismic data: fringes appear during daytime and pause at the weekends (Figure 4.4c, e and g). No clear pattern appears during ground frost building periods, such as at the end of February (Figure 4.4b). The reconstructed coefficients do not contain the fringes due to urban activity since these signals were probably encoded in one of the muted features (Figure 4.4d, f and h). The filtering effect reveals a slight amplitude decrease for the horizontal components at frequencies above 1 Hz during the end of February, coinciding with the coldest period of the dataset. During that time, a faint amplitude decrease can also be observed at the vertical component. At times with consecutive cold nights such as at the beginning of February or mid-March, these decreases are also faintly visible. These observations confirm that the wavefield experiences an energy decrease during ground frost

with a discrepancy between horizontal and vertical components. Indeed, the ratio of horizontal and vertical scattering coefficients show a clear broadband high-frequent decrease at the beginning and end of February for both original and reconstructed data (Figure 4.4i and j). It appears that the broadband decrease in the ratio becomes stronger with increasing time or amplitude of the freezing air temperature. The ratio of horizontal and vertical scattering coefficients resembles the classical Horizontal-to-Vertical-Spectral-Ratio (HVSR) based on the Fourier transform. The question rises if the observed change in the seismic data is due to a changing medium caused by freezing and thawing or due to changes in the seismic sources. First of all, we could argue that a source change would probably effect all three components similarly, which is not our case. Moreover, if a temperature related source would appear, it would probably increase the energy during times of freezing, which also does not fit our observations. In fact, it was shown before that ground frost can cause a broadband decrease in the HVSR for higher frequencies (Guéguen et al., 2017). Our observations suggest that less than 5 cm of ground frost has already an impact on the seismic wavefield. Indeed, models based on the diffusive field assumption (Sánchez-Sesma et al., 2011; Piña-Flores et al., 2016; García-Jerez et al., 2016) confirm an HVSR decrease due to a thin layer of ground frost (see Appendix 4.D and 4.E of this chapter, and Figure 4.D.1 and 4.E.1) in the supplementary materials). All these arguments suggest strongly that the revealed signature is indeed due to a medium change.

4.7 Conclusion

In this study, we made the first attempts towards inferring blindly medium changes from the wavefield recorded by a single station. For our case study, the medium continuously changes due to surface freezing and thawing, while anthropogenic activity creates a complex and non-stationary seismic wavefield. An AI-based approach, based on the deep scattering network, an ICA and hierarchical clustering, helped us explore the seismic data and search for possible patterns induced by the ground frost without assuming how the seismic data could be affected. One of the main outcomes of this study is that the AI-based approach blindly extracts a feature that isolates the seismic response due to the medium change and mutes other non-stationary processes. This opens new possibilities to utilize single station data for monitoring purposes, especially in environments with many source and medium processes such as permafrost (e.g. Köhler and Weidle, 2019) or volcanoes. AI-based strategies could complement other passive seismic methods used for permafrost monitoring (e.g. James et al., 2019; Lindner, Wassermann, and Igel, 2021; Cheng et al., 2022). This could give new insight into the response of permafrost to climate change given the decade-long availability of single seismic stations near permafrost areas. Future research could also investigate if other types of medium changes (e.g., groundwater fluctuations) could be directly extracted from the seismograms in a data-driven fashion.

Moreover, the revealed signature combined with the HVSR model indicates that superficial freezing might impact the modal energy distribution. This effect has been observed for other high-velocity surface layers at engineering sites (O'Neill and Matsuoka, 2005). However, to our knowledge, it has not yet been considered in permafrost studies using passive seismic methods. On the one hand, it could corrupt velocity variation measurements retrieved from surface waves in cross-correlograms. On the other hand, it would also be an opportunity since more modes increase the

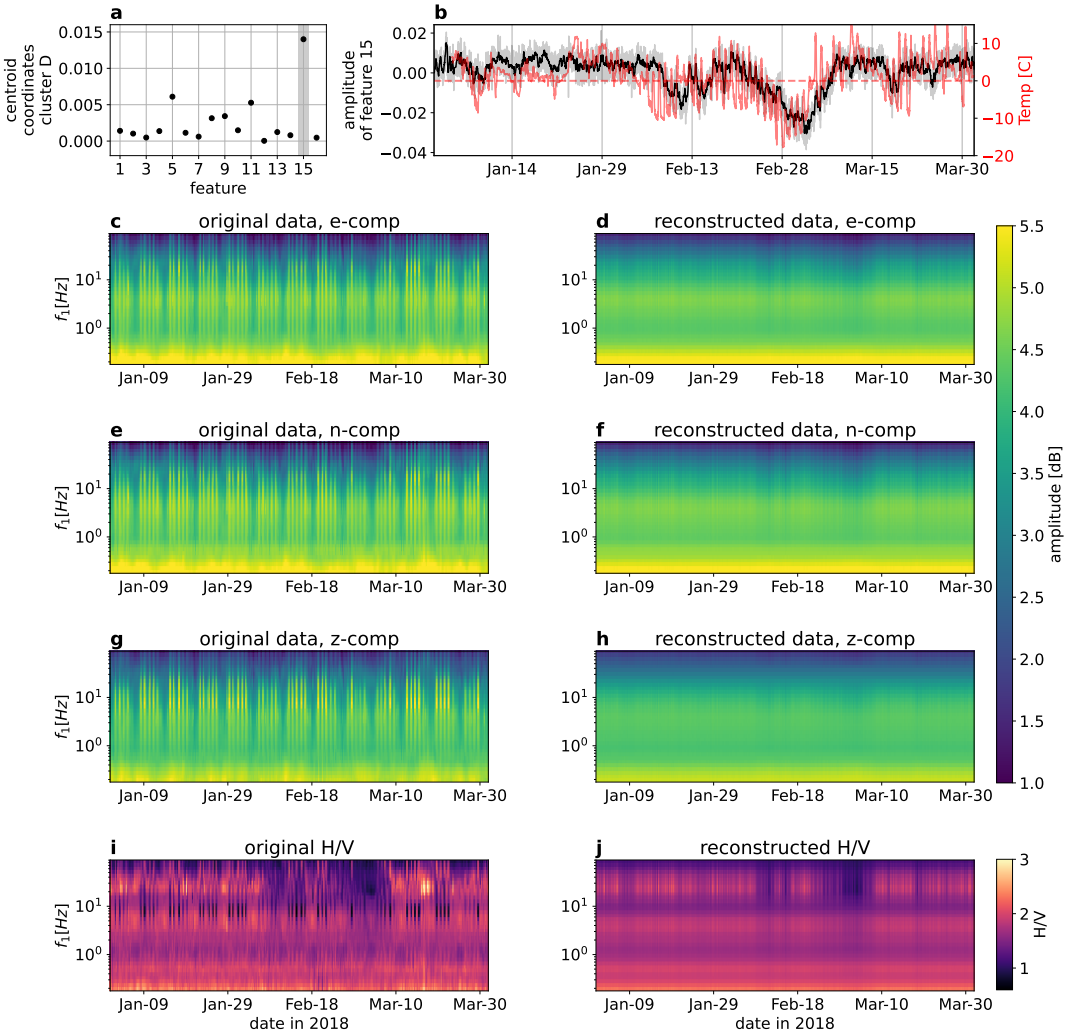

FIGURE 4.4: **The signature of freezing** (**a**) coordinates of the centroid of cluster D in the eight-dimensional feature space. (**b**) feature 15 as a smoothed time-series (black) compared to the temperature time-series recorded above ground (red). The orignal feature without smoothing is represented in grey. (**c,e,g**) Original first-order scattering coefficients for the east, north and vertical component, respectively. (**d,f,h**) Reconstructed first-order scattering coefficients based solely on feature 15 for the east, north and vertical component, respectively. (**i**) Ratio between horizontal and vertical components based on the original first order scattering coefficients. (**j**) Ratio between horizontal and vertical components based on the reconstructed first order scattering coefficients.

amount of information about the subsurface. Future research is needed to understand better the interaction between different surface wave modes in the presence of frozen surface layers.

4.8 Open Research

The seismic data was downloaded from Steinmannn, Hadziioannou, and Larose (2020) and the temperature data were provided by the Meteorological Institute of Hamburg. The temperature data can be retrieved by contacting the Meteorological Institute of Hamburg through wettermast. The main code for calculating the scattering coefficients, features and linkage matrix can be found under zenodo. The work relies heavily on the python packages ObsPy (Beyreuther et al., 2010), scikit-learn (Pedregosa et al., 2011a) and SciPy (Virtanen et al., 2020). The map was produced with map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

4.9 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). We want to thank the Meteorological Institute and the Institute of Soil Science at the University of Hamburg for providing the temperature data set, which we used in this publication. The seismic data was made available through the German Research Foundation (DFG) under Germany's Excellence Strategy—EXC 2037 'CLICCS - Climate, Climatic Change, and Society' - project number: 390683824, and the Cluster of Excellence 'CliSAP' (EXC177), contribution to the Center for Earth System Research and Sustainability (CEN) of Universität Hamburg.

Appendices

4.A Introduction

The seismic data is sampled with 200 Hz. Because the data was retrieved manually from the field, three data gaps of ca. 3 h occur in the dataset. Before applying the hierarchical waveform clustering, the data was demeaned and high-pass filtered with a corner frequency of 0.1 Hz. The data gaps were filled with zeroes. However, the scattering coefficients of the data gaps were removed before the feature selection. The supporting information provides details about:

- the design of the deep scattering network
- the number of releveant features retrieved with an ICA
- the cumulative detections for subcluster B.1, B.2 and the combination of cluster D and E
- the HVSR models with and without a thin layer of ground frost

4.B Design of deep scattering network

We design a deep scattering network with 36 complex-valued Gabor wavelets in the first layer and 9 Gabor wavelets in the second layer. A modulus operation retrieves real-valued scalograms. The first layer creates 36 scattering coefficients and the second layer creates 324 (as from 36×9) scattering coefficients per sliding window and component. The center frequencies of the first-layer wavelets range from 0.2 to 89 Hz and the center frequencies of the second layer wavelets range from 0.2 to 50 Hz. The number of wavelets was chosen specifically to cover a wide range of frequencies above the oceanic microseism. The upper frequency of the first layer is bounded by the sampling frequency of 200 Hz. The center frequencies are spaced logarithmically with four wavelets per octave in the first layer and one wavelet per octave in the second layer. The sliding window is set to 10 min to mimic the time resolution of the temperature data. In contrast to Steinmann et al. (2022), we apply average pooling instead of maximum pooling to the first and second layer scalograms since we are not searching for transient signals but changes in the ambient seismic wavefield.

4.C Extracting the most relevant features

After calculating the deep scattering spectrogram, we apply an ICA to retrieve the most relevant features. The ICA model can be written as:

$$\mathbf{x} = \mathbf{s}\mathbf{A}, \tag{4.1}$$

where $\mathbf{x} \in \mathbb{R}^{N \times F}$ are the N observations of dimension F, $\mathbf{A} \in \mathbb{R}^{F \times C}$ is the mixing matrix, and $\mathbf{s} \in \mathbb{R}^{C \times N}$ are the C independent components, representing the set of

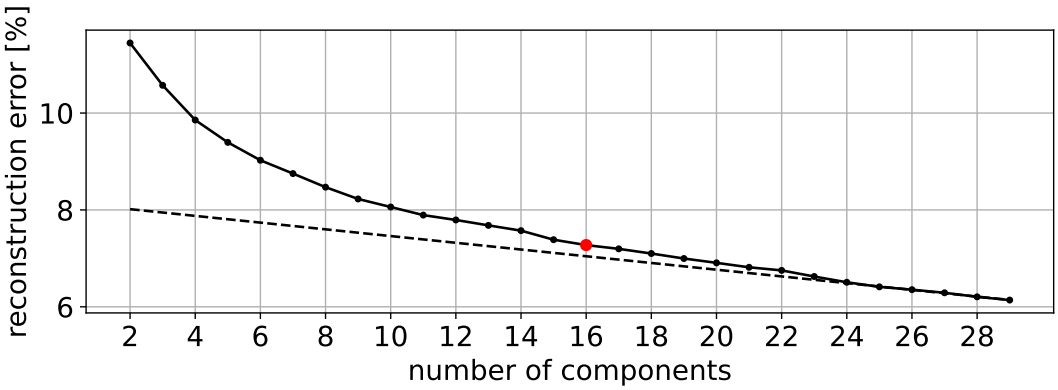

FIGURE 4.C.1: Reconstruction error for ICA-models with different number of independent components. The red dot marks the model we choose for further analysis. The dashed line fits a linear function based on the last seven points.

features for hierarchical clustering. Equation 4.1 considers the observations $\mathbf{x}$ as a linear combination of the independent components $\mathbf{s}$, with the mixing weights gathered in $\mathbf{A}$. In our case, $\mathbf{x}$ are the whitened scattering coefficients. Setting the number of independent components is an exploratory task that can be seen as a trade-off between keeping the dimensionality low for clustering and retaining the most crucial data information. We use the reconstruction loss $\epsilon(C)$ between the original data $\mathbf{x}$ and the reconstructed data $\hat{\mathbf{x}}^{(C)}$, based on the C independent components, as a guideline for choosing an optimal number for C. The reconstruction loss is defined as following:

$$\epsilon(C) = \frac{\sum_{i=0}^{N} |x_i - \hat{x}_i^{(C)}|}{N}. \tag{4.2}$$

Figure 4.C.1 depicts the reconstruction loss $\epsilon(C)$ for an increasing number of independent components C. The reconstruction loss decreases rapidly with the first 14 components. With more than 14 components, the rate of error decrease becomes smaller and almost linear. However, a small jump occurs from 14 to 16 components. Therefore, 16 independent components, marking a kink in the reconstruction error curve, seem like a good choice to us and are the basis for building the linkage matrix for the dendrogram.

4.D Inverting for a 1D velocity model

To forward model the effect of ground frost on the HVSR, we need a 1D velocity model with the shear wave velocity v_s, the compressional wave velocity v_p, the thickness of the layer h and the density ρ. Steinmann, Hadziioannou, and Larose (2021) provides a 1D velocity model to a depth of less than 30 m based on a shear wave refraction profile. The forward modelled HVSR based on this velocity model together with the observed HVSR at the three stations at 15 April 2018 are shown in Figure 4.D.1. We chose this day for an HVSR measurement for two reasons. Firstly, the time of the year and the temperature data suggest that we do not have any ground frost (Figure 1a). Secondly, it is a Sunday and, thus, we have better conditions for an equipartitionned wavefield without anthropogenic activity (Figure 3).

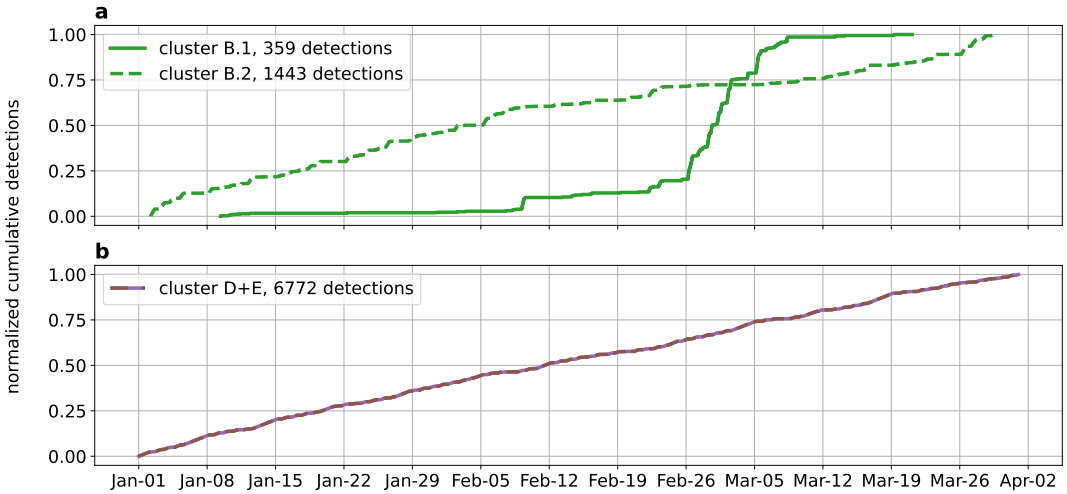

FIGURE 4.C.2: **Normalized cumulative detections for other cluster solutions.** Normalized cumulative detections for subcluster B.1 and B.2 **(a)** and the cluster-combination of D and E **(b)**. Note that each tick at the x-axis marks a Monday.

h [m]	v_s [m/s]	v_p [m/s]	ρ [g/cm^3]
172.82	394.54	1255.93	2000
611.60	520.96	2075.66	2000
∞	947.09	4250.25	2000

TABLE 4.D.1: 1D model of the subsurface at the measuring site based on the inversion of the HVSR with the diffusive field assumption

It is clear that the modelled HVSR does not fit the observations. Since the two resonance peaks below 1 Hz do not occur in the modelled HVSR, it appears that the velocity model is not deep enough. To update the velocity model, we invert the HVSR measurements based on the diffusive field assumption (Piña-Flores et al., 2016). We invert for a three-layer model with the observed HVSR between 0.1 and 1 Hz to fit the two resonance peaks. The higher frequency content seems unreliable, since the variations between the stations are too large given the fact that they are only 100 m apart (see map in Figure 1a). These variations at higher frequencies can be the result of different installation types. WM01 and WM02 are placed on a concrete slab while WM03 is inside a shed. We constrain the range of possible shear wave velocity of the first layer with the values given in Steinmann, Hadziioannou, and Larose (2021). The updated and deeper velocity model fits better the observations and, thus, is utilized for modelling the effect of the ground frost. The values of the updated model are presented in Table 4.D.1.

4.E Modelling the effect of a frozen surface on the HVSR

We model the effect of ground frost on the HVSR based on a 1D velocity model and diffuse wavefield assumption (Sánchez-Sesma et al., 2011; García-Jerez et al., 2016). Firstly, we derive a 1D velocity model from the inversion of H/V measurements (Piña-Flores et al., 2016) and constraints from a shear wave refraction profile (Steinmann, Hadziioannou, and Larose, 2021). To evaluate the effect of ground frost, we insert a centimeter thick high-velocity layer at the surface of the 1D model. Different

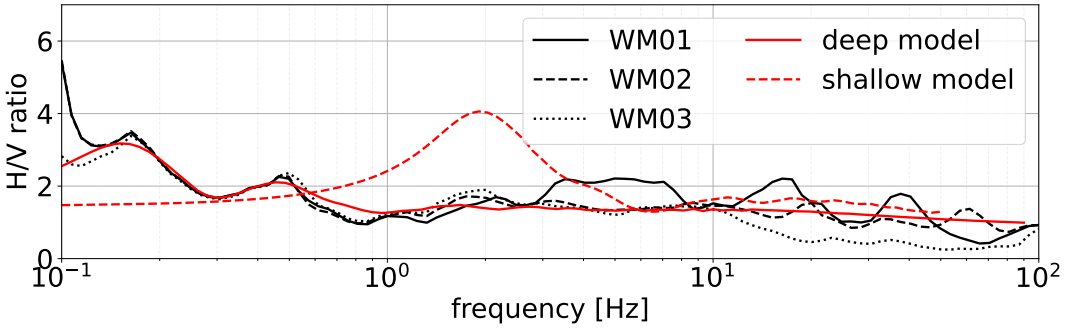

FIGURE 4.D.1: The observed HVSR at all three stations, the modelled HVSR based on the velocity model given in Steinmann, Hadziioannou, and Larose (2021) as the dashed red line and the modelled HVSR based on the inversion of the HVSR as the red solid line.

thicknesses and shear wave velocities account for different scenarios of the ground frost. The shear wave velocity of the ground frost depends strongly on the temperature and composition of the soil. A silt-clay mixture with a high water content as in our case can reach the eight-fold of its shear wave velocity with temperatures below $-8\,°C$ (Miao et al., 2019). Through the shear wave velocity and a constant Poisson's ratio of 0.33 (Zimmerman and King, 1986), we define the compressional wave velocity. We neglect changes in the density and set it to $2000\,kg\,m^{-3}$ for all layers.

Figure 4.E.1 shows the HVSR for different scenarios of ground frost and different number of considered surface waves modes. All models confirm the qualitative observation that the HVSR experiences a broadband decrease above 1 Hz due to a layer of ground frost with a certain thickness and increased shear wave velocity. Apart from the broadband decrease at higher frequencies, the two resonance peaks below 1 Hz do not seem to be effected. With increasing thickness and shear wave velocity the decrease is more pronounced and the maximum decrease moves to lower frequencies. Note that both parameters show a similar effect on the HVSR. Thus, it is difficult to disentangle the two effects in actual observations. We observe this scenario at the end of February and beginning of March marking the coldest and also the longest period of freezing air temperature (Figure 1b). During that time, the horizontal component and the HVSR experience the strongest decrease. However, we cannot say if an increasing thickness or decreasing temperature dominates the process. The number of surface modes considered in the wavefield has also an effect on the pattern of decrease. It has already been shown that large stiffness contrasts or reversal of velocity layers – that is high-velocity layer over low-velocity layer – can cause modal energy pertubation and dominant higher modes (O'Neill and Matsuoka, 2005). Freezing the soil from the surface downwards causes a reversal of velocity layers and might lead to modal energy pertubation. The broadband high-frequent HVSR decrease and its dependence on the number of modes suggest that this effect occurs. This would be important to consider when passive image interferometry is used for monitoring permafrosts. Dominant higher modes could appear on cross-correlograms during times of refreezing in autumn and corrupt measurements of velocity variations. A proper wavefield analysis would be needed to understand this process better, however, it is out of the scope of this work and, thus, subject to future research.

Overall, the model brings interesting insights to our observations retrieved from the seismic data. The observations and model agree qualitatively on a broadband high-frequent HVSR decrease due to ground frost. The decrease is more pronounced

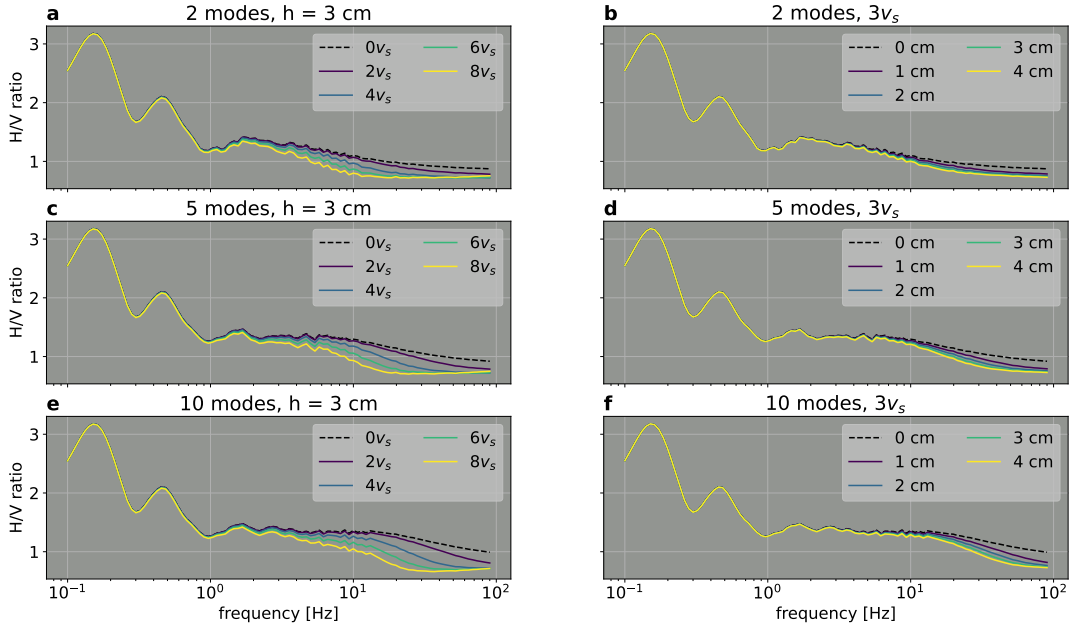

FIGURE 4.E.1: (**a,c,e**) The HVSR in the presence of a 3 cm thick frozen surface layer with varying shear wave velocities and varying number of Rayleigh and Love wave modes. The shear wave velocity of the frozen layer ranges between two-fold and eight-fold of the shear wave velocity of the first layer in the 1D model. The model without a frozen layer is depicted as a black dashed line. (**b,d,f**) The HVSR in the presence of a frozen surface layer with a thickness ranging from 1 to 4 cm and varying number of Rayleigh and Love wave modes. The shear wave velocity is fixed to the three-fold shear wave velocity of the first layer. The model without a frozen layer is depicted as a black dashed line.

for deeper and colder ground frost. Moreover, the model shows that it is difficult to entangle the interaction between the thickness and temperature of the ground frost and surface wave modes present in the wavefield. It is also clear that the HVSR of the seismic data contains many different source and medium effects (Figure 4i) and, thus, the diffusive wavefield assumption is not valid for the data. This highlights the strength of our data-driven approach, which isolates a pattern in the continuous seismograms related to the freezing and thawing process despite all the other source and medium effects affecting the data.

Chapter 5

Exploring seismo-volcanic signatures with machine learning at Klyuchevskoy, Kamchatka, Russia

Article in preparation

In the first application of hierarchical waveform clustering, we mainly focused on the dendrogram for the data exploration task (see Chapter 3). At this stage of the work, we noted that the independent components (ICs) contain useful information which might be also usable for data exploration. In the second application, we identified an interesting pattern in the IC space through the location of the centroids of the clusters (see Chapter 4). This motivated us to analyze more in depth the IC space in the following chapter. However, we did not limit ourselves to ICA and, thus, we also explore other techniques for dimensionality reduction such as PCA and UMAP. This chapter depicts the limitation of linear techniques such as PCA and ICA and gives an outlook on how manifold learning techniques such as UMAP might be an interesting tool for future applications.

5.1 Abstract

We explore and analyze the signal content of continuous three component seismograms recorded in the vicinity of the Klyuchevskoy volcano with methods of machine learning. A non-learnable wavelet-based network, called scattering network, retrieves a stable and time-invariant representation of the seismic time series. With methods of dimensionality reduction, namely manifold learning and principal and independent component analyzes, we retrieve meaningful signal patterns from this high-dimensional data representation in a data-driven fashion. The signal patterns indicate that the recorded wavefield is strongly non-stationary with ever-changing signal characteristics and no repeating patterns. In particular, the months before the eruption in April 2016 are characterized by rapid changes of signal characteristics, while the signal patterns seem more stable after the eruption. Our results confirm the idea that volcanic tremors are not a single class or a set of classes but a continuously changing signal, witnessing many different types of phenomena. Data-driven methods as we present in this study hold the potential to provide new information about known and unknown signal classes, perhaps enhancing our understanding of volcanic systems. In particular manifold learning techniques seem to be a great tool to visualize the signal content of large time series in a two-dimensional map.

5.2 Introduction

Volcanoes are complex active geological objects which are often described with a conceptual model. Detailed knowledge about the storage and transportation of magma at different depths remains limited and many different types of instruments are utilized to close these knowledge gaps. Continuous seismograms are part of the heterogeneous dataset recorded at volcanic environments where a large variety of seismic signals with different characteristics are recorded. The identification of these signals and their underlying physical mechanism contribute to the knowledge gain about the inner workings of a volcano. The seismo-volcanic research community has established different families of seismo-volcanic signals according to their observed signal characteristics. The analogous to pure tectonic earthquakes are the volcanic-tectonic earthquakes (VTs) which have a broadband signature and usually a clear *P* and *S*-wave arrival. Resonances of fluid-filled cracks or conduits can result in so-called long period events (LPs), which have less high-frequency energy than the VT events and a less clean signature of phase arrivals (Chouet, 1996). Both VTs and LPs are transient signals lasting mostly a few seconds. Hybrid events are a mixture of LPs and VTs and make the boundary between LPs and VTs fuzzy (White et al., 1998). Other transient signals related to volcanic activity are explosions, tornillos or rockfalls at the flank of a volcano. Besides transient events lasting a few seconds, volcanoes are also known to produce volcanic tremors which can last from minutes to years (Konstantinou and Schlindwein, 2003). Their appearance in frequency and amplitude can vary largely: volcanic tremors can be of monochromatic nature, cover a narrow frequency band or glide across different frequencies (Julian, 1994; Hotovec et al., 2013). Some studies observed a continuous transition from LPs to tremor episodes and back. This indicates that some types of tremors can be described as a rapid succession of LPs where no clear onset of an individual LP is visible in the seismogram (e.g. Latter, 1979; Fehler, 1983). Although all those mentioned signals belong to human-defined meta-classes, they may have subtle differences that witness the dynamic processes of the volcanic system. In particular, the class of volcanic tremors indicate a non-stationary phase of the volcanic system and change their signal characteristics accordingly. Moreover, the characteristic of tremor signals vary from volcano to volcano, making a generalization of this signal class even more difficult (Konstantinou and Schlindwein, 2003). Studying and analyzing volcanic tremors is difficult since the seismological processing tools adapted to earthquake seismology are not well-suited for tremor-like signals. The lack of sharp amplitude attacks complicates the detection task and the lack of clear phases complicates the localization of tremor sources. Recent studies have developed a network-based method with a sliding window to identify and locate the most prominent tremor source within the considered window (Soubestre et al., 2018; Soubestre et al., 2019).

The large variety of known signal classes and their possible subtle changes indicating changes within the volcanic complex motivate the application of exploratory data analysis to seismic time series recorded in the vicinity of volcanoes. Identifying these subtle changes in signal characteristics in a data-driven fashion can reveal new insights about the stationary or non-stationary state of a volcano. In the following we apply the strategy presented in Chapter 2 to continuous three-component seismograms recorded in the vicinity of the Klyuchevskoy volcano during an active period which resulted in an eruption in April 2016. We stop the strategy shown in Figure 2.1 at the feature generation and utilize both PCA and ICA as techniques for

exploratory data analysis to identify interesting patterns in the data. Moreover, we analyze two different pooling operations to understand better how the pooling acts as a filter of information. This study has multiple aims: (1) understanding better the representation of the scattering coefficient matrix with respect to the pooling operation, (2) testing PCA and ICA as methods for seismic data exploration and (3) revealing interesting patterns in the seismic data related to volcanic activity.

5.3 Exploratory data analysis with principal and independent component analysis

PCA and ICA have been widely applied to various facial recognition tasks with images (Turk and Pentland, 1991; Draper et al., 2003; Delac, Grgic, and Grgic, 2005). While PCA finds components which describe global structures of the images such as lighting from different angles, ICA finds components which describe local structures such as the mouth or eyebrows. Both methods are also considered blind source separation techniques, aiming at disentangling a recorded mixed signal into its superposed source signals. In particular, ICA is designed as such a method and has been largely applied in identifying source signals for functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and magnetoencephalography (MEG) time series data. While it successfully recovered interesting components for fMRI data, it mostly identified artifacts in the EEG/MEG data due to blinking or facial muscle movements (see e.g. Jung et al., 2000). The authors of Hyvärinen et al., 2010 argue that this is a result of ICA searching for non-Gaussian distributions. The artifacts have a very non-Gaussian distribution and the interesting brain activity are more Gaussian, since they resemble amplitude-modulated oscillatory activity. Even if the brain activity has some degree of non-Gaussianity, the artifacts are super non-Gaussian and, thus, a higher number of components is needed to recover the interesting signals related to brain activity. As an alternative, the authors of Hyvärinen et al., 2010 propose to apply ICA on the short-time Fourier transform (STFT) of the EEG/MEG time series, which resembles our approach of applying ICA to the scattering coefficient matrix. With their approach, they are able to identify modulated signals related to rhythmic brain activity. They conclude that ICA is always biased towards certain sources. If applied in the time domain, it mostly finds sources with non-Gaussian amplitudes such as artifacts. If applied in the Fourier domain, it mostly finds narrow-banded sources. The discussed applications motivates us to apply PCA and ICA not only as a technique to reduce the dimensions of the scattering coefficient matrix for clustering but also as a method to explore the data.

5.4 The data and setup of the scattering network

From July 2015 to July 2016 a temporary network called *KISS* was installed around the Klyuchevskoy Volcano group (KVG) in Kamchatka, Russia, in order to better understand the crustal magmatic plumbing system. During that time an eruption unfolded at Klyuchevskoy in April 2016, preceded and accompanied by a range of different seismic activity (Shapiro et al., 2017). The authors of Journeau et al. (2022) detect and locate volcanic tremor sources, revealing a trans-crustal magmatic system beneath the KVG (see Figure 5.1). The locations cover a wide region and multiple depths, picturing a large and dynamic trans-crustal magma system. Their results

strengthen the concept that tremor signals are of dynamic nature, containing interesting information about the state of the volcano. Besides tremors, they also detect deep long period events (DLPs) and VTs. Their catalogs provide additional information aiding the exploration and interpretation of the data with PCA and ICA.

In this study, we analyze the three component continuous seismograms recorded by station SV13, which is located directly above the tremor and DLP locations (see Figure 5.1B). We cover the complete operational time period of this station ranging from August 2015 to July 2016. The seismic data is demeaned, detrended and down-sampled to a sampling frequency of 25 Hz. We set up a two-layered scattering network with Gabor wavelets (see Figure 5.2). We define a sliding window of 20 min with 10 min overlap, which corresponds to the same time resolution of the tremor catalog provided by Journeau et al. (2022). The first layer wavelets are adapted to the possible frequency content of the tremors; their center frequencies range from 0.78 to 10 Hz with a logarithmic grid (see Figure 5.2a and b). The second layer wavelets start at much lower frequencies since they gather information about the modulation and shape of the signal (Figure 5.2c and d). The first layer covers 4 octaves and is densely spaced with 4 wavelets per octave. The second layer covers 8 octaves and is sparsely sampled with 1 wavelet per octave.

5.5 Results

5.5.1 Pooling: an information filter

The pooling operation turns the scalograms of the wavelet transform into scattering coefficients, providing a translation invariant representation of the seismic data. However, the pooling operation also reduces the information by collapsing the time axis of the pooling window. Different pooling operations filter the information differently, favoring different signal characteristics. Before we apply the scattering network to the complete time series, we analyze the scattering coefficient retrieved with maximum and median pooling for a 20 min seismogram recorded at station SV13 (Figure 5.3). The dominant signal in this 20 min seismogram is a broadband transient event arriving after 800 s and lasting for 100 s. Moreover, there are also persistent harmonic tremor signals around 0.8 and 2 Hz with a lower amplitude than the transient event. Besides the broadband transient and the tremors, we can identify changing amplitudes at frequencies around 10 Hz. This example shows the variety of signals of a single seismogram and we must acknowledge that any representation without time information - such as the Fourier spectrum or scattering coefficients - will simplify the data. The information retrieved by the scattering coefficients depends largely on the settings of the scattering network: number of wavelets, frequency range of wavelets and pooling operation. Figure 5.3c, d and e show the median and maximum pooled scattering coefficients together with the Fourier spectrum. The first-order maximum pooled coefficients resemble a smoothed Fourier spectrum (Figure 5.3c). The first-order median pooled coefficients are lower in amplitude and contain different local maxima and minima. They show larger amplitude for the two frequency bands with the harmonic tremors and lower amplitudes in between. The transient event with large amplitudes between 0.2 and 10 Hz seems to have no influence on the median pooled coefficients. In contrast, the maximum pooled coefficients have an amplitude distribution which matches much better the transient event. The type of pooling operation, which transforms the scalogram into

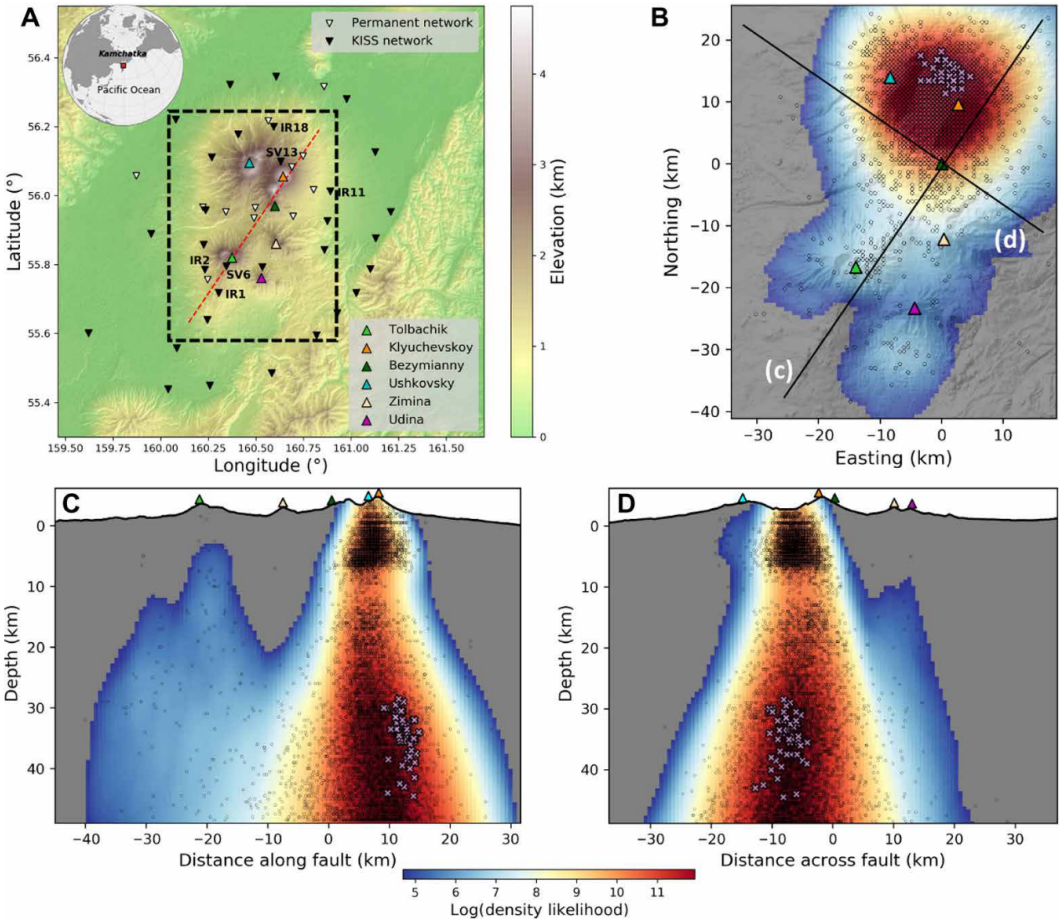

FIGURE 5.1: Map and tremor and DLP locations retrieved from Journeau et al. (2022). **(A)** shows a map of the Klyuchevskoy Volcano Group (KVG) together with the permanent stations and the temporary KISS network. The red dashed line marks the fault across the KVG. **(B)** shows a zoom onto the black dashed box in **(A)**, **(C)** shows a cross-section along the fault with the and **(D)** shows a cross-section across the fault. The locations of the tremors are depicted in black and the locations of DLPs are depicted in purple.

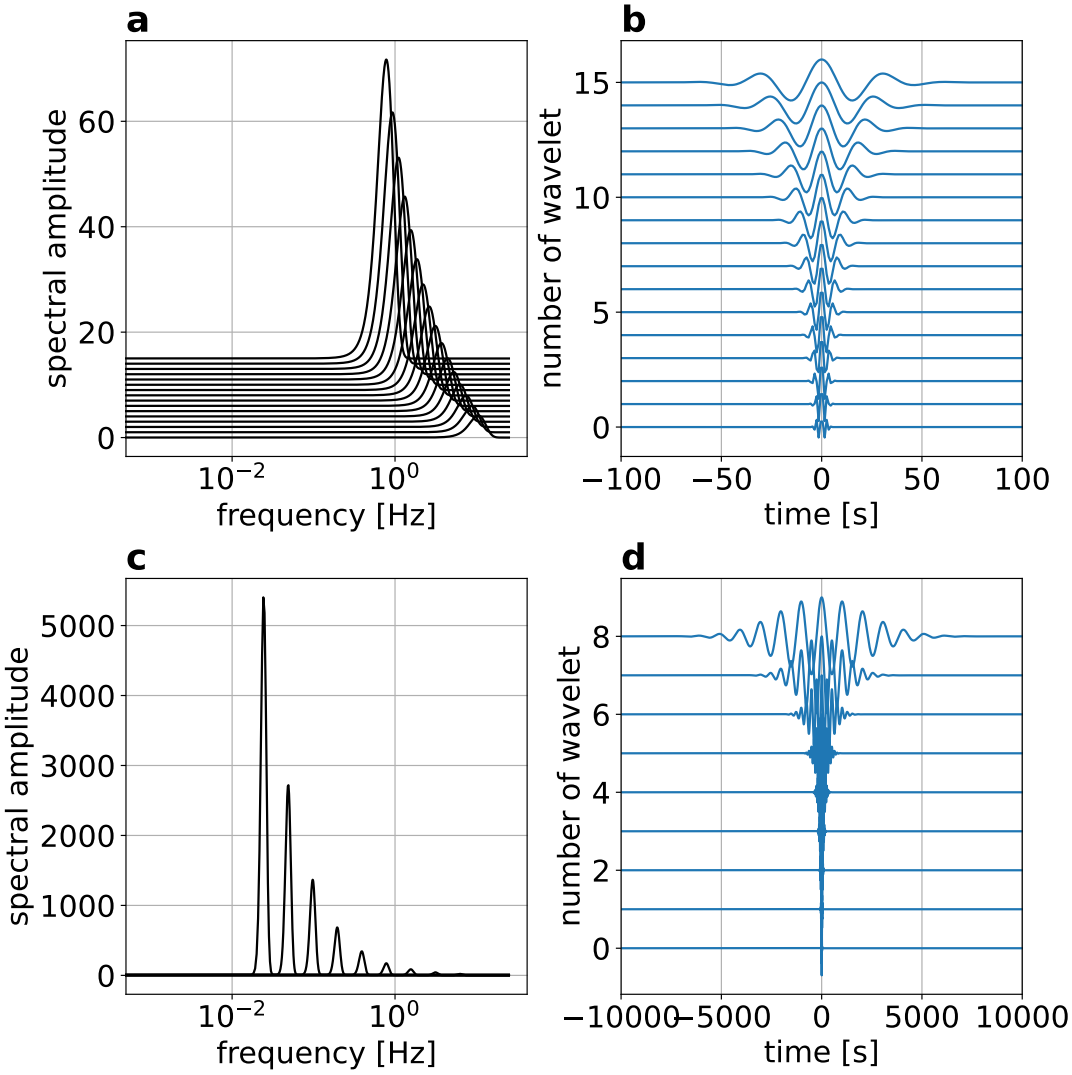

FIGURE 5.2: Setup of wavelets for the two-layer scattering network. **(a)** Amplitude spectra of the first-layer Gabor wavelets. **(b)** The real part of the first-layer Gabor wavelets in time-domain. **(c)** Amplitude spectra of the second-layer Gabor wavelets. **(d)** The real part of the second-layer Gabor wavelets in time-domain.

scattering coefficients, filters the data and stores different type of information. Median pooled coefficients contain the information of the background wavefield and ignore any short lived transients in the seismogram. Maximum pooled coefficients are sensitive to any type of short-lived transient in the seismogram which mask the background wavefield. Note also that maximum pooling would save the information of two transient events, if they appear in different frequency ranges. Thus, it could be a representation of a mixture of large amplitude events with different frequency content. Both pooling operations are valid, however, we need to acknowledge that both representations are biased and simplify the nature of the seismic data. This is important to consider if we apply exploratory data analysis techniques. For the further analysis, we consider median pooling, since we want to focus on tremor signal.

5.5.2 Visual analysis of the scattering coefficient time series

Figure 5.4 shows the time series of the median pooled first- and second-order coefficients of the east component data recorded at station SV13 together with the tremor and DLP catalog from Journeau et al. (2022). The first-order coefficients are easy to analyse visually and resemble a STFT (Figure 5.4a). Note that the presentation of the second-order coefficients is different to the second-order coefficients shown in Figure 5.3, since we vectorized the coefficients in order to show the time information (Figure 5.4b). The top of the y-axis shows the scattering coefficients of all f_2 with the highest frequency f_1. The bottom shows the scattering coefficients of all f_2 with the lowest frequency f_1. Due to the cascading of wavelet operation from first to second layer, the information of the first layer is encoded in the second layer. Visually, these coefficients do not seem informative, however, they are important for classification or clustering tasks (Andén and Mallat, 2014). Generally, the coefficients show larger amplitudes during intense tremor periods. For instance, the onset of tremor at the beginning of December 2015 is clearly marked by a rapid amplitude increase.

The visual inspection of the scattering coefficients of the east component indicate that they contain meaningful and interpretable information related to the tremor activity. However, we also realize that the data, in particular the second-order coefficients, are too large to observe interesting patterns in detail. This demonstrates the limitations of visual inspections of large datasets and motivates the need for tools which retrieve the interesting information in a human-readable fashion. Note also that we only show the east component of the three-component data. The three component data increases the data size and a visual analysis becomes even less feasible.

5.5.3 Principal components (PCs)

By applying PCA to the whole scattering coefficient matrix (all three channel and first- and second-order coefficients), we explore the data as an ensemble and retrieve information which are hard to spot by just looking at the whole data matrix. The explained variance retrieved with a PCA estimates how much information each principal component (PC) contains and is directly linked to the eigenvalues. Figure 5.5a shows the cumulative variance ratio with an increasing number of PCs. The curves show how much additional information is added with an increasing number of components. If the cumulative variance ratio reaches 1, we have enough components to explain all the variance of the original data and we can reconstruct the original data without any information loss. The first PC explains more than 85 % of

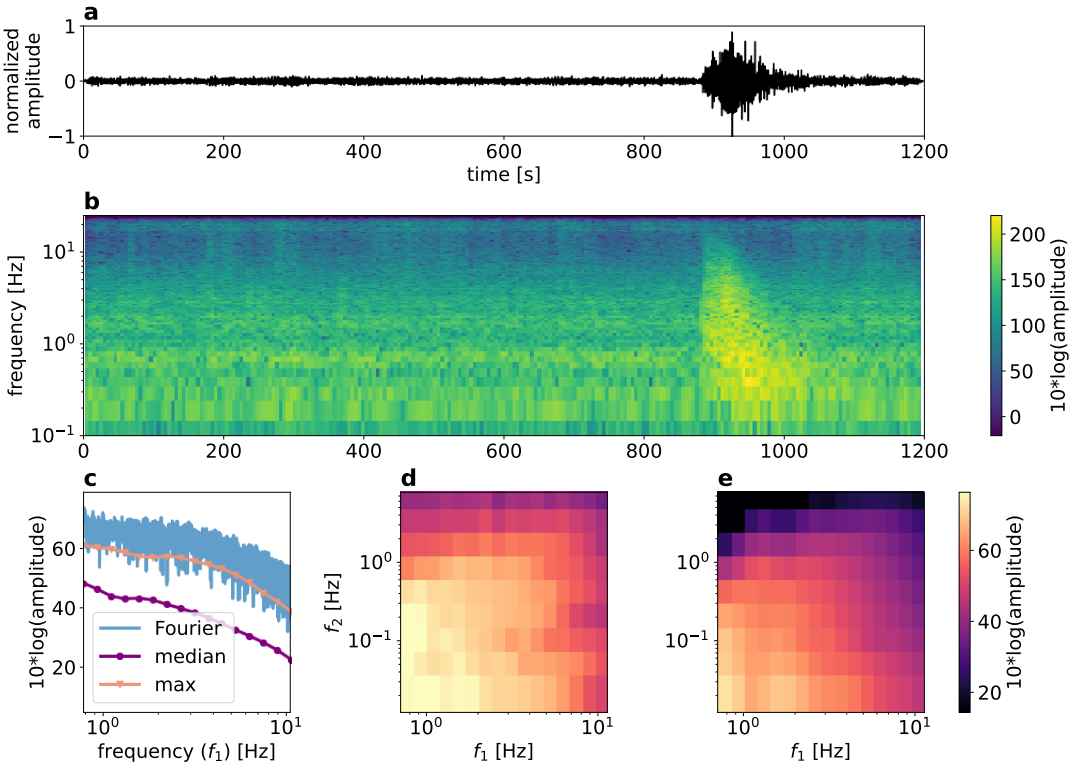

FIGURE 5.3: Comparison between Fourier spectrum and scattering coefficients of a seismic signal. **(a)** shows an example seismogram with normalized amplitude in time domain. **(b)** shows its corresponding Fourier spectrogram. **(c)** shows the Fourier amplitude spectrum and the first order median and maximum pooled scattering coefficients of the signal shown in **(a)**. **(d)** shows the second order maximum pooled scattering coefficients and **(e)** shows the second order median pooled scattering coefficients as a function of the center frequencies f_1 and f_2.

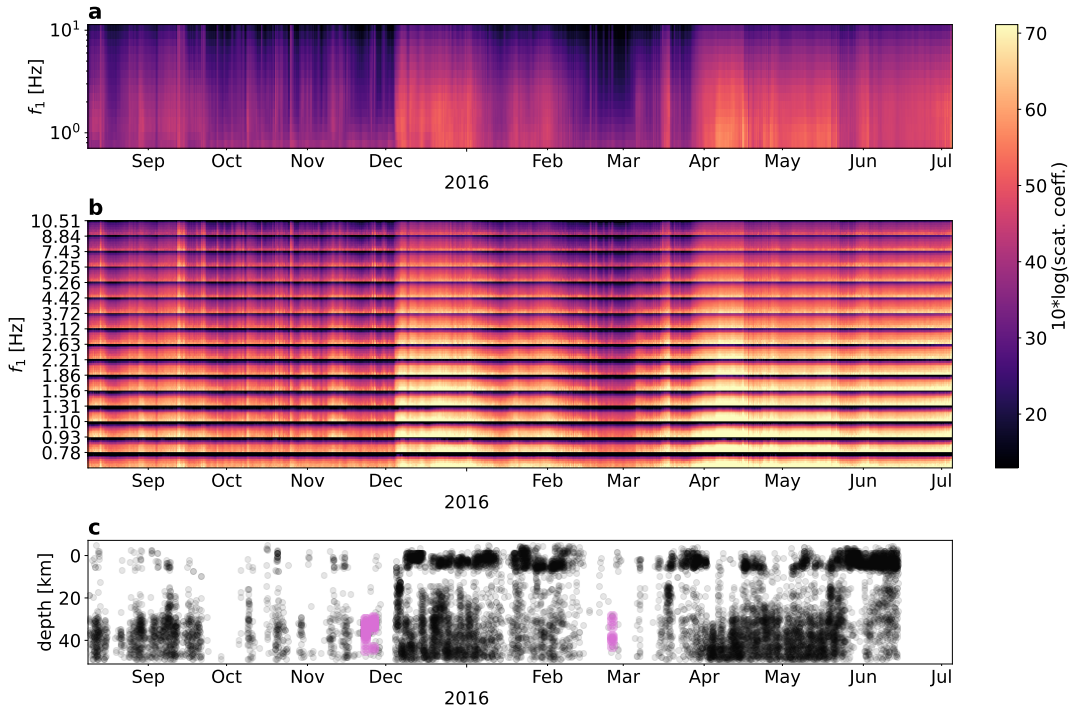

FIGURE 5.4: **Time series of scattering coefficients and catalogs (a)** Time series of first-order scattering coefficients of the east components of SV13. **(b)** Time series of second-order scattering coefficients of the east components of SV13. **(c)** The tremor and DLP catalog as a function of calendar time and depth. The black dots correspond to tremor detections and the light magenta crosses correspond to DLP detections occurring at the 23 and 26 November 2015 and 25 February 2016. The eruption unfolded in April 2016.

the variance. Thus, we could reconstruct the scattering coefficient matrix with the first PC and lose about 15 % of information. The second component is already much less informative with about 5 % explained variance ratio. With increasing number of components, less and less information is added. This is expected, since PCA orders the PCs according to their eigenvalues.

Figure 5.5b shows some of the PCs, that is the rows of the feature matrix displayed in Figure 2.1. The first 10 PCs show a diverse set of patterns containing slow and fast variations of its amplitude of arbitrary unit. These patterns are likely related to interesting information about the seismic wavefield. The 51st to 60th components appear noisy, indicating less meaningful information than the first 10 components. It is possible that these components are related to patterns caused by instrumental noise or random fluctuations in the seismic wavefield. The 201st to 210th component are even noisier and less informative. The visual inspection of the components show that higher-order components contain less and less interesting information regarding patterns of interest in the seismograms. Often, the explained variance ratio is utilized to justify how many components to keep for further analysis such as clustering. The visual inspection of the components agrees that the explained variance ratio seems to be a good indicator for informative components. However, we would argue that the first 10 components contain interesting information despite the rapid decreasing explained variance ratio. The discriminative power might lay within higher-order components, especially for signals occurring rarely in the time series. This let us conclude that the decision on the number of components for further analysis should be based on the explained variance ratio together with a visual inspection of the components.

The loadings of the components

The PCs show us temporal patterns of the scattering coefficient matrix. To understand better these patterns, we can look at the matrix $\mathbf{A}$, which transforms the scattering coefficients to the PCs (see Equation 2.11). The columns of this matrix contain the right singular vectors of the centered scattering coefficients, which are parallel to its eigenvectors. They indicate the direction of the principal axis in the scattering coefficient space and are often called *loadings* in the literature. The loadings help us understanding what part of the scattering coefficient space is important in constructing the PCs and which coefficients correlate along the principal axes. Figure 5.6 shows the loadings assigned to the first- and second-order coefficients to construct the first 10 components and the 51st to 60th components. The blue and red color indicate which scattering coefficients are correlated along a given principal axis. Two pixels showing the same color are correlated and two pixels showing a different color are anti-correlated. The intensity of a color shows us how strongly these pixels are either correlated or anti-correlated.

The first PC, explaining more than 85 % of the data, shows high correlation between all scattering coefficients for both scattering coefficients representation. This means that if we move along this axis, we increase or decrease all scattering coefficients simultaneously. The red colors of the loading pattern of the second-order coefficients also indicate a high correlation between all coefficients. It seems that the first PC reflects the variation of broadband energy. The loadings of the second PC show an anti-correlation between high and low frequencies on all three components for the first- and second-order scattering coefficients (Figure 5.6). The scattering coefficients with f_1 below 3 Hz are correlated (colored blue) and the scattering coefficients with f_1 larger than 3 Hz are correlated (colored red). However, both regimes

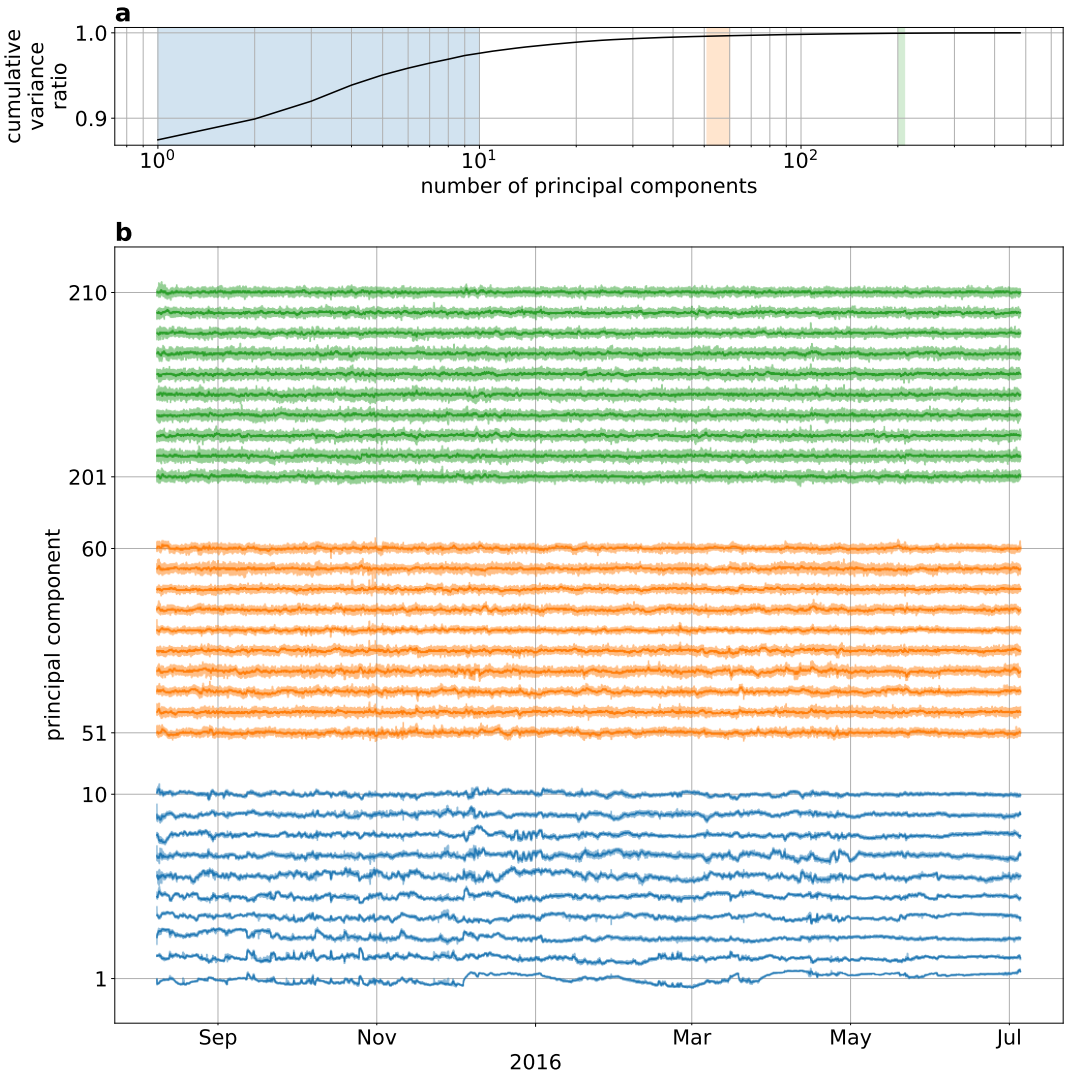

FIGURE 5.5: **(a)** cumulative variance ratio for the the PCs of the scattering coefficients matrix. The color-coded areas indicate which 30 PCs are shown in **(b)**: the first 10 PCs are shown in blue, the 51st to 60th PCs are shown in orange and the 201st to 210th PCs are shown in green. The saturated lines show the PCs smoothed by a median filter considering 51 samples.

are anti-correlated, indicated by red and blue colors. Therefore, high-frequency and low-frequency signals are separated along the axis of the second PC. Note that the loadings of the first 10 components show large alternating amplitudes for the second order coefficients with $f_2 > f_1$. These scattering coefficients have no physical meaning, since the envelope frequency f_2 is larger than the signal frequency f_1. The alternating patters indicates that they cancel each other out, giving them little importance in the PCA.

The interpretation of the loadings for the first two PCs revealed that the components can be related to global signal characteristics in the data. The following 8 components show similar patterns with smooth loading patterns where neighboring points have similar values. Hence, they seem to contain meaningful and physical information describing seismic signals. In contrast, the loadings for the 50th PC and higher are less smooth and neighboring points have large differences. This confirms the assumption that these components might be related to random wavefield fluctuations or instrumental noise. Our interpretation resembles the interpretation of PCA applied to facial images where the first components find global image characteristics and higher order components depict complex and not interpretative patterns (Draper et al., 2003). The authors of Unglert and Jellinek (2017) applied PCA to spectra of tremor signals recorded by single stations at different volcanoes and, interestingly, their results indicate a similar interpretation for their first two PCs. In the following, we want to take the interpretation further and compare the PCs to the appearance of tremors.

Comparison of principal components with the tremor catalog

Figure 5.7 show the tremor catalog in time and depth compared to first three PCs and their reconstruction of the first-order scattering coefficients of the east component. We can reconstruct the scattering coefficients by taking the outer product of the PCs and the pseudo inverse of the loading matrix $\mathbf{A}$. By zeroing all components except one, we isolate the reconstruction based on a single component. The comparison of the first PC and its reconstructed first-order scattering coefficients with the tremor catalog shows how the broadband energy is varying and how it correlates with the appearance of tremor (Figure 5.7a and b). The first PC marks clearly the onset of tremors in December and shows that the broadband energy is stable until mid January. Then, it decreases slowly and unsteady until it reaches its minimum at the beginning of March, when the catalog indicates an absence of tremors. Interestingly, the broadband energy increases at the beginning of March, before the catalog indicates the next longer-lasting tremor period later in March. It is possible that the tremors start earlier than the catalog indicates, since the catalog relies on a certain amount of tremor energy for localizing the tremor source (Journeau et al., 2022).

Similarly to the first PC, we can reconstruct the scattering coefficients based solely on the second PC and compare with the tremors (see Figure 5.7c and d). This comparison shows us which tremor periods contain more than usual high- or low-frequency content. For example a large part of deep tremors in August and September contain more low frequency energy. On the contrary, the beginning of tremor period in December has more high-frequency content. After a slow change, mid-december to mid-january indicate a balance between high- and low-frequency energy, since the component is centered around 0. At mid January the amplitude of the component decreases rapidly correlating with a slight spatial change of the tremors indicated by the catalog.

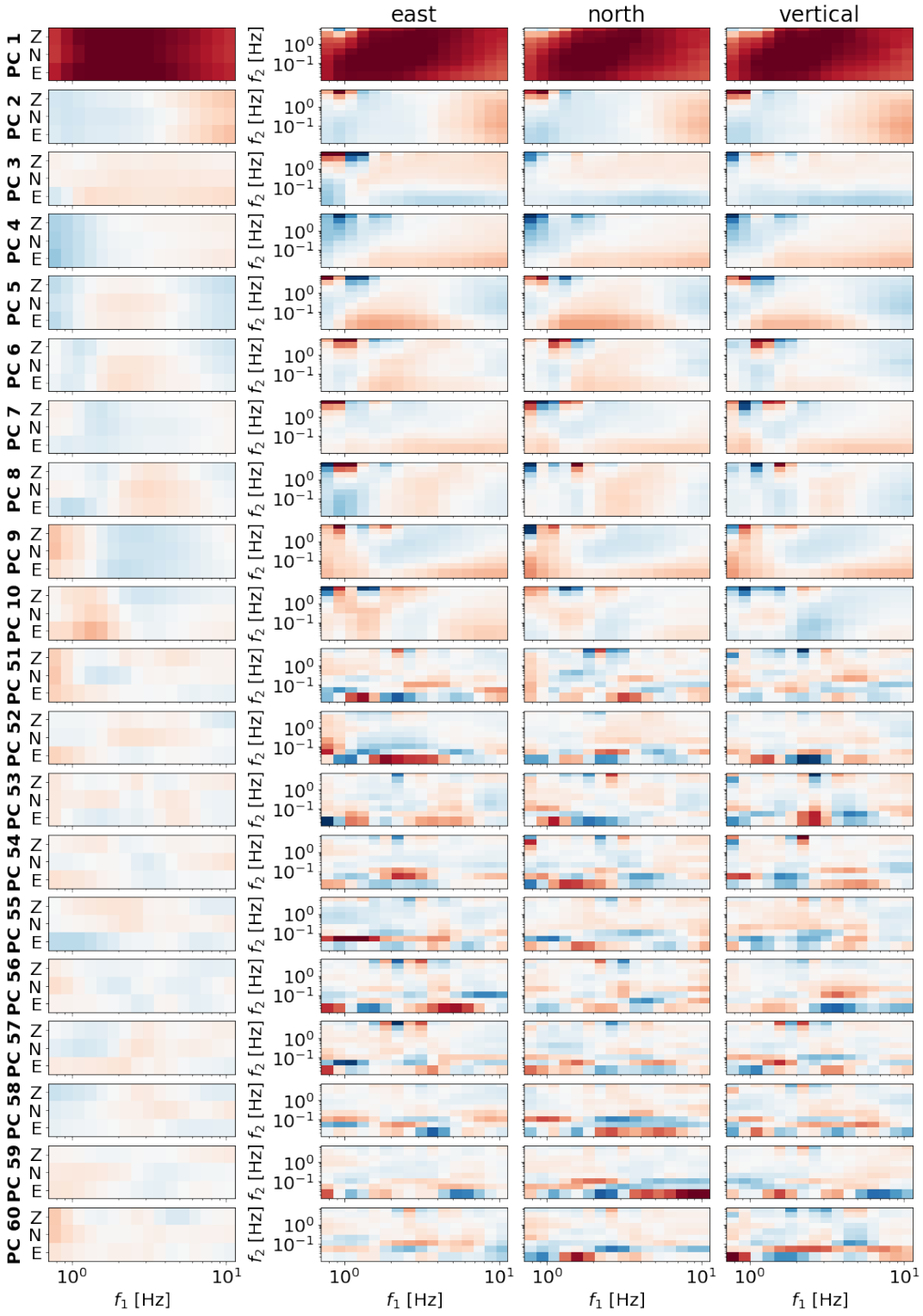

FIGURE 5.6: The loadings of the first 10 PCs and of the 51st to 60th PC, indicating the direction of the respective PC in the scattering coefficient space. Left column shows the loadings corresponding to first-order scattering coefficients of all three components E (east), N (north) and Z (vertical). The columns to the right show the loadings corresponding to second-order scattering coefficients of the east, north and vertical component, respectively. Red colors indicate positive values and blue colors indicate negative values.

The third PC seems to mimic the activity of deep tremors during the beginning of the recording time (see Figure 5.7e and f). Positive amplitudes of the PC correlate well with the appearance of deep tremors before December (Figure 5.7e). At later times, the amplitude variation of this component is much smaller and stays around 0. The loadings of this component show that the east components has a slightly different pattern than the north and vertical components (PC 3 in Figure 5.6). Indeed, the spectrograms which correspond to the largest positive values of the third PC show a strong monochromatic tremor signal around 1 Hz, particularly strong on the east component (see Figure 5.A.1 in the Appendix of this chapter). The interpretation of the third component shows that, besides global signal characteristics, we also find components which correspond to a certain type of tremor signal. We stop the detailed analysis of the components here, since we just wanted to show that the PCs find meaningful directions in the scattering coefficients space, making them interpretive and revealing interesting patterns.

5.5.4 Independent components (ICs)

Compared to PCA, ICA does not rank the components, since no eigenvalues are attached to them. Therefore, we can not attribute a quantity of information gain for the individual component such as the explained variance ratio. As an alternative, we can compute the overall reconstruction loss, measuring the difference between the original data and the reconstructed data based on a given number of components. Figure 5.8 shows the decreasing reconstruction error with increasing number of ICs. Thus, the information gain becomes smaller with increasing number of components. Around 7 components the curve shows a little kink and the decrease becomes flatter afterwards. Often, a model around the kink is used for further analysis, since a larger number of components adds less and less information. However, the choice of one model is mainly interesting if there is a specific task to solve. In our case of exploratory data analysis, it is interesting to look at different models, since ICA finds different solutions for different number of components. In contrast, PCA only adds components for an increasing number of components, it does not affect the already found components.

5.5.5 Low number vs. high number of independent components

Figure 5.9 shows the ICs for an ICA model with six components. The components are not ordered and contain no information about the amplitude or sign. Similarly to PCA, we obtain a loading matrix which can aid the interpretation of the ICs (Figure 5.10). For ICA, this matrix is often refereed to as the *mixing* matrix, since this matrix is the linear operator which mixes the ICs to retain an estimation of the original data. Compared to the PCs shown in Figure 5.5b, we can identify similar type of patterns within the ICs. For instance, IC 2 resembles the PC 3 but flipped (Figure 5.9). The similarity of these two components are backed up by the similarity of their loadings, which reveal that their axes are closely aligned in the scattering coefficient space (see PC 3 in Figure 5.6 and IC 2 in Figure 5.10). On the other hand, we also retrieve new patterns such as the third IC which is relatively sparse and shows large amplitudes only in December 2015.

Since ICA is a generalization of PCA, it is not surprising that ICA finds some components similiar to PCA. However, this changes drastically if we increase the number of components. Figure 5.11 shows an ICA solution with 50 components

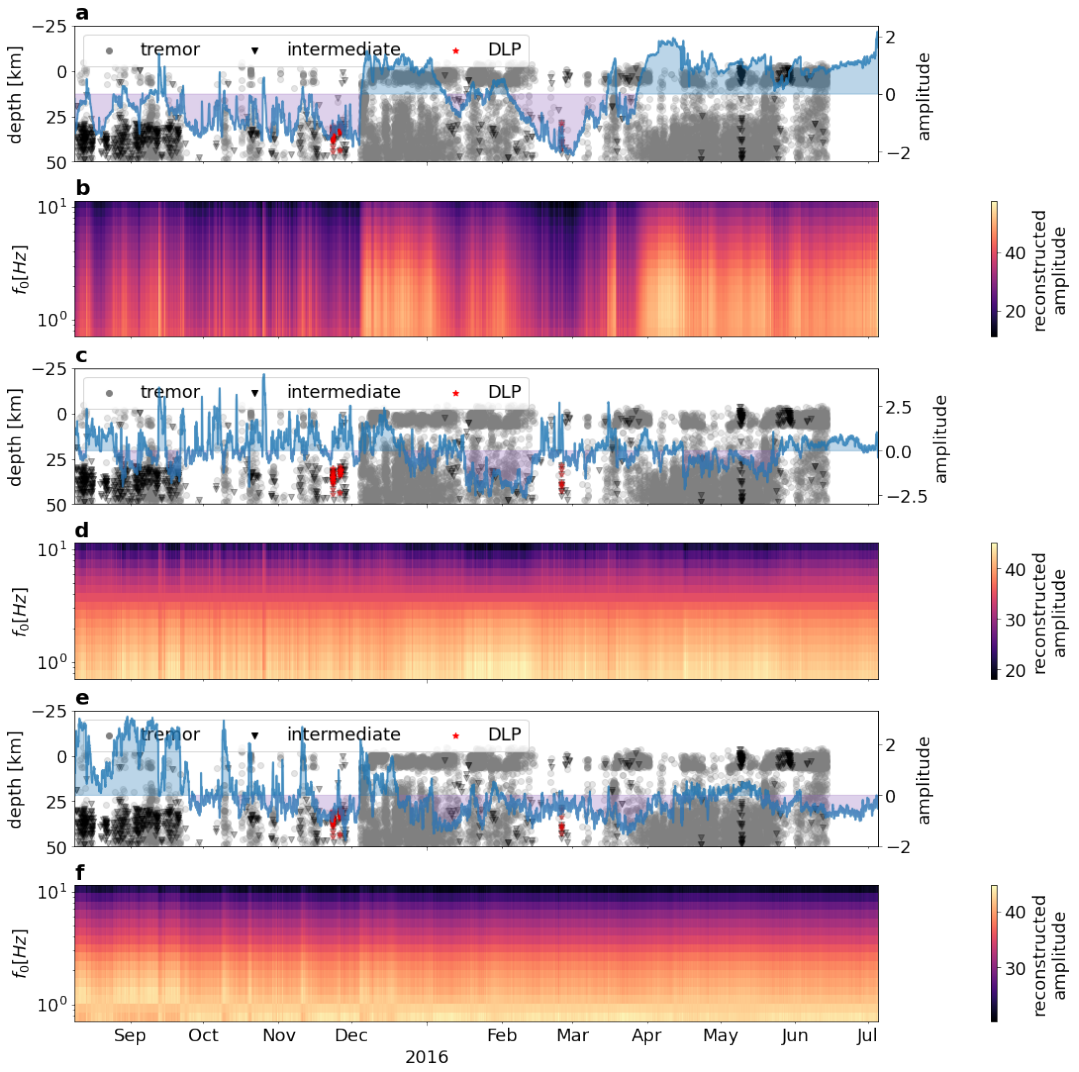

FIGURE 5.7: Comparison between the catalog describing the tremors and DLPs and the third PC. **(a,c,f)** Grey points are tremor detections, black triangles are tremor detections containing DLP swarms, red triangles are DLP detections. The blue line shows the first PC in time with blue-shaded area for positive amplitude and red-shaded area for negative amplitude. **(b,d,f)** Reconstruction of the first-order coefficients of the east component based on the first, second and third PC, respectively.

which are either sparse or noisy. This is very different to the 6 component ICA solution displayed in Figure 5.9 and to the PCA solution from Figure 5.5b. In fact, not a single component of the 6 component ICA solution can be found in the 50 component ICA solution. This example shows how ICA aims at unmixing different patterns and maximizing independence between its components. A larger number of components allows ICA to look for finer details in the patterns and separate them accordingly on different components. Figure 5.12 reveals more clearly what is happening with an increasing number of ICs. The components are median filtered to suppress the noise and to reveal better the long-term trend of each component. Then, the modulus of the components are normalized by its cumulative sum for each time step. This reveals how sparse the data representation is. If the representation is not sparse, one observation in time is equally explained by all components. Thus, the amplitudes of the components will be of similar value between 0 and 1. If the representation is sparse, one component dominates and all components except the dominating one are closer to 0. Moreover, we sort the processed components by its maximum absolute amplitude in time to reveal a potential evolution in time. The 6 component ICA model is not a sparse representation and most of the observations are explained by multiple components. Moreover, all components seem to be important throughout the time. The 20 component model already becomes sparser and each component shows a larger localized amplitude at a given time. In particular, the tremor dominated time period between December and April is described by a succession of sparse components. This characterization is even stronger for the 50 component model where almost each component seems to be associated to a time period. The time period until December is described by ca. 15 sparse components, which seem to be important not only at one time period but at multiple. Then, the start of the main tremor period at the beginning of December is represented by a rapid succession of components which are only activated once. This indicates that the seismic signal characteristics are changing rapidly due to the start of tremors. The quick signal evolution seems to slow down after mid December. The interplay between a rapid succession of components and a slower change of components seems to continue until the end of the recording time. Note that some components are black throughout the whole time, contributing not much information. These components are the noisy and Gaussian ones such as component 9 in Figure 5.11. Nevertheless an overwhelming amount of components contain interesting information for further analysis and interpretation. This shows the limitations of linear dimensionality reduction technique: a large number of components is needed to describe the interesting structure of the data. We can easily miss interesting information if we only retrieve a low number of components. The same holds true for PCA, where at least 10 components seem to contain interesting information. Nonlinear techniques such as UMAP are able to capture more information on a smaller number of components by preserving the local structure of the manifold. In the following we only give a short outlook on the potential of these techniques for future analysis.

5.6 Creating seismic signal maps with PCA and UMAP

Figure 5.13a and c attempt to deliver the broader picture of the data by plotting the first against the second PC, which accounts for more than 90 % of the data's variance. Earlier we related the first PC to the variation of broadband energy and the second PC to the variation of low- vs. high-frequency content. If we apply this knowledge

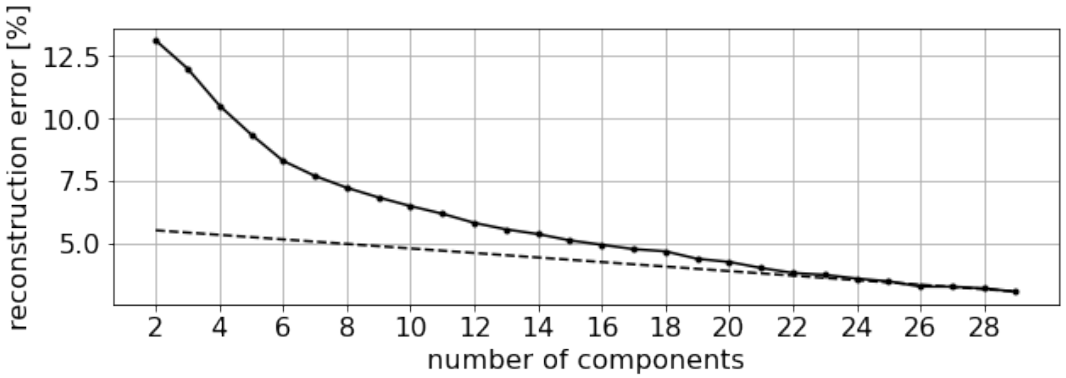

FIGURE 5.8: Reconstruction error for ICA models with different number of components.

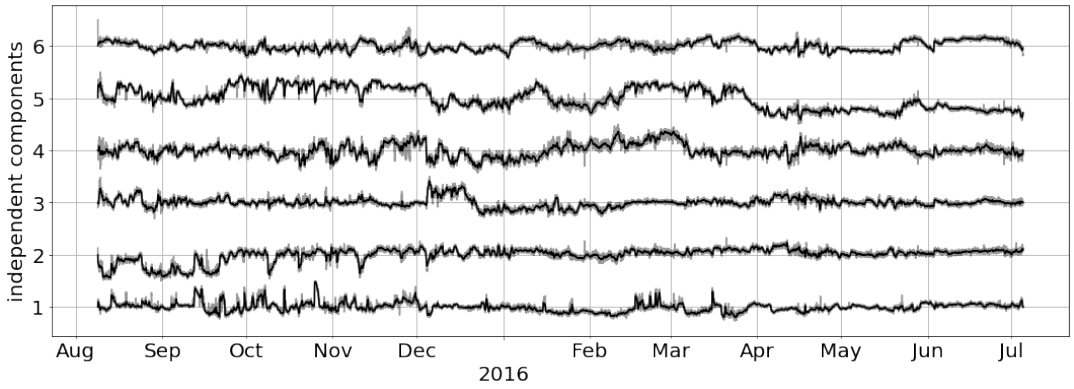

FIGURE 5.9: ICA model with 6 independent components shown in grey. The black lines are the components smoothed with a median filter.

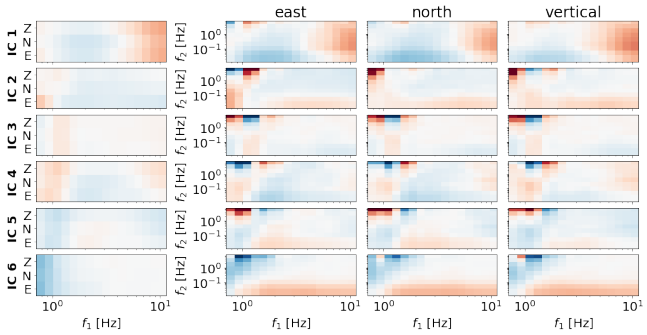

FIGURE 5.10: The columns of the unmixing matrix, indicating the direction of the respective independent component (IC) in the scattering coefficient space. The left column shows the loadings applied to the first-order scattering coefficients of all three components. The right columns show the loadings applied to second-order scattering coefficients of the east, north and vertical component, respectively. Red colors indicate positive values and blue colors indicate negative values.

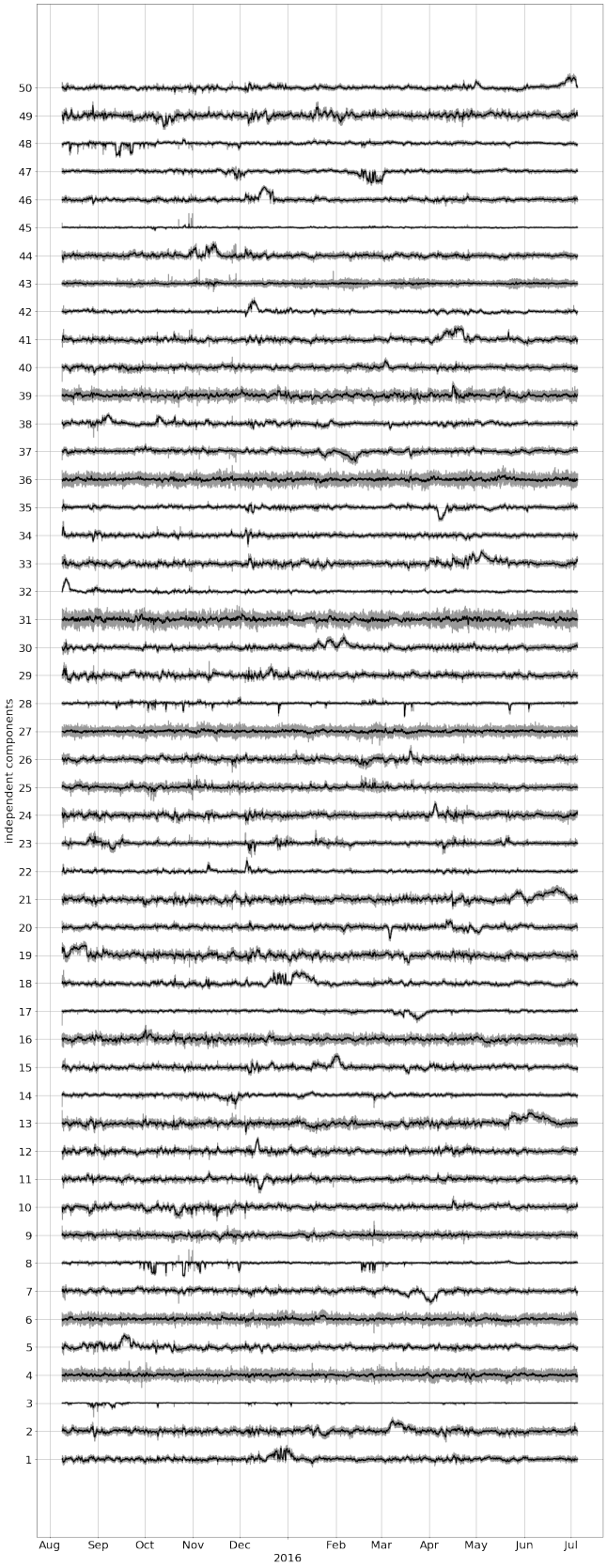

FIGURE 5.11: ICA model with 50 independent components shown in grey. The black lines are the components smoothed with a median filter.

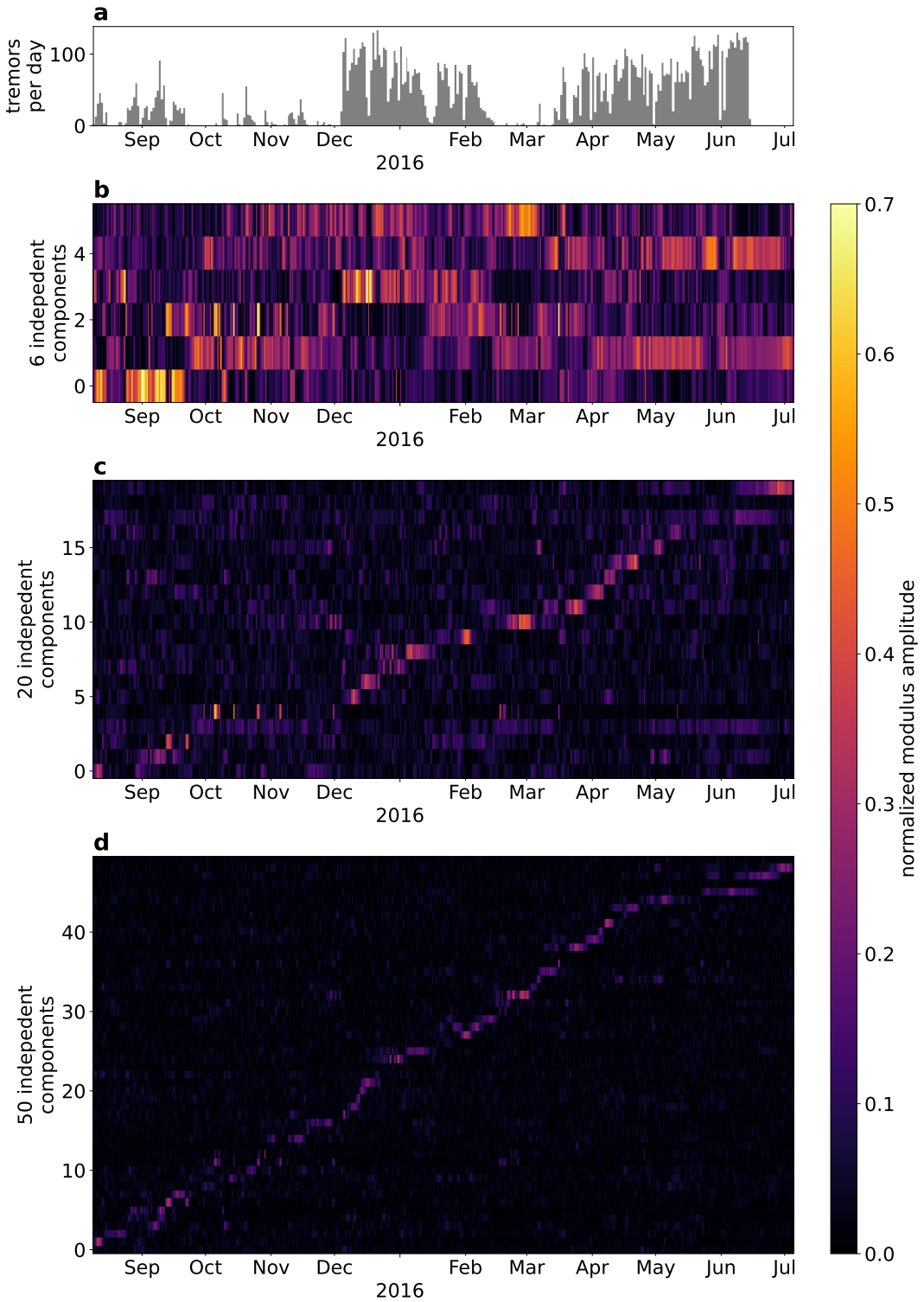

FIGURE 5.12: **(a)** Daily number of tremor detections. The normalized modulus amplitude of independent compontents from ICA models with 6 components in **(b)**, 20 components in **(c)** and 50 components in **(d)**. The components have been normalized per observation in time.

to the two-dimensional PC space depicted in Figure 5.13a and c, we can identify northern area with high-frequency signals, the eastern area with large-amplitude signals, the southern area with low-frequency signals, and the western area with low-amplitude signals. We could call it a seismic signal map where different regions correspond to different types of seismic signals. With this map, we can make two general statements about the data. First of all, the time axis seems to be encoded in the seismic data, which is similar to the observation we made with ICA in Figure 5.12. We have agglomerations of similar color-coded data points which seem to overlap. Secondly, the cloud of data points shows barely any hard cluster boundaries, suggesting smooth transition of signal characteristics. Hard clustering algorithm would not account for these smooth transitions and draw hard boundaries through the cloud of data points. An appropriate clustering algorithm for this dataset would be of fuzzy nature and adaptable to clusters of different shapes and size. As a sanity check, we marked the data points containing DLP signals (Figure 5.13c). In a meaningful and good representation, these data samples should be close to each other, since they contain similar signals. In this case, they are located in the same region but are also a spread out a little. Note that the data points represent 20 min of waveform data and one DLP lasts only a few seconds, since their source location are close to the station (see the map in Figure 5.1). Hence, it is likely that information from other signals dominate the pooling window. The DLP signals are distributed in a limited area in the north-west, corresponding to low-amplitude and high-frequency signals. Remember that these terms are relative to the signal content of the data. If the main signal type are volcanic tremors it is not surprising to describe DLP swarms of low amplitude and high frequency. We can assume that if more and uncataloged DLP events exists, they would be located in the neighborhood of the cataloged events. However, the absence of hard boundaries make it difficult to identify more DLP events.

UMAP captures ever-changing signal characteristics

The two-dimensional PC space is an informative map but limited in details. Higher-order components contain also interesting information, which can potentially discriminate different types of signals or create a hard cluster for the DLP events. We tried to access this information by analysing each component, however, this becomes quickly overwhelming. Figure 5.13b and d show the two dimensional space found by UMAP, revealing very interesting structures. We can identify islands and trajectories of data points, which are sometimes continuous and at other times well separated. If consecutive data points in time are placed close to each other, the signal characteristics are changing slowly. If consecutive data points in time are placed far apart, the signal characteristics are changing rapid or even abruptly. We can see both occurring in the UMAP space but not so well in the PC space. The information of the time axis is strongly encoded within the UMAP space, suggesting that the seismic wavefield is highly non stationary. This delivers a similar picture to the ICA model with a large number of components, which identified many sparse ICs activated at different times (see Figure 5.12). To strengthen our point here, we apply UMAP to the Hamburg dataset presented in Chapter 4 and show that the encoding of the calendar time seems to be something unique to the volcanic environment (see Appendix 5.B of this chapter). The locations of the DLP swarms are limited to an area in the UMAP space, where we could also potentially find similar type of signals (Figure 5.13d). Note that the UMAP results were retrieved with a minimum

distance of 0.2 and 50 neighboring points. UMAP shows similar results with similar conclusions for varying hyperparameters (see Appendix 5.C of this chapter).

5.7 Conclusion

With a scattering network and methods of dimensionality reduction, we analyzed continuous three component seismograms recorded in the vicinity of the Klyuchevskoy volcano. The introduction of this chapter stated three aims which we want to revisit now.

Median vs. maximum pooling

Both pooling operations filter the the data and give a biased but valid view point. Maximum pooling preserves information about short transients with large amplitude and median pooling preserves information about long-lasting signals without being sensitive for the amplitude. In this study, we choose median pooling since we were mainly interested in the information given by the volcanic tremor signals. However, maximum pooling is the better choice if the signals of interests are of short duration with a large amplitude such as volcanic-tectonic earthquakes.

Interpretation of the feature space given by PCA and ICA

PCA finds meaningful and interesting components of the scattering coefficient matrix which can be related to either global characteristics such as broadband energy or specific signal types. The ICA finds similar components to PCA if the number of components is low. For a larger number of components the representation becomes sparser and different time periods were described by different components. This suggests that ICA does not find global characteristics but separates specific patterns in the data on its ICs. The difference we found in ICA and PCA resembles the difference when the methods are applied to facial images (Draper et al., 2003). However, both linear methods PCA and ICA demand a large number of components to capture all interesting information. Thus, interpretation can be overwhelming and it seems more difficult to get a bigger picture of the data. Manifold learning techniques such as UMAP seem to solve this issue by preserving more information on a smaller number of components. The first results look promising but further research needs to be done to exploit its full potential.

Insights about the volcanic complex

All methods (PCA, ICA and UMAP) revealed that the recorded seismogram witness a strongly non-stationary wavefield with ever-changing signal characteristics and no repeating patterns. In particular, time periods with a high detection rate of tremors seem to be highly non-stationary. The volcanic complex seems to undergo a continuous change without going back to its initial state. A comparison to seismic data recorded in an urban environment in the absence of a volcano supports the interpretation that the ever-changing wavefield characteristic is related to the volcanic environment. Our results strengthen the concept that volcanic tremors - which dominate the signal content of our data - should not be considered as a single class or a set of classes. Instead, volcanic tremors change continuously and define a spectrum of signal classes with no hard boundaries. The tremor catalog provides source locations of the tremor signals but only differentiates between two classes of

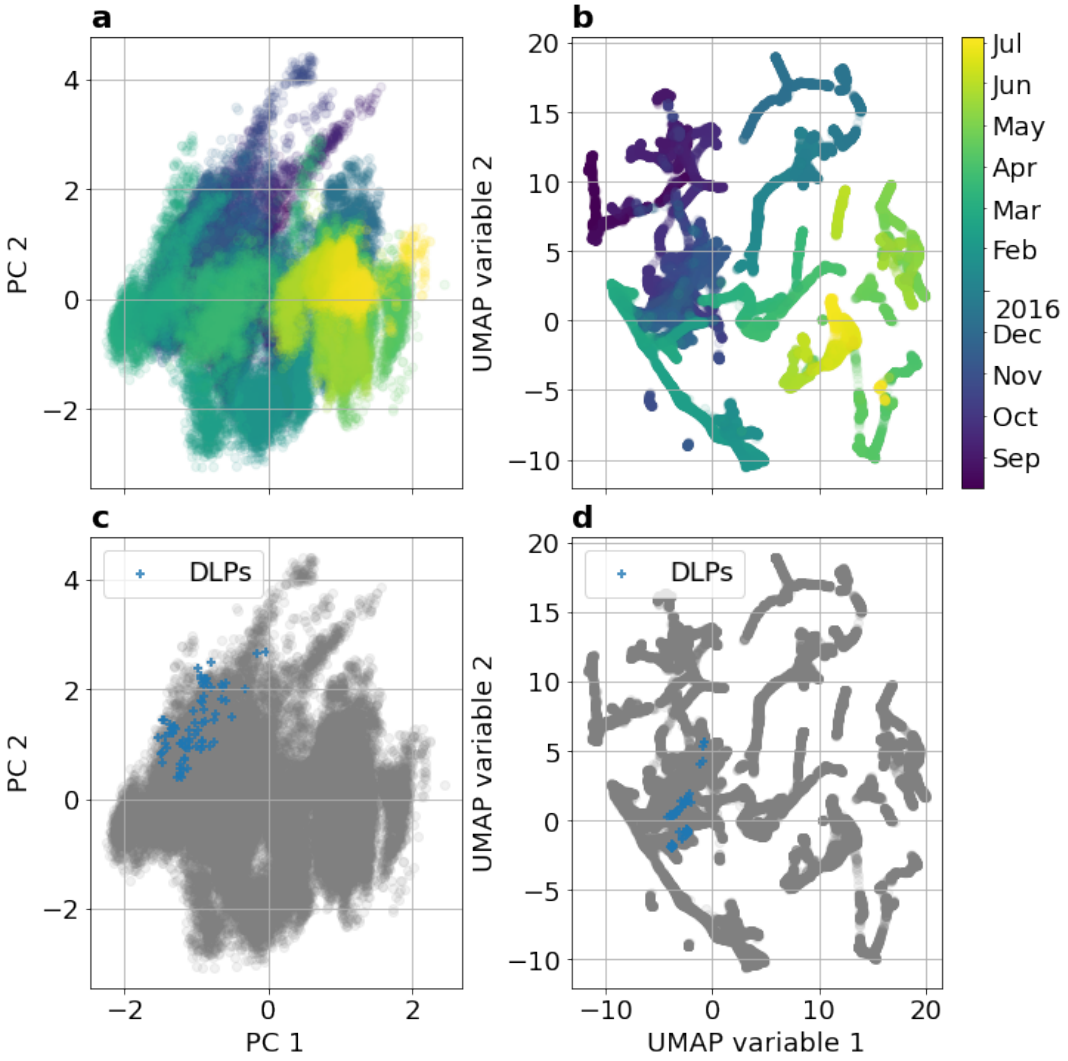

FIGURE 5.13: Two dimensional representation of the scattering coefficients **(a)** PC space based on the first two PCs with color-coded time axis. **(b)** UMAP space for with color-coded time axis. **(c)** PC space with blue crosses marking the data points containing signals from cataloged DLP events. **(d)** UMAP space with blue crosses marking the data points containing signals from cataloged DLP events.

tremors: continuous tremor and intermediate tremor dominated by DLP swarms. Data-driven methods as we present in this study can provide new insights about the varying signal characteristics and compare different tremor episodes in time. This is particular interesting for very large time series, witnessing many different types of tremors. Holtzman et al. (2018) clustered seismic events related to geothermal activity and found a cyclic behaviour in the spectral properties of these events, indicating cyclic phases in the Geysers geothermal fields. We did not identify any cyclic behaviour in the volcanic environment of Klyuchevskoy but larger time series might reveal re-occurring patterns as it is the case for the Geyser.

Appendices

5.A Example spectrograms representing third PC

5.B UMAP applied to station WM01 in Hamburg, Germany

The ICA-analysis and UMAP applied to the data of station SV13 show an ever-changing wavefield with no repeating patterns. To strengthen that point, and the utility of UMAP, we apply UMAP with the same parameters to the scattering coefficients retrieved from station WM01 in Hamburg, Germany (see Chapter 4). Figure 5.B.1a shows the two-dimensional UMAP space with colors indicating the calendar time. Data points with different calendar times are overlapping mostly and, thus, the calendar time is not encoded as it is the case for SV13 (see Figure 5.13 for comparison).

We did not expect a strong encoding of the calendar time in the data, since the continuous freezing and thawing process and the anthropogenic activity were the main identified patterns in the data (see Chapter 4). It is very likely that we find these patterns also in this UMAP representation. Indeed, the freezing process is visible when we color-code the UMAP space with the temperature recorded at 5 cm depth (see Figure 5.B.1b). The freezing process indicated by temperature close to 0 °C does not create a separated cluster but it places data points on unused areas which are connected to the data points color-coded with warmer temperatures. This is expected since the freezing and thawing process is continuous, altering also the wavefield in a slow and continuous fashion. The main characteristic of the dataset becomes clear if we color-code the data points with the hour of the day (Figure 5.B.2a-c). We can identify four clusters characterized by different hours of the day. A small cluster located in the south-east (Figure 5.B.2a) is clearly showing the effect of the taper applied to the daily data streams (Figure 5.B.2b). The largest cluster, located in the east, can be associated with ambient seismic noise recorded mostly at nighttime (Figure 5.B.2c) and during weekends (Figure 5.B.2d). This cluster transitions towards a smaller cluster at the center of the UMAP space associated with daytime activity ranging continuously from 5:00 to 16:00 local time (Figure 5.B.2c). Further west, we have the last cluster of the four identified cluster, which is also related to daytime activity but with clear pauses around 9:00 in the morning and noon. With this information we now understand better the temperature imprint shown in Figure 5.B.1b: the freezing and thawing process modulates the dominating anthropogenic seismic imprint in the data. In fact, this confirms exactly our interpretation of the hierarchical clustering result in Chapter 4. Remember that we found subclusters (cluster D and subcluster B.1) related to the freezing process within the large ambient seismic noise cluster (cluster D+E) and the smaller urban activity cluster (cluster B). Cluster A, associated with labour activity, showed exactly the same pauses in the morning hours and at noon as the far-west cluster in the UMAP representation.

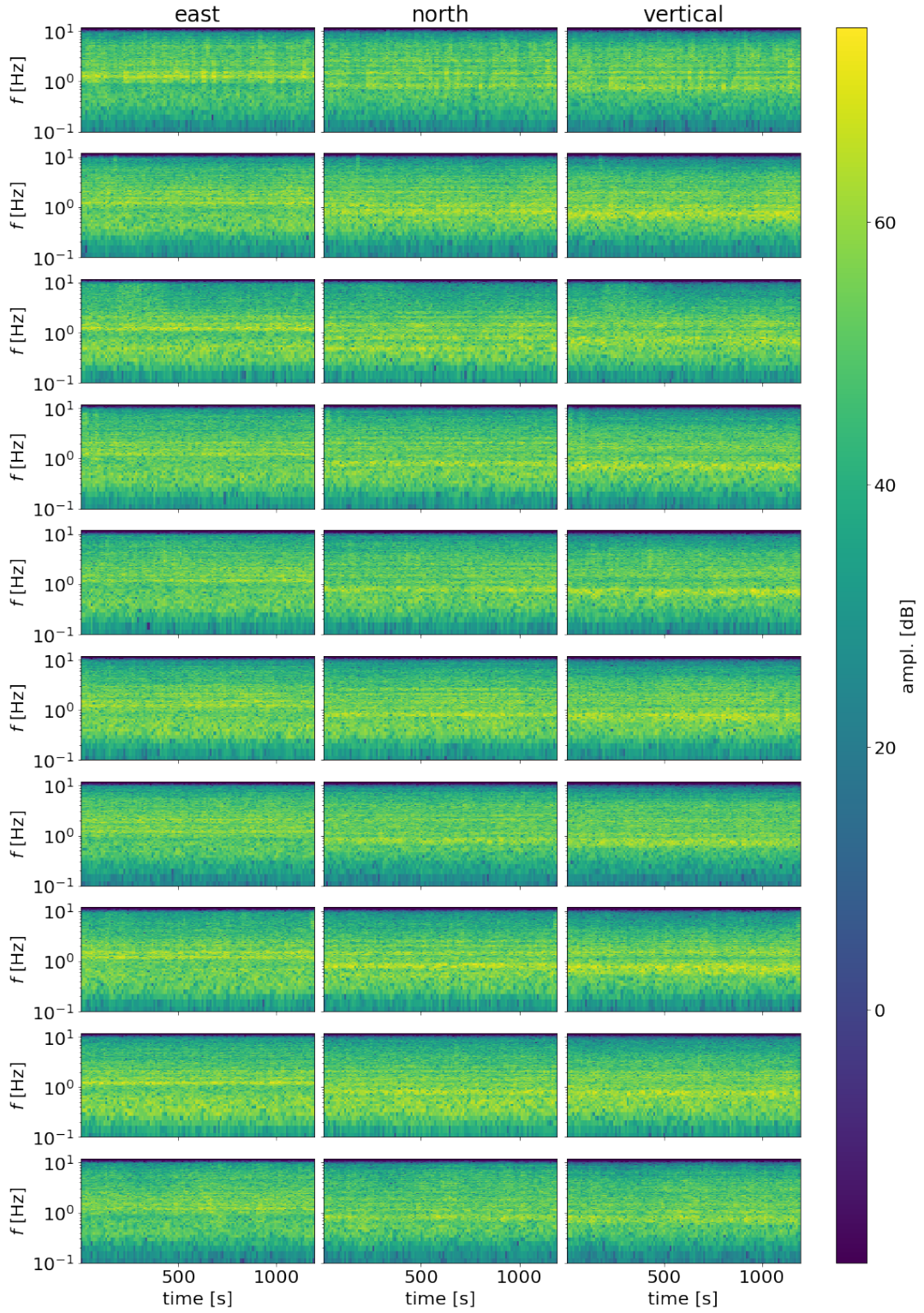

FIGURE 5.A.1: Spectrograms of the three-component seismograms of data points corresponding to large positive values on the third PC

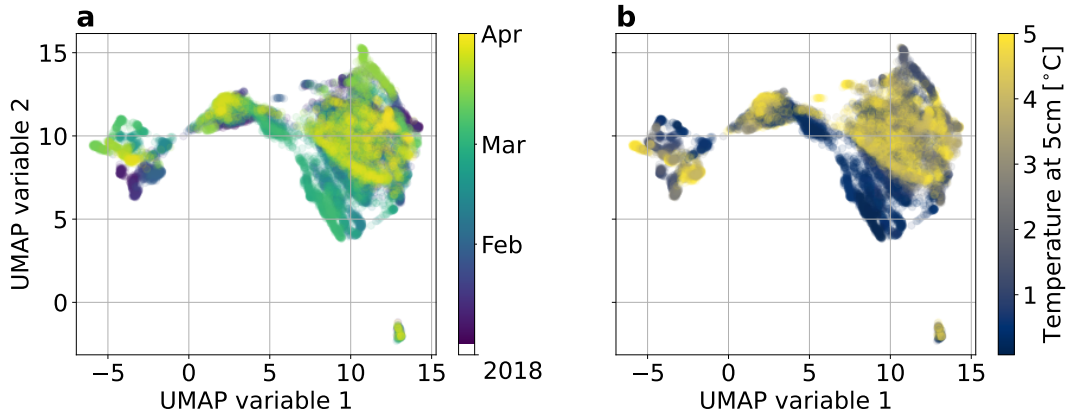

FIGURE 5.B.1: UMAP results for the scattering coefficient matrix of station WM01 in Hamburg, Germany (see Chapter 4). In **(a)** the UMAP results are color-coded according to the calendar time of each data point and in **(b)** the UMAP results are color-coded according to the temperature measured at 5 cm depth. Both the temperature time series and the scattering coefficients matrix have the same sampling rate of 10 min

This short analysis strengthens our interpretation that the ever-changing wavefield is a unique characteristic of the seismic data recorded in the vicinity of a volcano. Moreover, the results and interpretation of UMAP align with the results and interpretation of the hierarchical waveform clustering approach, confirming the utility of this manifold learning technique.

5.C Hyperparameter test for UMAP

5.C.1 Case 1: SV13, Kamchatka, Russia

As pointed out in Chapter 2, the number of neighboring data points and the minimum distance for putting neighboring points together are the main hyperparameters which control outcome of UMAP. Since we are interested in a visualization of the high dimensional scattering coefficient matrix, we only retrieve two UMAP variables. Figure 5.C.1 shows the results of UMAP for 9 different combinations of the minimum distance and number of neighbors. The data cloud comes in different shapes with regard to the hyperparameters, but all the examples confirm that the time axis seems to be encoded in the seismic data. We can confirm that a larger minimum distance results in a larger and less clustered data cloud. While a small number of neighbors seems to produce a round and unstructured data cloud, a larger number of neighbors results in some elongated shapes and more complex patterns.

5.C.2 Case 2: WM01, Hamburg, Germany

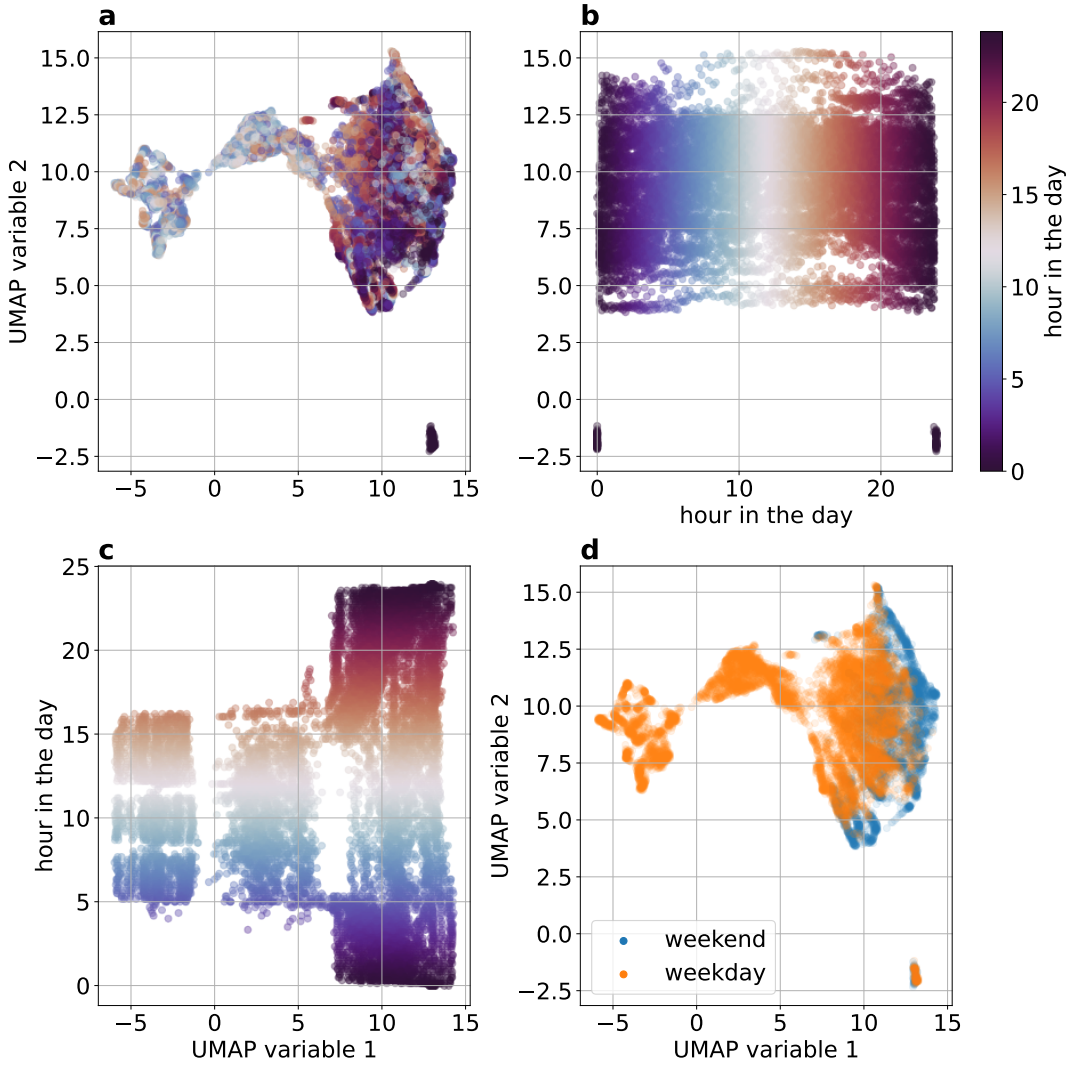

FIGURE 5.B.2: UMAP results for the scattering coefficient matrix of station WM01 in Hamburg, Germany (see Chapter 4). **(a)** shows the two UMAP variables color-coded with the time of the day. **(b)** shows the time of the day against the second UMAP variable, color-coded with the time of the day. **(c)** shows the first variable against the time of the day, color-coded with the time of the day. **(a)** shows the two UMAP variables with a color-code indicating weekend (Sa.-So.) and weekday (Mo.-Fr.).

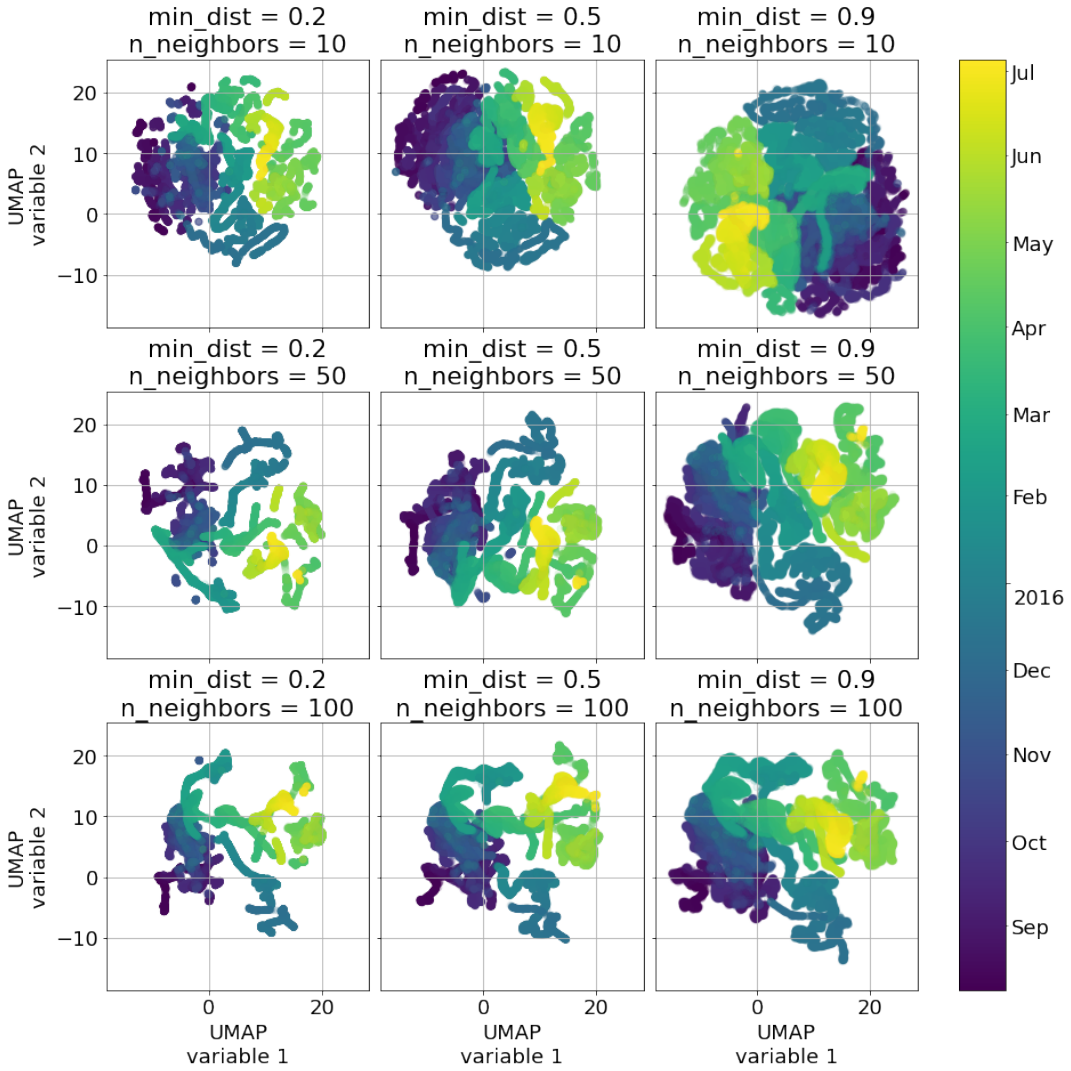

FIGURE 5.C.1: Different UMAP results with changing hyperparameters for the data recorded at SV13. The results shown in Figure 5.13 correspond to min_dist = 0.2 and n_neighbors = 50.

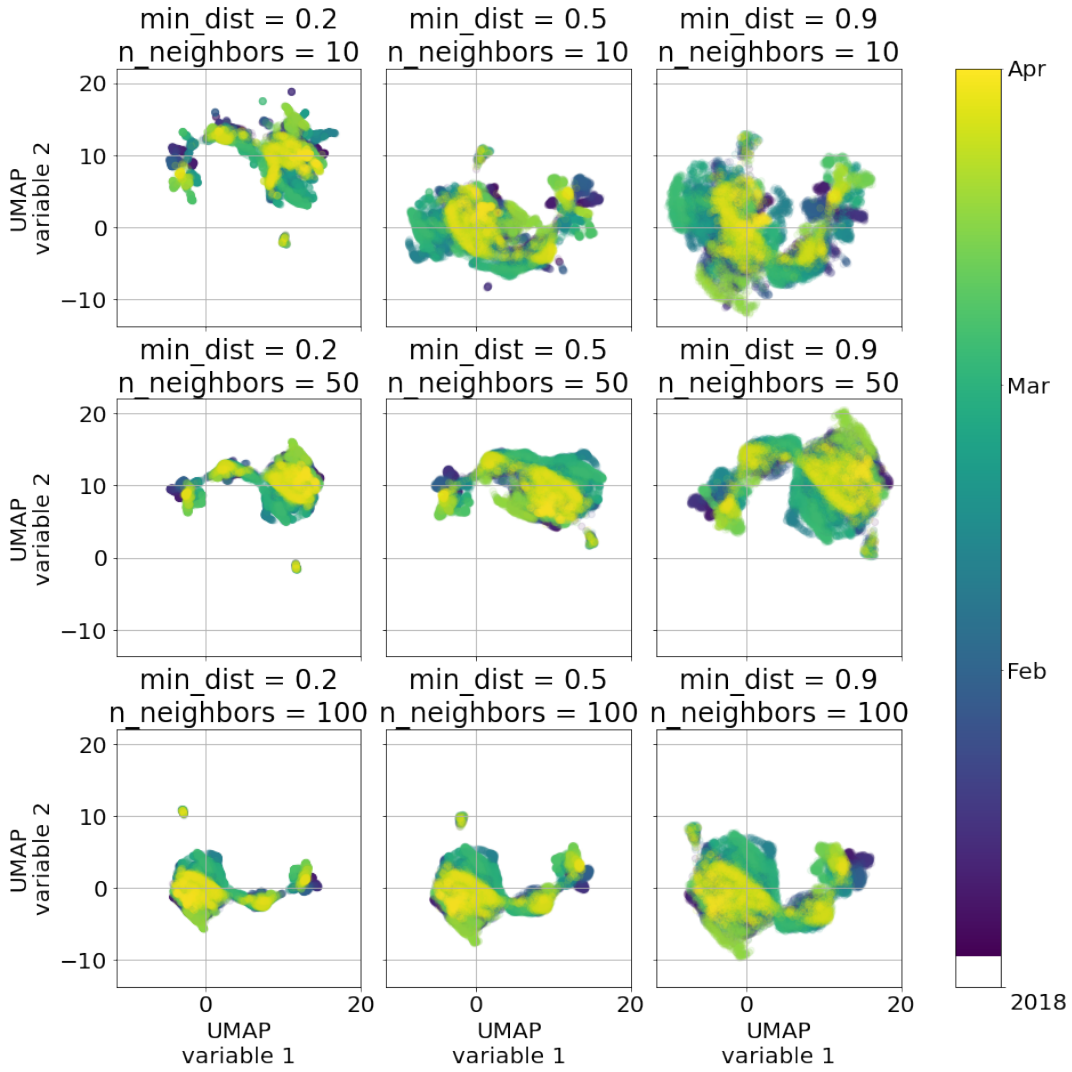

FIGURE 5.C.2: Different UMAP results with changing hyperparameters for the data recorded at WM01, Hamburg, Germany. The results shown in Figure 5.B.1 and 5.B.2 correspond to min_dist = 0.2 and n_neighbors = 50.

Chapter 6

Conclusion and outlook

6.1 Conclusion

With hierarchical waveform clustering we present a strategy for data-driven exploration of continuous seismograms, aiming at the discovery of overlooked patterns in seismic time series. The applications focused on three component seismograms recorded by a single seismic station. The scattering transform proved to be an interesting and meaningful representation of continuous seismograms in regard to pattern recognition tasks such as clustering or matrix factorization. The scattering coefficients offer interpretability while still delivering a richer representation than spectral coefficients based on the Fourier transform. Nevertheless, the second- and higher-order coefficients can challenge the interpretability. The design of the scattering network is intuitive and has to be adapted to the task at hand. For instance, the design of the wavelets is guided by the potential targeted signals in the time series. Another important choice is the pooling operation which in one way or the other filters information and favors certain signal characteristics. With PCA and ICA, we are able to retrieve interesting and meaningful patterns in the time series of scattering coefficients, which can be either directly used for interpretation or other pattern recognition tasks such as clustering. While PCA identifies global signal characteristics, ICA finds sparse components with the potential to separate different seismic source patterns. In Chapter 4 we even identified blindly a medium change and were able to isolate the seismic signature of freezing and thawing from the urban seismic activity. These temporal data-driven patterns describe also the stationarity of the recorded wavefield. The application shown in Chapter 5 challenges the concept of volcanic tremors as one signal class and confirms the ever-changing nature of these signals, holding information about the current state of the volcanic complex. Besides interpreting component by component, hierarchical clustering offers a data-driven way to explore the complete low dimensional PC or IC space. Chapter 3 and 4 proofed that the dendrogram offers a meaningful hierarchical overview of all the recorded signals in the dataset. Within the dendrogram we could identify clusters of similar low-magnitude seismic burst events belonging to the larger cluster of general seismicity. The presented strategy is in stark contrast to most other unsupervised learning approaches which aim at finding a fixed number of seismic signal clusters by optimizing an objective function (such as in Snover et al., 2020; Seydoux et al., 2020). Instead, the main goal of our approach is data exploration which includes the exploration and interpretation of multiple possible solutions by an expert. A similar conclusion was drawn in Köhler, Ohrnberger, and Scherbaum (2010): "Manual inspection of suggested clustering solutions and interpretation based on expert knowledge is an integral part of this approach. Selecting automatically the number of clusters based on a validity measure for example, may not tap the full discrimination potential of the approach."

6.2 Outlook

Hunting for tectonic signals in plate boundary observatories

The proposed strategy could help to identify new and overseen patterns in many different types of seismic experiments. However, one very interesting application considers continuous seismograms from borehole stations within plate boundary observatories. Rouet-Leduc, Hulbert, and Johnson (2019) observed a continuous tremor-like chatter in the ambient seismic wave field recorded close to the Cascadian subduction zone. Our approach might identify this pattern blindly and would strengthen the assumption that there is valuable information about tectonic processes in the ambient seismic wave field. The approach could also aid in the hunt for tremors at the NAF with less constraints on the tremor's signal characteristics. Recently, Ben-Zion et al. (2022) called for the next generation of plate boundary observatories at different sites in the world in order to better understand earthquake processes and their nucleation phase. Pattern recognition algorithms such as hierarchical waveform clustering could aid in the search of potentially new tectonic signals recorded by these observatories.

Another interesting application would follow the idea presented in Holtzman et al. (2018), which identified a cyclic behaviour of the Geyser due to the signal properties of the cataloged events. Recent developments in earthquake detection and location with deep learning create large catalogs which might be best analyzed and explored with unsupervised learning methods (Beroza, Segou, and Mostafa Mousavi, 2021). A UMAP representation retrieved from the scattering coefficients of the earthquake waveform data might reveal new insights about the earthquake processes, which are not necessarily encoded in the spatial distribution of the earthquakes.

Towards data-driven descriptors for tremor signals and understanding their underlying mechanism

With UMAP and ICA we are able to estimate how stationary a wavefield is at a given time in a data-driven fashion. Until now, we know that something is changing in the seismogram on a certain timescale. However, we did not yet identify what is exactly the change and this would be an important step towards understanding the underlying mechanism for the signal change. Clustering the UMAP space could help to identify common signal characteristics and connect them to certain areas in the UMAP space.

Towards detecting medium changes directly from continuous seismograms

Chapter 4 showed the potential of detecting medium changes directly in continuous seismograms. It seems possible that other medium changes such as groundwater fluctuations or perhaps even healing processes after large Earthquakes modulate the ambient seismic wavefield as we have seen for the freezing and thawing process. Accessing this information directly from the continuous seismograms could give new insights about the medium changes and how they affect seismic wave propagation. This would enhance monitoring strategies based on a single station.

Enhancing the strategy with manifold learning and hierarchical spectral clustering

Hierarchical waveform clustering is based on relatively simple and easy to interpret tools, which have been studied and utilized since multiple decades. Our work showed that PCA and ICA capture interesting patterns of the seismograms. However, due to their linear mapping many components are needed to describe the interesting structure of the data, and this poses difficulties for visualization and interpretation. Moreover, if too many components are retained, clustering algorithms will again struggle with the curse of dimensionality. The applications of UMAP showed that manifold learning techniques seem to capture a lot of interesting structure in only two dimensions. The two-dimensional UMAP space resembles a map for navigating through the individual worlds of each dataset, revealing different islands of seismic signals and their potential connection. In this thesis, we applied UMAP mainly for a visual exploration of the dataset and we did not provide an extensive study on the hyperparameters of UMAP. We believe that UMAP holds a large potential for exploration of large seismic time series and recent examples have even shown successful application of clustering time series data in UMAP spaces (e.g. Ali et al., 2019).

Besides using non linear ways to reduce the dimensions, other more advanced clustering algorithms might offer new insights, too. We utilized hierarchical clustering with the Ward's method in order to explore the data with the dendrogram. However, it seems that this approach might not be the best suited method for seismological time series. As pointed out in Chapter 2, hierarchical clustering with the Ward's method tends to find even-sized ball-shaped clusters. Seismic time series often contain large class imbalances such as between the class of Earthquakes and the ambient seismic wave field. This results in a large point cloud with distributed outliers in the two dimensional PC representation (see for example Figure 5.13a in Chapter 5) or sparse independent components (see for example Figure 3.B.2 in Chapter 3). Density-based hierarchical clustering (HDBSCAN) is an extension of hierarchical clustering, which is able to identify clusters of various shape and size, while still offering a dendrogram for data exploration. Recent applications have also shown that HDBSCAN performs well in the UMAP space, delivering new insights about the data (Herrmann et al., 2022).

From single station to array data

An extension to array data is another interesting and challenging future research direction. Adding more stations poses questions about how to organize the data. Instead of creating a scattering coefficient matrix, it might be more suitable to store the data as a tensor. Thus, matrix factorization methods such as PCA or ICA would be replaced by tensor decompositions, which take more computer resources and are often more difficult to solve. Similarly, hierarchical waveform clustering could be also adapted to the large amount of data generated by distributed acoustic sensing, which has the potential to reveal new types of seismicity (Klaasen et al., 2021).

Bibliography

Aggarwal, Charu C, Alexander Hinneburg, and Daniel A Keim (2001). "On the surprising behavior of distance metrics in high dimensional space". In: *International conference on database theory*. Springer, pp. 420–434.

Ali, Mohammed et al. (2019). "TimeCluster: dimension reduction applied to temporal data for visual analytics". In: *The Visual Computer* 35.6, pp. 1013–1026.

Allen, Rex V (1978). "Automatic earthquake recognition and timing from single traces". In: *Bulletin of the seismological society of America* 68.5, pp. 1521–1532.

Andén, Joakim and Stéphane Mallat (2014). "Deep scattering spectrum". In: *IEEE Transactions on Signal Processing* 62.16, pp. 4114–4128.

Baker, Frank B (1974). "Stability of two hierarchical grouping techniques case I: sensitivity to data errors". In: *Journal of the American Statistical Association* 69.346, pp. 440–445.

Balestriero, Randall et al. (2018). "Spline filters for end-to-end deep learning". In: *International Conference on Machine Learning*. PMLR, pp. 364–373.

Beaucé, Eric et al. (2019). "Systematic detection of clustered seismicity beneath the Southwestern Alps". In: *Journal of Geophysical Research: Solid Earth* 124.11, pp. 11531–11548.

Becht, Etienne et al. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1, pp. 38–44.

Bellman, Richard (1966). "Dynamic programming". In: *Science* 153.3731, pp. 34–37.

Ben-Zion, Yehuda et al. (2022). "A Grand Challenge International Infrastructure for Earthquake Science". In: *Seismological Research Letters*.

Beroza, Gregory C, Margarita Segou, and S Mostafa Mousavi (2021). "Machine learning and earthquake forecasting—next steps". In: *Nature communications* 12.1, pp. 1–3.

Beyreuther, Moritz et al. (2010). "ObsPy: A Python toolbox for seismology". In: *Seismological Research Letters* 81.3, pp. 530–533. DOI: https://doi.org/10.1785/gssrl.81.3.530.

Bocchini, GM et al. (2021). "Does deep tectonic tremor occur in the central-eastern Mediterranean basin?" In: *Journal of Geophysical Research: Solid Earth* 126.1, 2020JB020448.

Bonnefoy-Claudet, Sylvette, Fabrice Cotton, and Pierre-Yves Bard (2006). "The nature of noise wavefield and its applications for site effects studies: A literature review". In: *Earth-Science Reviews* 79.3-4, pp. 205–227.

Bruna, Joan and Stéphane Mallat (2013). "Invariant scattering convolution networks". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1872–1886.

Chami, Ines et al. (2020). "From trees to continuous embeddings and back: Hyperbolic hierarchical clustering". In: *arXiv preprint arXiv:2010.00402*.

Cheng, Feng et al. (2022). "Watching the Cryosphere Thaw: Seismic Monitoring of Permafrost Degradation Using Distributed Acoustic Sensing During a Controlled Heating Experiment". In: *Geophysical Research Letters* 49.10, e2021GL097195. DOI: https://doi.org/10.1029/2021GL097195.

Chouet, Bernard A (1996). "Long-period volcano seismicity: its source and use in eruption forecasting". In: *Nature* 380.6572, pp. 309–316.

Ciaramella, A et al. (2004). "Characterization of Strombolian events by using independent component analysis". In: *Nonlinear Processes in Geophysics* 11.4, pp. 453–461.

Comon, Pierre (1994). "Independent component analysis, a new concept?" In: *Signal processing* 36.3, pp. 287–314.

Cosentino, Romain and Behnaam Aazhang (2020). "Learnable group transform for time-series". In: *International Conference on Machine Learning*. PMLR, pp. 2164–2173.

DANA (2012). *Dense Array for North Anatolia*. DOI: 10.7914/SN/YH_2012. URL: http://www.fdsn.org/doi/10.7914/SN/YH_2012.

Delac, Kresimir, Mislav Grgic, and Sonja Grgic (2005). "Independent comparative study of PCA, ICA, and LDA on the FERET data set". In: *International Journal of Imaging Systems and Technology* 15.5, pp. 252–260.

Deparis, J et al. (2008). "Analysis of rock-fall and rock-fall avalanche seismograms in the French Alps". In: *Bulletin of the Seismological Society of America* 98.4, pp. 1781–1796.

Dewey, James and Perry Byerly (1969). "The early history of seismometry (to 1900)". In: *Bulletin of the Seismological Society of America* 59.1, pp. 183–227.

Diaz, Jordi (2020). "Church bells and ground motions". In: *Journal of Seismology*, pp. 1–10.

Domingos, Pedro (2012). "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10, pp. 78–87.

Draper, Bruce A et al. (2003). "Recognizing faces with PCA and ICA". In: *Computer vision and image understanding* 91.1-2, pp. 115–137.

Dumoulin, Vincent and Francesco Visin (2016). "A guide to convolution arithmetic for deep learning". In: *arXiv preprint arXiv:1603.07285*.

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

Emre, Ö et al. (2011). "1: 250,000 scale active fault map series of Turkey". In: *Kayseri (NJ36-8) Quadrange, Ankara*.

Estivill-Castro, Vladimir (2002). "Why so many clustering algorithms: a position paper". In: *ACM SIGKDD explorations newsletter* 4.1, pp. 65–75.

Ezugwu, Absalom E et al. (2022). "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects". In: *Engineering Applications of Artificial Intelligence* 110, p. 104743.

Fehler, Michael (1983). "Observations of volcanic tremor at Mount St. Helens volcano". In: *Journal of Geophysical Research: Solid Earth* 88.B4, pp. 3476–3484.

García-Jerez, Antonio et al. (2016). "A computer code for forward calculation and inversion of the H/V spectral ratio under the diffuse field assumption". In: *Computers & geosciences* 97, pp. 67–78.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Guéguen, Philippe et al. (2017). "How sensitive are site effects and building response to extreme cold temperature? The case of the Grenoble's (France) City Hall building". In: *Bulletin of earthquake engineering* 15.3, pp. 889–906.

Herrmann, Moritz et al. (2022). "Enhancing cluster analysis via topological manifold learning". In: *arXiv preprint arXiv:2207.00510*.

Holtzman, Benjamin K et al. (2018). "Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field". In: *Science advances* 4.5, eaao2929.

Hotovec, Alicia J et al. (2013). "Strongly gliding harmonic tremor during the 2009 eruption of Redoubt Volcano". In: *Journal of Volcanology and Geothermal Research* 259, pp. 89–99.

Hutchison, Alexandra A and Abhijit Ghosh (2017). "Ambient tectonic tremor in the san jacinto fault, near the anza gap, detected by multiple mini seismic arraysambient tectonic tremor in the san jacinto fault, near the anza gap, detected by multiple mini seismic arrays". In: *Bulletin of the Seismological Society of America* 107.5, pp. 1985–1993.

Hyvärinen, Aapo and Erkki Oja (2000). "Independent component analysis: algorithms and applications". In: *Neural networks* 13.4-5, pp. 411–430.

Hyvärinen, Aapo et al. (2010). "Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis". In: *NeuroImage* 49.1, pp. 257–271.

Ide, Satoshi et al. (2007). "A scaling law for slow earthquakes". In: *Nature* 447.7140, pp. 76–79.

Inbal, Asaf et al. (2018). "Sources of long-range anthropogenic noise in Southern California and implications for tectonic tremor detection". In: *Bulletin of the Seismological Society of America* 108.6, pp. 3511–3527.

Isacks, Bryan, Jack Oliver, and Lynn R Sykes (1968). "Seismology and the new global tectonics". In: *Journal of geophysical research* 73.18, pp. 5855–5899.

Jain, Anil K (2010). "Data clustering: 50 years beyond K-means". In: *Pattern recognition letters* 31.8, pp. 651–666.

Jain, Anil K, M Narasimha Murty, and Patrick J Flynn (1999). "Data clustering: a review". In: *ACM computing surveys (CSUR)* 31.3, pp. 264–323.

James, SR et al. (2019). "Insights into permafrost and seasonal active-layer dynamics from ambient seismic noise monitoring". In: *Journal of Geophysical Research: Earth Surface* 124.7, pp. 1798–1816.

Jenkins, William F et al. (2021). "Unsupervised deep clustering of seismic data: Monitoring the Ross Ice Shelf, Antarctica". In: *Journal of Geophysical Research: Solid Earth*, e2021JB021716.

Johnson, Christopher W et al. (2020). "Identifying different classes of seismic noise signals using unsupervised learning". In: *Geophysical Research Letters* 47.15, e2020GL088353.

Johnson, Stephen C (1967). "Hierarchical clustering schemes". In: *Psychometrika* 32.3, pp. 241–254.

Joswig, Manfred (1990). "Pattern recognition for earthquake detection". In: *Bulletin of the Seismological Society of America* 80.1, pp. 170–186.

Journeau, Cyril et al. (2022). "Seismic tremor reveals active trans-crustal magmatic system beneath Kamchatka volcanoes". In: *Science advances* 8.5, eabj1571.

Julian, Bruce R (1994). "Volcanic tremor: Nonlinear excitation by fluid flow". In: *Journal of Geophysical Research: Solid Earth* 99.B6, pp. 11859–11877.

Jung, Tzyy-Ping et al. (2000). "Removing electroencephalographic artifacts by blind source separation". In: *Psychophysiology* 37.2, pp. 163–178.

Kaufman, Leonard and Peter J Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kennett, BLN and ER Engdahl (1991). "Traveltimes for global earthquake location and phase identification". In: *Geophysical Journal International* 105.2, pp. 429–465.

Klaasen, Sara et al. (2021). "Distributed acoustic sensing in volcano-glacial environments—Mount Meager, British Columbia". In: *Journal of Geophysical Research: Solid Earth* 126.11.

Köhler, Andreas, Matthias Ohrnberger, and Frank Scherbaum (2010). "Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps". In: *Geophysical Journal International* 182.3, pp. 1619–1630.

Köhler, Andreas and Christian Weidle (2019). "Potentials and pitfalls of permafrost active layer monitoring using the HVSR method: a case study in Svalbard". In: *Earth Surface Dynamics* 7.1, pp. 1–16.

Konstantinou, Konstantinos I and Vera Schlindwein (2003). "Nature, wavefield properties and source mechanism of volcanic tremor: a review". In: *Journal of Volcanology and Geothermal Research* 119.1-4, pp. 161–187.

Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek (2009). "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering". In: *Acm transactions on knowledge discovery from data (tkdd)* 3.1, pp. 1–58.

Kuiper, F Kent and Lloyd Fisher (1975). "391: A Monte Carlo comparison of six clustering procedures". In: *Biometrics*, pp. 777–783.

Lacroix, Pascal and Agnes Helmstetter (2011). "Location of seismic signals associated with microearthquakes and rockfalls on the Séchilienne landslide, French Alps". In: *Bulletin of the Seismological Society of America* 101.1, pp. 341–353.

Larose, Eric et al. (2015). "Environmental seismology: What can we learn on earth surface processes with ambient noise?" In: *Journal of Applied Geophysics* 116, pp. 62–74.

Latter, John H (1979). "Volcanological observations at Tongariro National Park. II: Types and classification of volcanic earthquakes, 1976-1978". In: *Report-Geophysics Division* 150.

Lay, Thorne and Terry C Wallace (1995). *Modern global seismology*. Elsevier.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.

Lindner, Fabian, Joachim Wassermann, and Heiner Igel (2021). "Seasonal freeze-thaw cycles and permafrost degradation on Mt. Zugspitze (German/Austrian Alps) revealed by single-station seismic monitoring". In: *Earth and Space Science Open Archive ESSOAr*.

Maaten, Laurens Van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11.

Maggi, Alessia et al. (2017). "Implementation of a multistation approach for automated event classification at Piton de la Fournaise volcano". In: *Seismological Research Letters* 88.3, pp. 878–891.

Malfante, Marielle et al. (2018). "Automatic classification of volcano seismic signatures". In: *Journal of Geophysical Research: Solid Earth* 123.12, pp. 10–645.

McInnes, Leland, John Healy, and James Melville (2018). "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426*.

Meng, Haoran and Yehuda Ben-Zion (2018). "Characteristics of airplanes and helicopters recorded by a dense seismic array near Anza California". In: *Journal of Geophysical Research: Solid Earth* 123.6, pp. 4783–4797.

Met Office (2010 - 2015). *Cartopy: a cartographic python library with a matplotlib interface*. Exeter, Devon. URL: http://scitools.org.uk/cartopy.

Miao, Y et al. (2019). "Influence of Seasonal Frozen Soil on Near-surface Shear Wave Velocity in Eastern Hokkaido, Japan". In: *Geophysical Research Letters* 46.16, pp. 9497–9508.

Milligan, Glenn W and Paul D Isaac (1980). "The validation of four ultrametric clustering algorithms". In: *Pattern Recognition* 12.2, pp. 41–50.

Mousavi, S Mostafa and Gregory C Beroza (2022). "Deep-learning seismology". In: *Science* 377.6607, eabm4470.

Mousavi, S Mostafa et al. (2019). "Unsupervised clustering of seismic signals using deep convolutional autoencoders". In: *IEEE Geoscience and Remote Sensing Letters* 16.11, pp. 1693–1697.

Müllner, Daniel (2011). "Modern hierarchical, agglomerative clustering algorithms". In: *arXiv preprint arXiv:1109.2378*.

— (2013). "fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python". In: *Journal of Statistical Software* 53, pp. 1–18.

Nadeau, Robert M and David Dolenc (2005). "Nonvolcanic tremors deep beneath the San Andreas Fault". In: *Science* 307.5708, pp. 389–389.

Obara, Kazushige (2002). "Nonvolcanic deep tremor associated with subduction in southwest Japan". In: *Science* 296.5573, pp. 1679–1681.

Oliver, Jack and Leonard Murphy (1971). "WWNSS: Seismology's Global Network of Observing Stations: Standardized collection and efficient distribution of earthquake data yield social and scientific rewards." In: *Science* 174.4006, pp. 254–261.

Outcalt, Samuel I, Frederick E Nelson, and Kenneth M Hinkel (1990). "The zero-curtain effect: Heat and mass transfer across an isothermal region in freezing soil". In: *Water Resources Research* 26.7, pp. 1509–1516.

O'Neill, Adam and Toshifumi Matsuoka (2005). "Dominant higher surface-wave modes and possible inversion pitfalls". In: *Journal of Environmental & Engineering Geophysics* 10.2, pp. 185–201.

Peddinti, Vijayaditya et al. (2014). "Deep scattering spectrum with deep neural networks". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 210–214.

Pedregosa, F. et al. (2011a). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. DOI: 10.5555/1953048.2078195.

Pedregosa, Fabian et al. (2011b). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Pfohl, Anna et al. (2015). "Search for tectonic tremor on the Central North Anatolian fault, Turkey". In: *Bulletin of the Seismological Society of America* 105.3, pp. 1779–1786.

Piña-Flores, José et al. (2016). "The inversion of spectral ratio H/V in a layered system using the diffuse field assumption (DFA)". In: *Geophysical Journal International*, ggw416.

Podolskiy, Evgeny A and Fabian Walter (2016). "Cryoseismology". In: *Reviews of geophysics* 54.4, pp. 708–758.

Poyraz, Selda Altuncu et al. (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". In: *Tectonophysics* 656, pp. 190–201.

Riahi, Nima and Peter Gerstoft (2015). "The seismic traffic footprint: Tracking trains, aircraft, and cars seismically". In: *Geophysical Research Letters* 42.8, pp. 2674–2681.

Ross, Zachary E et al. (2018). "Generalized seismic phase detection with deep learning". In: *Bulletin of the Seismological Society of America* 108.5A, pp. 2894–2901.

Rouet-Leduc, Bertrand, Claudia Hulbert, and Paul A Johnson (2019). "Continuous chatter of the Cascadia subduction zone revealed by machine learning". In: *Nature Geoscience* 12.1, pp. 75–79.

Rouet-Leduc, Bertrand et al. (2020). "Probing slow earthquakes with deep learning". In: *Geophysical research letters* 47.4, e2019GL085870.

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

Rubinstein, Justin L, David R Shelly, and William L Ellsworth (2009). "Non-volcanic tremor: A window into the roots of fault zones". In: *New Frontiers in Integrated Solid Earth Sciences*. Springer, pp. 287–314.

Sánchez-Sesma, Francisco J et al. (2011). "A theory for microtremor H/V spectral ratio: application for a layered medium". In: *Geophysical Journal International* 186.1, pp. 221–225.

Sens-Schönfelder, Christoph and Ulrich Wegler (2006). "Passive image interferometry and seasonal variations of seismic velocities at Merapi Volcano, Indonesia". In: *Geophysical research letters* 33.21.

Seydoux, Léonard et al. (2020). "Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning". In: *Nature communications* 11.1, pp. 1–12.

Shapiro, Nikolai M et al. (2017). "Understanding Kamchatka's extraordinary volcano cluster". In: *Eos, Transactions American Geophysical Union* 98.

Shelly, David R, Gregory C Beroza, and Satoshi Ide (2007). "Non-volcanic tremor and low-frequency earthquake swarms". In: *Nature* 446.7133, pp. 305–307.

Sick, Benjamin, Matthias Guggenmos, and Manfred Joswig (2015). "Chances and limits of single-station seismic event clustering by unsupervised pattern recognition". In: *Geophysical Journal International* 201.3, pp. 1801–1813.

Snover, Dylan et al. (2020). "Deep Clustering to Identify Sources of Urban Seismic Noise in Long Beach, California". In: *Seismological Research Letters*.

Soubestre, J et al. (2019). "Depth migration of seismovolcanic tremor sources below the Klyuchevskoy volcanic group (Kamchatka) determined from a network-based analysis". In: *Geophysical Research Letters* 46.14, pp. 8018–8030.

Soubestre, Jean et al. (2018). "Network-based detection and classification of seismovolcanic tremors: Example from the Klyuchevskoy volcanic group in Kamchatka". In: *Journal of Geophysical Research: Solid Earth* 123.1, pp. 564–582.

Steinmann, René, Céline Hadziioannou, and Eric Larose (2021). "Effect of centimetric freezing of the near subsurface on Rayleigh and Love wave velocity in ambient seismic noise correlations". In: *Geophysical Journal International* 224.1, pp. 626–636.

Steinmann, René, Léonard Seydoux, and Michel Campillo (2022). "AI-Based Unmixing of Medium and Source Signatures From Seismograms: Ground Freezing Patterns". In: *Geophysical Research Letters* 49.15, e2022GL098854.

Steinmann, Rene et al. (2022). "Hierarchical exploration of continuous seismograms with unsupervised learning". In: *Journal of Geophysical Research: Solid Earth* 127.1, e2021JB022455.

Steinmannn, Rene, Celine Hadziioannou, and Eric Larose (Aug. 2020). *Data of seismic urban noise in the city of Hamburg, Germany 2018*. DOI: 10.5281/zenodo.3992631. URL: https://doi.org/10.5281/zenodo.3992631.

Turk, Matthew A and Alex P Pentland (1991). "Face recognition using eigenfaces". In: *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*. IEEE Computer Society, pp. 586–587.

Unglert, K and AM Jellinek (2017). "Feasibility study of spectral pattern recognition reveals distinct classes of volcanic tremor". In: *Journal of Volcanology and Geothermal Research* 336, pp. 219–244.

Unglert, Kathi, Valentina Radić, and A Mark Jellinek (2016). "Principal component analysis vs. self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra". In: *Journal of Volcanology and Geothermal Research* 320, pp. 58–74.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Ward Jr, Joe H (1963). "Hierarchical grouping to optimize an objective function". In: *Journal of the American statistical association* 58.301, pp. 236–244.

Wech, AG et al. (2012). "Tectonic tremor and deep slow slip on the Alpine Fault". In: *Geophysical Research Letters* 39.10.

White, Randall A et al. (1998). "Observations of hybrid seismic events at Soufriere Hills volcano, Montserrat: July 1995 to September 1996". In: *Geophysical Research Letters* 25.19, pp. 3657–3660.

Xu, Rui and Don Wunsch (2008). *Clustering*. Vol. 10. John Wiley & Sons.

Zimmerman, Robert W and Michael S King (1986). "The effect of the extent of freezing on seismic velocities in unconsolidated permafrost". In: *Geophysics* 51.6, pp. 1285–1290.