Real-time seismic monitoring using DAS fiber-optic instrumentation and machine learning : towards autonomous classification of natural and anthropogenic events

Camille Huynh

Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale Sciences de la terre et de l'environnement | ED 413 Université de Strasbourg

Ecole Doctorale 413 : Sciences de la Terre et de l'Environnement

Institut Terre et Environnement de Strasbourg

THÈSE présentée par / DISSERTATION presented by :

Camille HUYNH

soutenue le /defended on : **11 Février 2025 / 11 February 2025**

pour obtenir le grade de / to obtain the grade of :
Docteur de l'Université de Strasbourg / Strasbourg University Doctor

Discipline/Spécialité / Discipline/Specialty : **Géophysique**

Real-time seismic monitoring using DAS fiber-optic instrumentation and machine learning: towards autonomous classification of natural and anthropogenic events

THÈSE dirigée par / DISSERTATION supervisor :

MALET Jean-Philippe	Directeur de Recherche, ITES/EOST (Université de Strasbourg)
LANTICQ Vincent	Directeur Technique, FEBUS Optics

RAPPORTEURS :

METAXIAN Jean-Philippe	Directeur de recherche IRD, IPGP
CHALJUB Emmanuel	Physicien, ISTerre (Université Grenoble Alpes)

AUTRES MEMBRES DU JURY / OTHER MEMBERS OF THE JURY :

PELLETIER Charlotte	Maitresse de Conférences, IRISA (Université de Bretagne Occidentale)
JOUSSET Philippe	Chercheur, Deutsche GeoForschungs Zentrum
WEBER Jonathan	Professeur des Universités, IRIMAS (Université Mulhouse-Haute-Alsace)

CO-ENCADRANTS/TES :

HIBERT Clément	Physicien Adjoint, ITES/EOST (Université de Strasbourg)
JESTIN Camille	Chercheuse, SHOM

INVITÉS (le cas échéant) / INVITED MEMBERS (if applicable) :

REBEL Estelle	Chercheuse, TotalEnergies

Institut Terre & Environnement de Strasbourg | ITES | UMR 7063 de l'Université de Strasbourg & CNRS & ENGEES

FEBUS optics

Résumé

Mots-clés: Distributed Acoustic Sensing (DAS), Apprentissage machine, Sismologie, Espace latent, Tremblements de terre, Sources sismiques anthropiques, Apprentissage supervisée, Apprentissage non-supervisée

La sismologie permet d'étudier de nombreux phénomènes naturels par les ondes sismiques qu'ils génèrent. À partir de ces ondes nous pouvons remonter à de nombreuses informations, à commencer par le temps précis d'occurrence des phénomènes, mais aussi leur localisation, les propriétés de leur source et les propriétés du milieu. Les enregistrements de ces ondes sismiques sont basés principalement sur l'utilisation des réseaux de sismomètres. L'observation et l'analyse des données acquises permet d'identifier des caractéristiques communes associées à un type d'événement particulier, qui pourront être exploités pour comprendre les mécanismes physiques impliqués et potentiellement quantifier plus précisément l'aléa. Depuis quelques années, en parallèle d'une approche par sismomètre classique, une nouvelle technologie basée sur l'utilisation de fibres optiques a émergé pour la surveillance d'événements acoustiques naturels ou anthropiques: le DAS (Distributed Acoustic Sensing). Cette technologie innovante permet de mesurer des vibrations sismiques avec une très haute résolution spatiale sur des distances allant de quelques dizaines de mètres à plusieurs centaines de kilomètres. Bien que ces données soient plus volumineuses et complexes à traiter que celles issues des sismomètres traditionnels, elles offrent des perspectives prometteuses, en particulier pour l'analyse des champs d'ondes générés par les tremblements de terre, la détection des glissements de terrain, de divers événements anthropiques (par exemple des mouvements piétons, des mouvements de véhicules, ou des signaux sismiques générés par des travaux tiers), d'événements de faible amplitude ou très localisés (par exemple l'observation de microsismicité ou d'événements naturels environnementaux) ou encore une localisation précise de l'origine de ces événements sismiques.

L'objectif de ce travail de thèse vise à lever les verrous qui portent sur l'élaboration puis l'application d'une chaîne de traitement en proposant une approche basée sur l'apprentissage machine pour analyser en temps quasi réel des données sismologiques collectées à l'aide d'une fibre optique. L'objectif est consacré à la surveillance à échelle locale puis régionale de zones spécifiques, afin de permettre, à terme, de détecter et identifier en temps réel des événements naturels tels que les tremblements de terre et les glissements de terrain.

Au sein de cette thèse, nous avons choisi d'explorer dans un premier temps l'utilisation d'algorithmes classiques en apprentissage machine, comme les forêts aléatoires et le gradient boosting. Nous nous inspirons du travail réalisé par Provost et al. (2017), proposant l'utilisation d'un ensemble d'une soixantaine d'attributs pour la classification automatisée par algorithme des forêts aléatoires de signaux sismiques naturels (chute de roches, tremblements de terre, microséismes) mesurés par deux stations sismologiques permanentes. Ces attributs décrivent le comportement temporel et fréquentiel des signaux sismiques observés. Dans une première étude, nous avons utilisé ces attributs pour entraîner un algorithme d'apprentissage automatique à reconnaître plusieurs types d'événement d'origine anthropiques (mouvement piéton, impact avec une pelle, compacteur en marche, présence d'une pelle mécanique, fuite de pipe) acquis à l'aide d'un DAS, le long d'une fibre optique et en milieu contrôlé. Afin de considérer dans la prise de décision le caractère redondant de l'information transportée par les traces mesurées à des positions spatialement proches dans

la donnée DAS, nous avons complété l'algorithme d'apprentissage automatique par un algorithme d'agrégation de voisinage appelé champs de Markov aléatoires. Cette chaîne de traitement est également construite afin de fonctionner en temps quasi-réel sur des flux de données (données de durée constante, découpé indépendamment de la présente ou non d'un événement) et a produit un taux de bonne classification de 87%.

Suite à cette première étude, nous avons observé que l'utilisation des attributs purement temporelles ne capturent qu'une partie de la complexité de la donnée DAS ce qui a motivé l'exploration de nouveaux attributs incluant l'information spatiale de la donnée DAS, ainsi que l'information de similarités entre traces sismiques mesurées par positions différentes. Elles complètent la liste préexistante d'attributs temporels et leur efficacité est testée sur une base de données acquise le long d'une fibre optique de 91 km déployée dans les Pyrénées pendant une durée de trois semaines. Cette étude a montré la capacité de notre algorithme à reconnaître l'ensemble des 13 tremblements de terre repérés à l'aide du service en ligne proposé par le BCSF-RENASS ainsi que les 3 tirs de carrières les plus proches (< 4 km) parmi les 6 repérés. Afin d'entraîner notre modèle d'apprentissage automatique, il a été nécessaire d'effectuer un relevé manuel des événements sur les données à l'aide de la liste des événements fournie par le service en ligne. Sur une période d'observation plus étendue ou dans une région à forte sismicité, la labellisation manuelle pourrait devenir plus laborieuse.

Une des solutions permettant de pallier partiellement à cette problématique de labellisation massive est l'utilisation d'algorithmes non supervisés. Les approches par clustering de partitionnement puis par clustering hiérarchique offrent une exploration approfondie du jeu de données, en organisant ces dernières en clusters et en établissant une hiérarchie entre eux. Le cœur de notre travail de recherche sur cette approche a d'abord été d'identifier la meilleure représentation de la donnée DAS. Nous avons proposé deux représentations : l'utilisation des attributs préalablement utilisées par l'approche basée sur l'algoritme des forêts aléatoires, et une représentation en image combinant 4 transformations des données brutes (représentation en énergie, par le rapport STA/LTA, par moyennage des spectrogrammes calculés pour chaque position de fibre, et par représentation des spectres calculés pour chaque position de fibre). Les résultats montrent qu'il est possible de clusteriser la donnée fibre optique (FO), et que les deux ensembles d'attributs donnent des résultats intéressants. L'approche utilisant les attributs pré-calculés est intéressante car elle est beaucoup moins coûteuse en temps de calcul. Ces résultats très encourageants permettent d'imaginer de futurs systèmes de classification autonome, avec une intervention humaine réduite à la labellisation des clusters, et non plus à des millions de segments individuels dans la donnée mesurée par fibre optique.

En conclusion, cette thèse explore de manière approfondie l'application de techniques d'apprentissage automatique pour l'analyse de données sismiques mesurées par fibre optique, en tenant compte des spécificités liées à la nature dense et redondante de la donnée DAS. Nous avons démontré que l'utilisation d'attributs temporels et spatiaux, en association avec des algorithmes d'agrégation comme les champs de Markov aléatoires, permet d'améliorer la classification d'événements sismiques d'origine anthropique et naturelle. Les tests réalisés dans une zone de faible à moyenne sismicité, comme les Pyrénées, ont démontré qu'une labellisation manuelle réalisée avec l'aide du service en ligne proposé par le BCSF-RENASS était suffisante sur des périodes relativement courtes. Toutefois, dans des contextes à plus forte sismicité ou sur des périodes de mesure plus longues, cette approche pourrait devenir laborieuse à mettre en place. Dans ce cas, l'utilisation d'une approche auto-supervisée, permettant une labellisation par clusters plutôt que par événements individuels, apparaît comme une solution prometteuse pour automatiser la classification. Ces différentes pistes de réflexion offrent un potentiel considérable pour le développement futur de méthodes d'analyse de données DAS plus robustes et autonomes, avec des applications prometteuses pour la surveillance sismique et la compréhension des phénomènes naturels.

Abstract

Keywords: Distributed Acoustic Sensing (DAS), Machine Learning, Seismology, Latent Space, Earthquakes, Anthropogenic Seismic Sources, Supervised Learning, Unsupervised Learning

Seismology allows the study of numerous natural phenomena through the seismic waves they generate. From these waves, we can infer a wealth of information, starting with the precise timing of events, as well as their location, source properties, and the properties of the medium. The recording of these seismic waves is primarily based on the use of seismometer networks. Observing and analyzing the acquired data enables the identification of common characteristics associated with specific types of events, which can then be exploited to understand the physical mechanisms involved and potentially quantify the hazard more precisely.

In recent years, alongside traditional seismometer-based approaches, a new technology based on the use of optical fibers has emerged for monitoring natural or anthropogenic acoustic events: Distributed Acoustic Sensing (DAS). This innovative technology enables the measurement of seismic vibrations with very high spatial resolution over distances ranging from tens of meters to several hundred kilometers. Although these data are larger and more complex to process than those from traditional seismometers, they offer promising perspectives, particularly for analyzing the wavefields generated by earthquakes, detecting landslides, monitoring various anthropogenic events (such as pedestrian movements, vehicle movements, or seismic signals from third-party activities), low-amplitude or highly localized events (such as monitoring microseismicity or environmental natural events), and precisely locating the origin of these seismic events.

The goal of this thesis is to overcome the challenges associated with developing and applying a processing chain by proposing a machine learning-based approach to analyze near-real-time seismic data collected via optical fiber. The objective is focused on local and regional monitoring of specific areas to ultimately enable the real-time detection and identification of natural events such as earthquakes and landslides.

Within this thesis, we initially explore the use of classic machine learning algorithms, such as random forests and gradient boosting. We draw inspiration from the work by Provost et al. (2017), which proposes the use of a set of about sixty attributes for the automated classification of natural seismic signals (rockfalls, earthquakes, microseisms) measured by two permanent seismological stations. These attributes describe the temporal and spectral behavior of the observed seismic signals. In a first study, we used these attributes to train a machine learning algorithm to recognize various types of anthropogenic events (pedestrian movement, shovel impact, running compactor, presence of a mechanical shovel, pipe leak) acquired using DAS along an optical fiber in a controlled environment. To account for the redundancy of information carried by traces measured at spatially close positions in the DAS data, we enhanced the machine learning algorithm with a neighborhood aggregation algorithm called random Markov fields. This processing chain is also designed to function in near-real-time on data streams (constant-duration data, independent of event occurrence) and achieved a classification accuracy rate of 87%.

Following this initial study, we observed that using purely temporal attributes only captures part of the complexity of DAS data, which motivated the exploration of new attributes incorporating spatial information from the DAS data, as well as similarity information between seismic

traces measured at different positions. These complement the existing list of temporal attributes, and their effectiveness was tested on a dataset acquired along a 91 km fiber optic cable deployed in the Pyrenees for three weeks. This study demonstrated the ability of our algorithm to recognize all 13 earthquakes detected using the online service provided by the BCSF-RENASS, as well as the 3 nearest quarry blasts (< 4 km) among the 6 identified. To train our machine learning model, manual event marking on the data was required using the event list provided by the online service. Over an extended observation period or in a high-seismicity region, manual labeling could become more labor-intensive.

One solution to partially address this massive labeling issue is the use of unsupervised algorithms. Partitioning clustering and hierarchical clustering approaches allow for an in-depth exploration of the dataset, organizing the data into clusters and establishing a hierarchy among them. The core of our research work on this approach has initially focused on identifying the best representation of DAS data. We proposed two representations: the use of attributes previously used in the random forest algorithm approach, and an image-based representation combining 4 transformations of the raw data (energy representation, STA/LTA ratio, averaging of spectrograms calculated for each fiber position, and spectrum representation calculated for each fiber position). The results show that it is possible to cluster the fiber-optic data (FO), and that both sets of attributes yield interesting results. The approach using pre-calculated attributes is appealing because it is much less computationally expensive. These very promising results open the door to future autonomous classification systems, with human intervention reduced to labeling the clusters rather than millions of individual segments in the fiber-optic measured data.

In conclusion, this thesis provides an in-depth exploration of the application of machine learning techniques for the analysis of seismic data measured by optical fiber, taking into account the specificities related to the dense and redundant nature of DAS data. We have demonstrated that the use of temporal and spatial attributes, combined with aggregation algorithms like random Markov fields, improves the classification of seismic events of both anthropogenic and natural origin. Tests conducted in areas of low to moderate seismicity, such as the Pyrenees, showed that manual labeling assisted by the online service provided by BCSF-RENASS was sufficient over relatively short periods. However, in areas with higher seismicity or over longer measurement periods, this approach may become labor-intensive. In such cases, the use of a self-supervised approach, which allows for labeling by clusters rather than individual events, appears as a promising solution for automating classification. These various avenues of thought offer considerable potential for the future development of more robust and autonomous DAS data analysis methods, with promising applications for seismic monitoring and the understanding of natural phenomena.

Acknowledgments

It has now been a little over three years since I embarked on this doctoral journey in October 2021. Today, as I reach this significant milestone, I am deeply grateful for the opportunity to reflect on the path I have taken and to express my thanks to those who have supported and accompanied me throughout this academic adventure. Completing a PhD is a challenging and transformative experience, and I could not have reached this point without the guidance, encouragement, and kindness of many remarkable individuals.

I would like to express my gratitude to the following individuals, directly involved in the supervision of this doctoral thesis: Jean-Philippe Malet, my PhD supervisor, who has been the guiding force behind this work. His overarching perspective, strategic input, and ability to steer the project in the right direction have been invaluable. I am particularly grateful for his efforts in securing the contacts and authorizations necessary for deploying the fiber optic setup in the field (Viella), as well as his thoughtful feedback for the article manuscript reviews and the PhD manuscript thesis review. Clément Hibert and Camille Jestin, my co-supervisors, who played a decisive role since the very beginning. From helping shape this PhD project during my master internship in 2020, to introducing me to the fascinating world of seismology and DAS (as I transitioned from a background in physics and computer science), their expertise has been a constant source of inspiration. Clément's knowledge in seismology and AI, combined with Camille's expertise in DAS and data processing, have been invaluable. I deeply appreciate their kind guidance, constructive advice, and above all, their tremendous patience. Vincent Lanticq, for graciously accepting the collaboration between the ITES laboratory and FEBUS Optics company.

I would like to thank those who were not directly involved in the supervision, but whose support was essential for the successful completion of this doctoral thesis: David Michéa for his contributions to scientific computing, including assistance with computational infrastructure, HPC workflows, and the development of parallelized processing codes. The EOST-A2S HPC facility at the University of Strasbourg, part of the Data-Terra Research Infrastructure, was used, with some resources funded by the Equipex + GAIA-DATA project and the CPER A2S program. Fortuné Bayoua, Miguel Villar, and Mathieu Champion from FEBUS Optics for their guidance on event simulation protocols and their help in simulating leaks for the DAS measurements in the test bench, as described in Chapter 3. Gaëtan Calbris (FEBUS Optics) for his expertise with the FEBUS A1-R system during the measurement period at the test bench and in the Pyrenees, as presented in Chapters 3 and 4. Estelle Rebel and TotalEnergies for their collaboration with FEBUS Optics, which enabled the Pyrenees measurements discussed in Chapter 4.

I also want to thank the thesis committee who took their time to read carefully my work and attended the thesis defense. Thank you for your time, expertise, and valuable feedback. Thank you very much Emmanuel Chaljub, Jean-Philippe Metaxian, Charlotte Pelletier, Philippe Jousset, and Jonathan Weber.

I am profoundly grateful to the organizations that made this research possible through their financial support: The French National Association for Research and Technology (ANRT), for approving the CIFRE thesis project and awarding me a scholarship to carry out this work in collaboration

between FEBUS Optics and the Institute of Earth and Environment of Strasbourg. FEBUS Optics, for their significant contribution to funding this PhD until its completion, as well as for supporting the dissemination of my research through the publication of my first scientific article, my participation in AGU conferences and my participation to the "Galileo Conference: Fibre Optic Sensing in Geosciences" workshop. The ANR HighLand project, for its financial support, which allowed me to present my work at the EGU conference and contributed to the preparation of my second scientific article.

I am deeply thankful to my fellow PhD students and colleagues who shared this journey with me, making these three years an enriching and enjoyable experience. To my office mates, Joachim and Charlotte, my true companions throughout this adventure: thank you for the countless coffee breaks, moments of relaxation, and for sharing your daily challenges so we could brainstorm and develop tools to make our work easier. A special thanks to Joachim for introducing me to VSCode, which I now consider an essential tool for prototyping. To Franck, Qinglin, Weiwei, Mérédith, Clément (for introducing me to the SUAPS hikes), Emmanuel, Bérénice, Rémi, Julien, Hugo, Bastien M, Bastien W, Roxane (for the hockey games you got me hooked on), Estelle, Flavien, and all the other PhD students from the doctoral school I may have missed: thank you for the engaging discussions, moments of relaxation, coffee breaks, lunches, and memorable outings.

Last but not least, I would like to express my deepest gratitude to my parents and my two sisters Alicia and Aline for their unwavering emotional support throughout this journey. To my parents, thank you for always encouraging me in my studies and believing in me every step of the way. To my sisters, I am immensely grateful for your support, and for being there to listen and share in moments of doubt. Your presence has been a source of strength and comfort during this challenging journey.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence. 9, 33, 34, 37–39, 42–44, 64, 81

AUC-ROC Area Under the Receiver Operating Characteristic Curve. 38

BCSF Bureau Central Sismologique Français. ii, iv, 10, 11, 63, 65–68, 83, 85, 87, 101–103, 105, 106, 109, 110, 145

BYOL Bootstrap Your Own Latent. xiv, 36, 84, 88, 90, 92–103, 105–107, 157, 159, 161, 162, 164, 166

CNN Convolutional Neural Network. 42

DAS Distributed Acoustic Sensing. i–iv, 4–11, 13–33, 38–40, 42–45, 47–49, 55, 58, 59, 61, 63–67, 69, 70, 72, 74, 75, 80–88, 90, 93, 102, 103, 105–109, 111, 112, 145, 151

DFOS Distributed Fiber Optic Sensing. xiii, 4, 63

DFT Discrete Fourier Transform. 53, 57, 143, 144

DL Deep Learning. 34, 42–45

DT Derivation Time. 7

DTW Dynamic Time Warping. xiv, 72, 73, 153, 154

EB Energy Band. 32, 68–70, 79, 80, 90, 91, 95, 96, 98–100, 145, 157–161

EMD Empirical Mode Decomposition. 39, 40

EQ Earthquake. 70, 71, 80, 90, 98, 101, 102

FFT Fast Fourier Transform. 39, 42, 68, 72, 93

FO Fiber Optic. ii, iv, 3, 4, 7–9, 11, 13–15, 17–21, 23–33, 38, 39, 43, 111, 112

GL Gauge Length. 5, 7, 48

HPC High Performance Computing. 45

LOOCV Leave-One-Out Cross-Validation. 74, 75, 77, 80

LSTM Long Short-Term Memory. 43

MAD Median Absolute Deviation. 101, 102

MAE Mean Absolute Error. 38

MFCC Mel-Frequency Cepstral Coefficients. 39

MFE Morphological Feature Extraction. 39, 41

ML Machine Learning. 10, 34, 38, 40, 42, 44, 45, 85

MRF Markov Random Field. 10, 48, 54–56, 58, 61, 71, 74, 75, 77, 80

MSE Mean Squared Error. 38

OMP Observatoire Midi-Pyrénées. 85, 103, 105, 106

PCA Principal Component Analysis. 35, 39, 95

PRF Pulse Rate Frequency. 5, 7

PSD Power Spectral Density. 32, 90, 91

QB Quarry Blast. 70, 71, 80, 90, 98, 101, 102

RENASS Réseau National de Surveillance Sismique. ii, iv, 10, 11, 63, 65–68, 83, 85, 87, 101–103, 105, 106, 109, 110, 145

RF Random Forest. 10, 35, 48, 53, 54, 56, 58, 59

SMOTE Synthetic Minority Oversampling Technique. 37, 53

SNR Signal to Noise Ratio. 28, 81, 109, 112

SR Strain Rate. 5, 16, 61, 65, 67–74, 79, 80, 86, 88, 90, 93, 145, 151

SSL Self-Supervised Learning. 85, 91, 107

STA/LTA Slow Time Average Over Long Time Average. ii, xiii, 29–31, 33, 42, 64, 85, 88, 90, 91

SVM Support Vector Machine. 35, 39, 41

t-SNE t-Distributed Stochastic Neighbor Embedding. xiv, 35, 95–98, 101, 157–161

VT Volcano Tectonic. 17, 41

WT Wavelet Transform. 39, 40

XGBoost Extreme Gradient Boosting. xiv, 61, 63, 64, 70, 73, 74, 77, 78, 80, 109

Chapter 1

Introduction

From Seismogenic Sources to Seismological Observations and Signals

Seismological monitoring offers information about the Earth internal structure and geohazard processes. Traditional seismometers captures ground vibrations and seismic waves across a broad spectrum of frequencies. When recording seismological signals, it is essential to consider that the measured seismic waveform $u(t)$ is shaped by multiple elements, grouped into three main categories and summarized in the Equation 1.1 (Aki & Richards, 2002) and Figure 1.1 .

$$u(t) = s(t) * g(t) * i(t) \quad \text{where } * \text{ denotes the convolution operator} \tag{1.1}$$

The properties of the seismogenic source $s(t)$ are fundamental in shaping the recorded waveforms. These properties include the magnitude, the duration, and the frequency content of the emitted seismic waves. As the seismic signal travel through the Earth subsurface, they are influenced by the medium they pass through $g(t)$. Changes in geological structures, rock types, and other subsurface features, as well as the distance between the source and the sensor, can lead to scattering, refraction, and attenuation of seismic energy, causing distortions in both amplitude and frequency content. Finally, the instrumental response of the sensor $i(t)$ modifies the recorded signal. Different seismometers exhibit variable sensitivities, frequency ranges, and noise characteristics, all of which influence the fidelity of the recorded waveform.

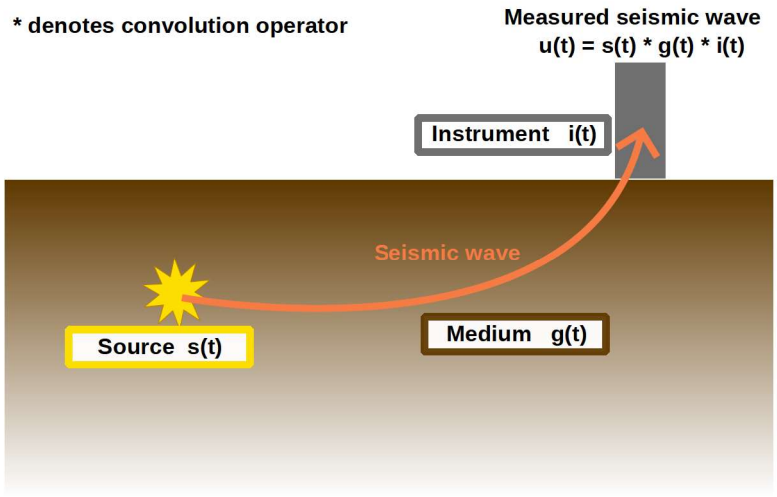


FIGURE 1.1: From seismogenic source to seismological aquisition. The seismic signal is shaped by the source $s(t)$, the propagation medium $g(t)$, and the sensor instrument $i(t)$.

Conventional Seismological Instruments and Their Applications

An instrument is a device used to measure a physical quantity. In seismology, the goal is to quantify ground deformation associated with wave propagating from a seismic source. Seismologists then aim at measuring ground displacement, acceleration, strain, or strain rate. These physical quantities are referred to as seismic signals. A seismological instrument consists in several essential components: a **sensor** that detects physical changes in the ground, a **data acquisition and processing system** (also called digitizer) that captures and analyzes the sensor signals, and a data storage system to record the collected data. These instruments can be **passive**, in which case they are not powered by an external energy source, or **active**, in which case they are powered by a battery or an electrical outlet. Among the most commonly used instruments are the active instruments called broadband seismometers, accelerometers, and geophones. Figure 1.2 illustrates the frequency responses of these instrument families.

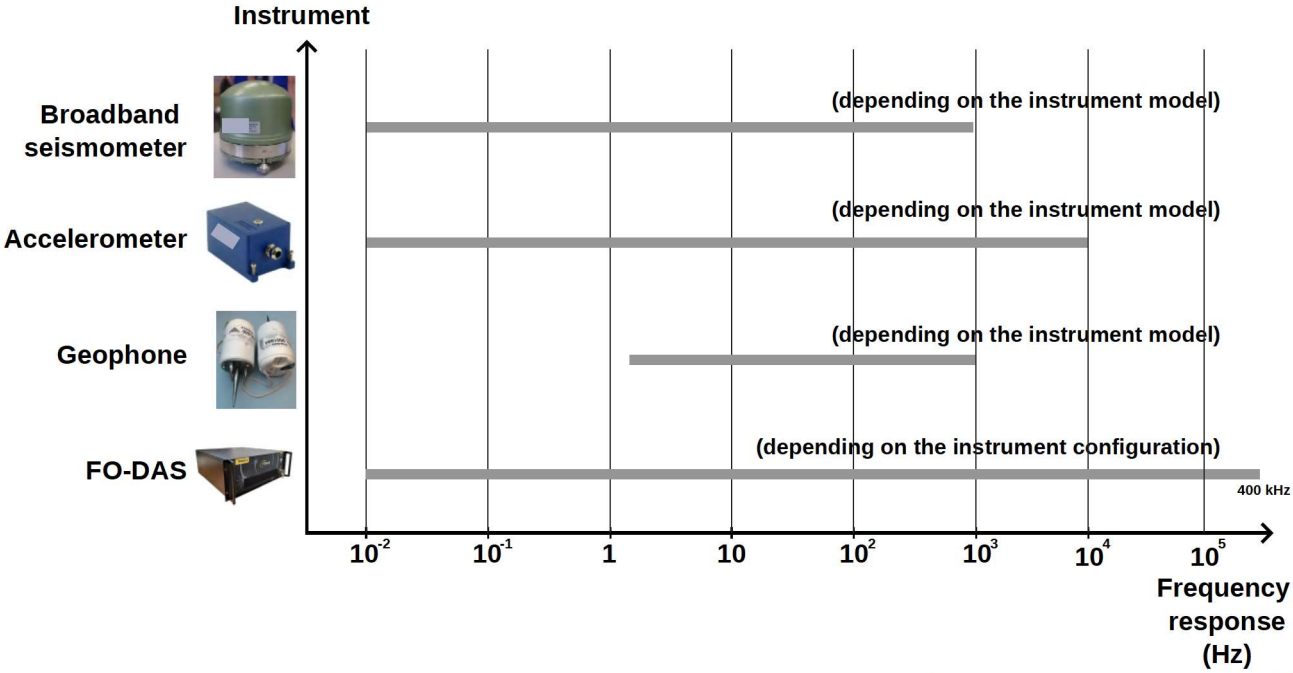


FIGURE 1.2: Comparison of the frequency responses of several families of seismological instruments. Broadband seismometers, accelerometers, geophones and distributed acoustic detection are included in the comparison.

- **Broadband seismometers** are sophisticated instruments widely employed in Earth science observatories for seismological monitoring. They are designed to measure ground movement velocities across a wide frequency range, from very low frequencies (as low as 0.01 Hz) to higher frequencies, up to several hundred of hertz (Ackerley et al., 2014). This versatility makes them ideal for capturing a large spectrum of seismic activity, from local seismogenic earthquakes (Havskov & Alguacil, 2004; Pechmann et al., 2007), seismicity around volcanoes (Benoit & McNutt, 1997; Kawakatsu et al., 2000; Takeo et al., 2010; Kaneko et al., 2018), seismicity under the oceans (Webb, 1998; Collins et al., 2001), to large-scale regional and global teleseismogenic earthquakes (Tsuboi et al., 1999; Havskov & Alguacil, 2004). These seismometers work using a precisely controlled inertial mass that responds to ground displacements. Because of their cost, their sensitivity and their wide frequency coverage, broadband seismometers are typically installed in permanent observatories, where they provide continuous and high-quality data for seismological research.

- **Accelerometers** are specifically designed to measure strong ground motions, such as those generated by nearby, high-magnitude seismogenic sources. They directly measure ground acceleration, making them well-suited for detecting rapid and intense movements (Geng et al., 2013; Zou et al., 2014). This makes accelerometers a preferred choice for monitoring seismogenic hazards in urban environments (D'Alessandro et al., 2014; Patanè et al., 2022), where they are often installed to assess the seismic response of critical infrastructure like buildings (Mita & Yokoi, 2001; Santana et al., 2012; Mahjoubi et al., 2020), bridges (Meng et al., 2007; Moschas & Stiros, 2011; Han et al., 2016; Xiong et al., 2017; Chilamkuri & Kone, 2020), and other structures. With their focus on low to medium frequencies and sensitivity to high-amplitude signals, accelerometers are essential in civil engineering and risk assessment, providing valuable data to design earthquake-resistant structures and to improve urban resilience against seismic risks.

- **Geophones** are portable seismological instruments designed to detect small, localized events, making them ideal for temporary deployments or rapid response scenarios. They are lightweight, easy to install, and sensitive to low-amplitude signals such as local microseisms or human-induced vibrations. Operating in a frequency range from 1 to several hundred of hertz (Hoover & O'Brien, 1980; Krohn, 1984), geophones are particularly useful for local-scale studies, such as monitoring microseismicity, detecting low-magnitude earthquakes, and studying natural hazards like landslides (Bordoni et al., 2007; Tonnellier et al., 2013), avalanches (Van Herwijnen & Schweizer, 2011; Heck et al., 2019), and rockfalls (Helmstetter & Garambois, 2010; Zimmer & Sitar, 2015). Due to their sensitivity, geophones can detect small earthquakes that might not be captured by broadband seismometer networks (Prugger & Gendzwill, 1988), offering higher resolution of local seismicity and improving the understanding of fault zones and seismic risk in tectonically active areas (Lienkaemper et al., 2006; Hansen & Schmandt, 2015). Dense geophone arrays in active fault regions, for example, can help illuminate fault behavior and detect foreshocks or aftershocks, aiding in the prediction of larger seismogenic events (Wurman et al., 2007; Allen & Stogaitis, 2022).

While geophones are versatile and cost-effective for local seismological monitoring, deploying and maintaining large networks of these instruments in diverse terrains presents logistical challenges, particularly in terms of power supply and data storage for long-term studies. Recent advances in Fiber Optic (FO) based technology offer a good solution, enabling continuous, large-scale monitoring over extensive distances with only a FO cable.

Fiber-Optic Based Instruments: Principles and Key Notions

Distributed Fiber Optic Sensing (DFOS) refers to instruments that integrate Fiber Optics (FO) and convert them into continuous, distributed sensors along their entire length. In DFOS instruments, the data acquisition and processing system is called an interrogator. Its function is to send laser pulses through the FO and to analyze the backscattered light (Figure 1.3). The backscattering happens due to tiny imperfections naturally present in the glass constituting the FO. When the FO is subjected to physical stress such as strain, temperature fluctuations, or vibrations, these imperfections shift slightly, causing a change in the phase of the backscattered light. By analyzing these changes, DFOS instruments can detect and pinpoint these physical variations along the full length of the cable. The three types of backscattering, known as Rayleigh, Raman, and Brillouin scattering, are based on several light wavelength shifts and carry different information (Hartog, 2017). Among the DFOS methods, **Distributed Acoustic Sensing** (DAS), relying on Rayleigh backscattering, excels at monitoring ground vibrations and seismic waves in real time across extensive areas and is one of the most widely used in seismology.

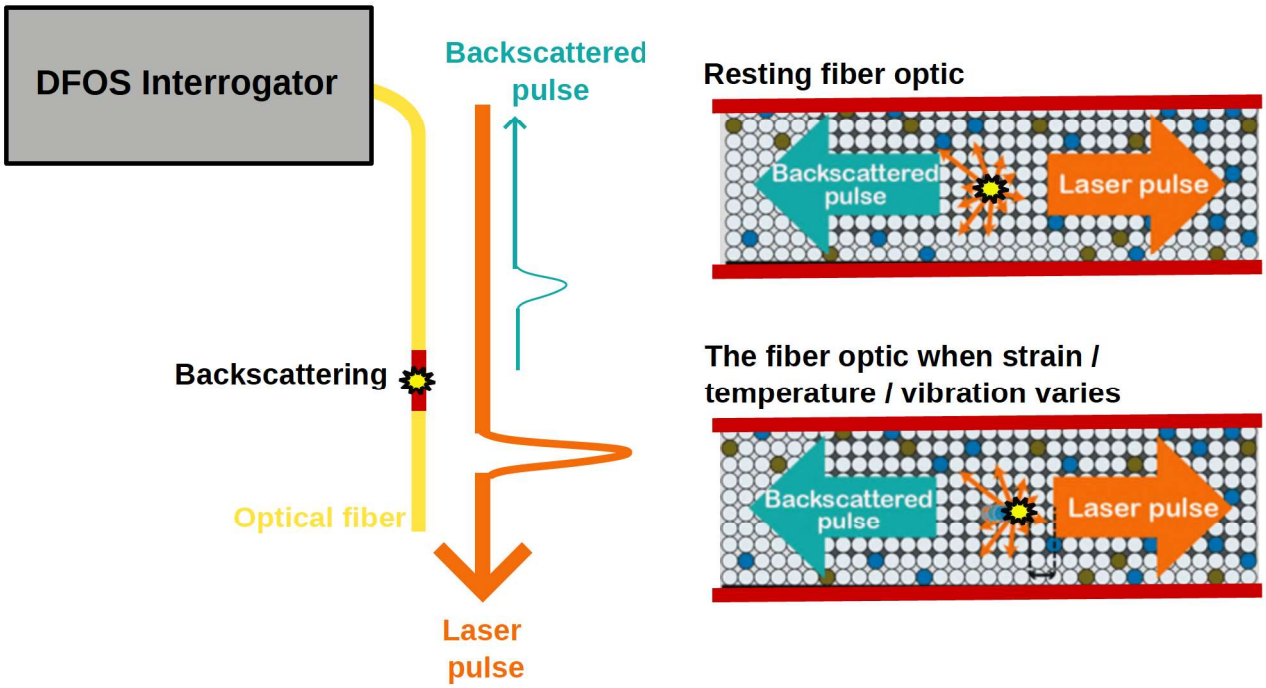


FIGURE 1.3: Distributed Fiber Optic Sensing (DFOS) general sketch. During acquisition, a laser pulse is sent through the fiber optic cable. The backscattered light is then analyzed to measure seismic vibrations (adapted from the Instituto Geografico Nacional website).

A measurement cycle runs as follows:

- First, a laser sends a light pulse into the FO.
- The light is continuously scattered by imperfections in the FO, and reflected back to the interrogator.
- A photo-receiver captures the scattered light, the phase shift is analyzed, and the corresponding ground deformation is calculated.
- The process is repeated to capture seismic waves at different times.

This measurement cycle has significant implications for both the spatial and temporal resolution of the resulting data. A key constraint in DFOS instruments is ensuring that the emitted laser pulse

has sufficient time to travel the entire length of the FO and return to the interrogator before the next measurement cycle begins. This is especially critical for the point farthest from the interrogator. Considering the speed of light $c = 3.0 \times 10^8$ m.s^{-1}, the refractive index of the medium n, and the fiber length L, the **maximum frequency of laser pulse emission** $f_{\mathbf{R\_max}}$ can be expressed by the Equation 1.2:

$$f_{\text{R\_max}} = \frac{c}{2.n.L} \tag{1.2}$$

For example, a commonly used rule to set $f_{\text{R\_max}}$ is that for a 10-km fiber optic length, a $f_{\text{R\_max}}$ of 10 kHz is applied.

The choice of the frequency of the laser pulse emission $f_R < f_{\text{R\_max}}$, also called Pulse Rate Frequency (PRF), allows the calculation of the maximum delay $\tau_{\text{max}} = \frac{1}{f_R}$ below which it can be confirmed that the light received by the photodetector originates from the most recent laser pulse. This light is captured by a photo-receiver and digitized at a sampling frequency f_d. Two critical temporal dependencies arise from this process. First, the laser pulse emission frequency directly impacts the **temporal resolution** t_Δ, as described in Equation 1.3:

$$t_\Delta = \tau_{\text{max}} = \frac{1}{f_R} \tag{1.3}$$

Second, the sampling frequency of the digitizer determines the **spatial sampling** x_Δ, given by Equation 1.4:

$$x_\Delta = \frac{c}{2.n.f_d} \tag{1.4}$$

Figure 1.4a,b illustrates these concepts with optical signals emitted by the laser and received by the photo-receiver. In Figure 1.4c, the measured optical signal is cut at intervals of τ_{max} to extract the positions along the fiber optic. This results in a "fast time" axis, whose duration is limited by τ_{max}, and a "slow time" axis, with a temporal sampling of τ_{max}.

The consequence is the relationship between the reception delay of the laser pulse after emission $\tau < \tau_{\text{max}}$ and the measured fiber position x is given in Equation 1.5. From this equation, it follows that the "fast time" axis corresponds to the distance axis of the DAS aquisition, while the "slow time" axis represents the temporal axis.

$$x(\tau) = \frac{c.\tau}{2.n} \tag{1.5}$$

Strain Rate (SR), processed by the DAS interrogator and expressed in nstrain/s, is estimated by analyzing phase variations in the optical signal to assess ground displacements and deformations. Phase variation analysis is performed within both spatial and temporal signal windows represented in Figure 1.5. The **Gauge Length (GL)** is the spatial parameter defining spatial resolution. A shorter GL allows for the detection of smaller deformations over shorter distances but may increase noise

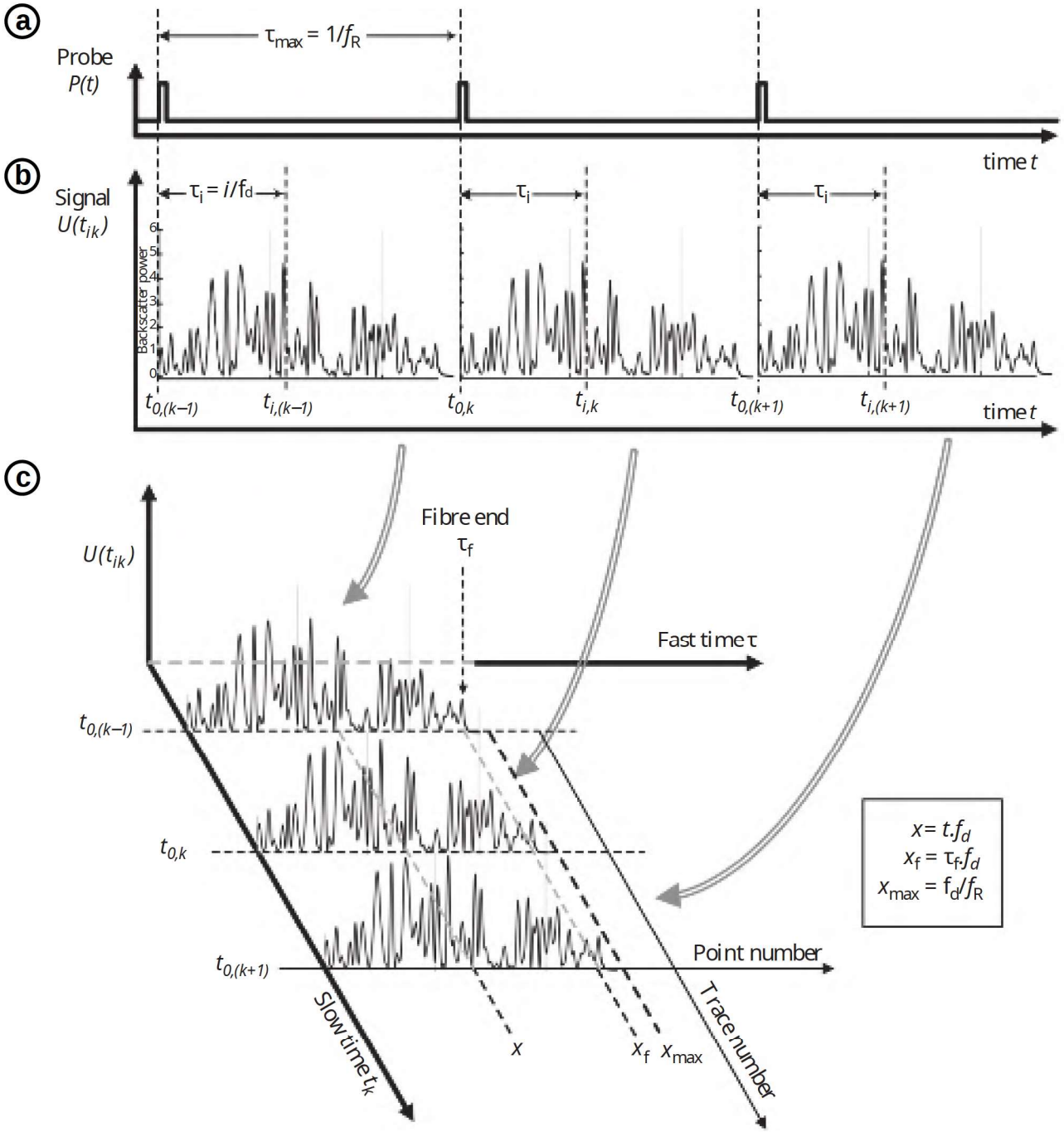


FIGURE 1.4: Time scale of DAS aquisition. The emitted laser pulse is represented in ***(a)*** *and the backscattered optical signal is measured by a photo-receiver and represented in* ***(b)****.* ***(c)*** *shows the same representation as* ***(b)*** *but represented using the "fast time" and "slow time" axes (adapted from Hartog, 2017).*

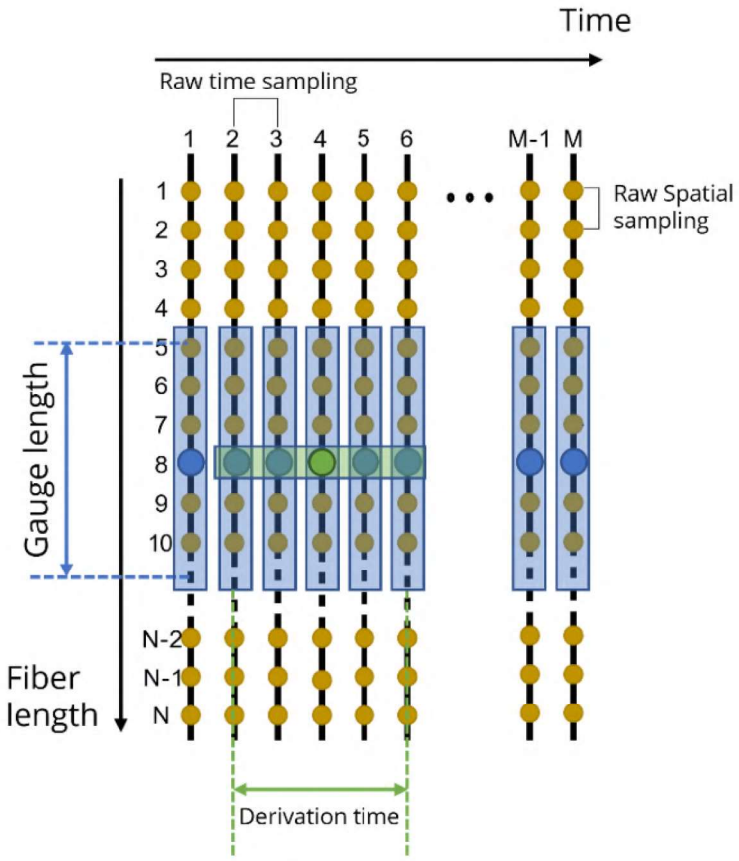


FIGURE 1.5: Definition of Gauge Length and Derivation Time. The yellow dotted line represents the raw optical signal acquired by the FO-DAS instrument (taken from FEBUS Optics documentation).

levels, while a longer GL offers better signal stability but reduced resolution. GL is, when allowed by the used DAS interrogator, set to exceed half the shortest anticipated seismic wavelength, ensuring accurate wavelet reconstruction and enhances the signal-to-noise ratio (Dean et al., 2017). For the temporal window, the **Derivation Time (DT)** is set to enable the measurement of the frequencies of interest, with the cutoff frequency being the inverse of DT. However, the corner frequency, where amplitude attenuation starts to distort the signal, occurs before the cutoff frequency. This corner frequency should be considered when selecting the appropriate DT to ensure accurate measurements. A shorter DT therefore enables the capture of higher-frequency seismic waves, which is useful for detecting rapid, high-frequency events such as local earthquakes or tremors. But it also increases noise levels, reducing the clarity of the signal. In contrast, a longer DT enhances the signal-to-noise ratio and is more suited for detecting lower-frequency seismic waves, such as those from larger, distant earthquakes. The choices are influenced by the seismogenic sources to be observed, but are also limited by the optical temporal sampling (t_Δ) determined by the PRF, and the optical spatial sampling (x_Δ).

Fiber-Optic Based Instruments: Documented Advantages and Drawbacks

FO-DAS instruments offer significant advantages over traditional seismological instruments in terms of data quantity, deployment and maintenance. In terms of data, FO sensor offers much **higher spatial resolution**, as FO-DAS instrument enables aquisitions with a resolution of a few tens of centimeters for small-scale configuration to a few tens of meters for a hundred kilometers of cable.

Each sample point along the fiber is referred to as a channel, with each channel representing a localized measurement of strain or deformation at that specific point in the fiber. FO-DAS instrument is sensitive to **strain rate from mHz** (Parker et al., 2014; Becker et al., 2017) **to several hundreds of kHz**, allowing for the detection of high-frequency seismic waves. In terms of deployment, FO-DAS systems can be more **cost-effective over large distances**, as a single cable combined with a single interrogator can cover extensive areas without the need for numerous individual instruments. Additionally, **existing deployed FO cable can often be repurposed for DAS acquisition**, reducing installation costs and environmental impact. FO-DAS instruments require significantly **less maintenance**, as data processing is managed by a single digitilizer, and data collection is carried out by a resilient FO cable. The FO cable itself is highly **durable and resistant to environmental and electromagnetic interference**, providing a robust and long-lasting solution for seismological monitoring. These advantages make FO-DAS instruments an attractive choice for large-scale and long-term seismic applications. FO-DAS instruments also present well-documented challenges. One key consideration is the **directional sensitivity** of the instruments, which requires careful planning of the fiber layout to ensure accurate data collection. Additionally, the **large volume** of data generated during long-term monitoring can place significant demands on storage and processing capabilities. A third challenge lies in the **higher noise levels** of FO-DAS instruments compared to traditional geophones, which can complicate the detection of weak seismic signals.

Fiber-Optic Based Instruments: Overview of the DAS Interrogator Industry and Key Players

The DAS interrogator industry is composed of companies specialized in several aspects of DAS technology. Some companies, such as OFs (US) and NKT Photonics (Denmark), focus exclusively on manufacturing interrogators. Other firms, like Halliburton (US), Baker Hughes (US), Senstar (Canada), Aragon Photonics (Spain), Cementys (France), and Bandweaver (China), produce complete solutions but rely on third-party interrogators. Other companies like Schlumberger (US), Omnisens (Switzerland), OptaSense (UK), Fotech Solutions (UK), Silixa (UK), AP Sensing (Germany), ASN (France), FEBUS Optics (France) offer both interrogators and integrated solutions. These firms specialize in monitoring pipelines, detecting leaks, securing perimeters, monitoring seismicity, and overseeing critical infrastructure, combining their own interrogators with tailored solutions for diverse industries. For the purpose of this thesis, we use the DAS FEBUS A1-R interrogator manufactured by FEBUS Optics.

Main Challenges and Research Objectives

The unique characteristics of DAS acquisition make it a powerful tool for seismological monitoring, but they also introduce significant challenges in data processing. The high spatial and temporal resolution of DAS data generates vast amounts of information, which demands efficient storage solutions, robust computational methods, and advanced techniques for post-acquisition analysis like data exploration. This introductory section already highlights the challenges associated with processing DAS data, which will be further elaborated at the end of Chapter 2. These challenges can be summarized by the following questions:

- One of the primary challenges lies in constructing an effective processing chain for DAS data, given their unique characteristics compared to conventional seismic data. **How can we adapt**

existing processing methods to accommodate the distributed spatial nature of DAS data?

- Another challenge is the identification of events of interest among the background noise, given the rich content of DAS data. **How can relevant seismogenic events be effectively selected?**

- For long-term monitoring systems, there is a need for extended recording durations to account for factors such as seasonal effects and environmental variations over time. In these cases, manually labeling the data becomes a significant challenge, as it can be time-consuming and labor-intensive. **How can training datasets for machine learning models be constructed more efficiently, especially when dealing with large volumes of continuous data that span across seasons or years?**

As a consequence of the challenges, this PhD thesis is dedicated to the development, enhancement and application on field data of advanced signal processing methods for the detection and classification of seismogenical events recorded by DAS data. Considering this goal, the research objectives are structured into three main topics:

- **Develop a complete DAS data processing chain** for real-time event detection and classification, using Artificial Intelligence (AI) techniques adapted from conventional seismological methods. This pipeline will incorporate spatial data coherency processing, specifically employing Markov Random Fields to account for spatial dependencies.

- **Integrate the spatial features of DAS data** and the relationships among virtual point sensors as features within the Machine Learning model, and apply the method to a DAS in-situ acquisitions.

- **Design a method to streamline the creation of training datasets for DAS acquisition** by clustering similar DAS signals, thereby eliminating the need of manual event-by-event annotation.

Outline of the thesis

The manuscript is divided into six chapters, each focusing on a specific aspect of the research. The chapters are structured around two published peer-reviewed papers and one article currently under review. Figure 1.6 provides an overview of the thesis structure.

Chapter 2 introduces the various types of measurable signals detected using FO-DAS instrument, including airborne acoustic, hydroacoustic, and seismic signals. It then explores the different seismogenic sources that can be identified in the field. Each source type is characterized by its waveform, frequency content, amplitude, and duration, which are important to understand for achieving accurate seismological monitoring. We then explore classical tools for seismogenic event detection and classification and the challenges they face when applied to DAS data. The limitations of traditional methods have led to the development of advanced signal processing techniques, particularly those based on Artificial Intelligence (AI). Recent advances in AI for seismology are presented in this chapter. Finally, the chapter places the thesis research within this context, outlining the main contributions of the thesis and the novel approaches proposed to address the challenges of seismogenic event detection and classification using DAS data.

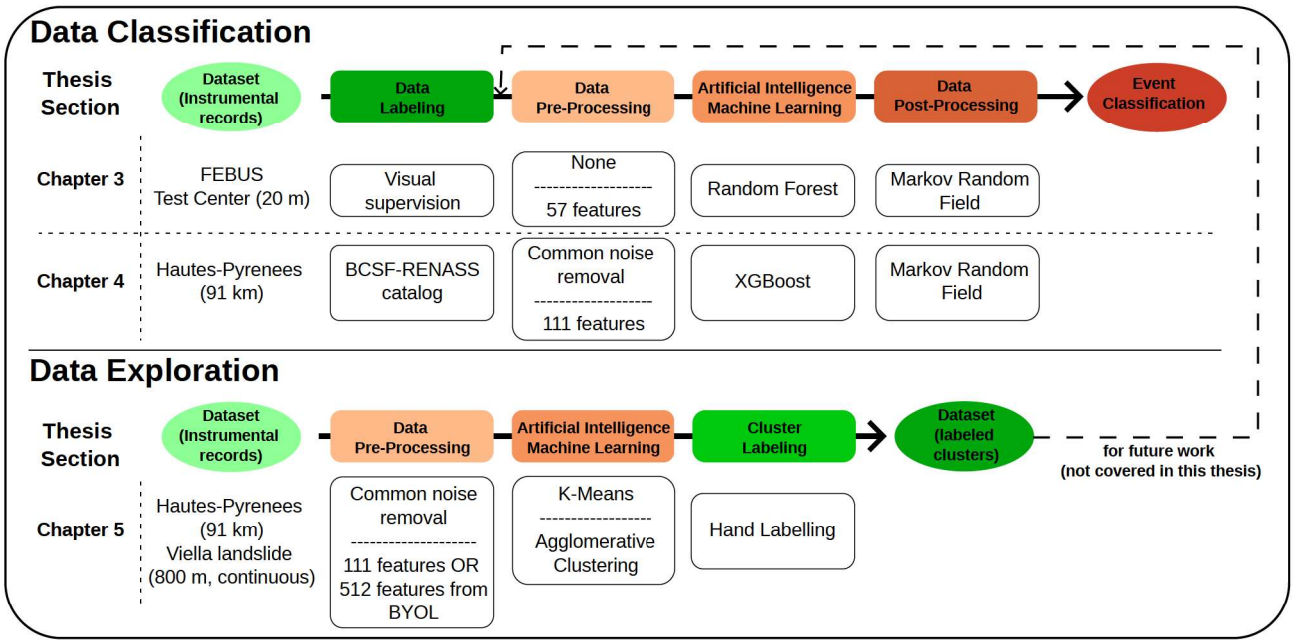


FIGURE 1.6: Overview of the goal of each chapter and the developed processing chain.

Chapter 3 presents the development of a DAS data processing pipeline designed for real-time event detection and classification. Drawing from successful methodologies that use handcrafted feature sets for data measured by conventional seismometers, this approach proposes a similar feature set tailored for DAS data, implemented with a Random Forest (RF) Machine Learning ML algorithm. The choice of RF was based on several key considerations, including its robustness, interpretability, and computational efficiency. To align with real-time processing requirements, a streaming approach is adopted, in which data is segmented into predefined windows for analysis. Additionally, to incorporate spatial dependencies between DAS data points, a Markov Random Field (MRF) model is applied after the initial classification stage, enhancing the accuracy of seismogenic event detection. The viability of this processing chain was tested on data acquired in a controlled test bench (CESG, FEBUS Optics). The events were accurately segmented in this controlled nature of the test environment. This approach has shown an accuracy of 87%. This first promising study gave us some keys of algorithm enhancement such as taking into acount one of the advantage provided by DAS systems: the continuous spatial sampling of the acquisition.

Chapter 4 focuses on integrating the spatial features of DAS data and the relationships between virtual point sensors, also called channels, as features within the ML model. The proposed method involves feature engineering to translate these DAS-specific spatial dependencies into features. The processing pipeline developed in Chapter 3 is maintained as the core structure in this chapter. For testing these new features, real-world DAS data are acquired from a 91-km fiber-optic cable deployed in the French Hautes-Pyrénées. We focus on 19 10-minutes-records, which include quarry blasts and earthquakes cataloged by the "Bureau Central Sismologique Français" and "RESeau NAtional de Surveillance Sismique" (BCSF-RENASS). This method provided very good results in low-magnitude event detection, achieving notable accuracy despite the presence of frequent anthropogenic disturbances along the fiber, such as moving vehicles, and meteorological activities that could locally degrade signal quality.

Chapter 5 addresses the need for semi-automated data annotation for continuous DAS data, as the previous methods require either manual annotation or pre-labeled datasets. When dealing with continuous data, manual annotation is highly time-consuming and impractical, unless a fully annotated dataset is available from an external source, an option that is rarely feasible. For the previously presented Hautes-Pyrénées dataset, only a partial annotation was possible. To streamline

the construction of training datasets, this chapter proposes a clustering-based approach for DAS data, helping for an efficient identification of the different seismogenic event types without requiring event-by-event examination. To pre-process DAS data, we leverage the handcrafted features used in previous chapters, and we introduce a comparison with features automatically generated using a self-supervised learning algorithm called BYOL. The method is validated on the Hautes-Pyrénées dataset and additional data acquired from a deployment at the Viella landslide (Hautes-Pyrenees, France), recorded from December 11, 2023, to January 24, 2024. The clustering approach has shown good results on a smaller scale, successfully identifying all seismogenic events with magnitudes greater than 1.5 in the Viella dataset, along with multiple previously uncatalogued events not listed in the BCSF-RENASS. Additionally, the method successfully identified anomalies caused by fiber breaks and revealed agricultural activity cycles near the deployment site.

Chapter 6 concludes the thesis work, by summarizing the main contributions and findings of the research. It also discusses the limitations of the proposed methods and suggests potential avenues for future work. The chapter concludes with a reflection on the broader implications of this research work and its potential impact on the field of DAS research and more generally the use of FO-DAS instrument in seismology .

Chapter 2

Fiber Optics as Seismological Sensors: A State-of-the-Art Review

2.1 Measurable Signals with Distributed Acoustic Sensing Instruments

The Distributed Acoustic Sensing (DAS) instrument uses Fiber Optic (FO) cables as sensors to detect vibrations in the surrounding environment. When vibrations occur, the FO cable deforms slightly, and the signal is detected and measured by the DAS interrogator and converted into a strain rate. The nature of the vibrations depends on the medium in which the fiber is deployed. Vibrations are classified into two main types: acoustic waves (which travel through liquids or gases) and seismic waves (which propagate through solid media like the ground).

2.1.1 Acoustic Waves

Acoustic waves are pressure waves that travel through liquids or gases due to particle oscillations. These waves are longitudinal, meaning the particles move in the same direction as the wave. Underwater acoustic waves typically have frequencies between 1 Hz and 1 MHz, as higher frequencies cannot propagate in water. Airborne acoustic waves does not have frequency limit. Human hearing ranges from 20 Hz to 20 kHz, but some animals communicate using frequencies outside this range, either infrasonic (below 20 Hz) or ultrasonic (above 20 kHz).

Hydroacoustic waves are generated by a diverse range of natural and human-induced sources. Marine life, including mammals, fish, and birds, produce these signals (Krieger & Wing, 1986; Wanzenböck et al., 2003), as do environmental phenomena such as surface waves, underwater currents, and tsunamis (Bolshakova et al., 2011; Davy et al., 2014). Seismogenic events, like ice detachment and underwater volcanic activity, also generate hydroacoustic waves (Fox & Dziak, 1998; Talandier et al., 2006; MacAyeal et al., 2008). Additionally, human activities, such as sonar used for marine life monitoring and various shipping or offshore operations, contribute to the generation of these waves (Stanton & Clay, 1986; Brehmer et al., 2003; Hildebrand, 2009). In contrast, **airborne acoustic waves** are produced by both natural and anthropogenic sources. These include sounds from animals, thunderstorms (Bedard, 2005), wind, and waterfalls, as well as human activities like speech, traffic, construction, and industrial noise (Grollmisch et al., 2019).

FO-DAS technology, though not yet widely adopted for acoustic applications, has demonstrated capabilities in detecting hydroacoustic signals. For instance, FO-DAS instrument has been successfully used to monitor air-gun shots, with studies comparing its performance to traditional hydrophones (Matsumoto et al., 2021). In addition to this, FO-DAS instrument has been employed to detect various marine phenomena, including whales, storms, ships, and earthquakes, using fiber optic cables in Arctic regions (Landrø et al., 2022). While its use for airborne acoustics remains relatively uncommon, FO-DAS instrument is increasingly being explored for air traffic monitoring, particularly in the detection of unmanned aerial vehicles (UAVs), as highlighted by recent studies (Fang et al., 2022; Chen et al., 2024).

2.1.2 Seismic Waves

Seismic signals cover a wide range of frequencies, with typical seismological frequencies ranging from 10 mHz, associated with Earth tides, to 100 Hz, which are typical for earthquakes (Hou et al., 2021). These seismic signals are the result of elastic waves that propagate through the Earth or along its surface, caused by the deformation of the medium through which they travel. Seismic waves can be broadly grouped into two main families: body waves and surface waves. Figure 2.1 illustrates the different types of seismic waves.

Body waves travel through the Earth interior and are further divided into two types: Primary (P) waves and Secondary (S) waves. P waves are compressional waves that move faster than S waves, and are often the first to be detected by seismographs. Due to their lower energy, P waves tend to cause less ground shaking and are generally less destructive. On the other hand, S waves move perpendicular to the direction of propagation, which makes them slower but more powerful. These waves tend to cause stronger ground shaking, making them more destructive during seismic events. Body waves, which travel through the Earth interior, are mainly caused by deep natural seismogenic events such as earthquakes and volcanic activity. Landslides and anthropogenic explosions, though significant, produce weaker body seismic waves.

Surface waves, such as Rayleigh and Love waves, travel along the Earth surface and can result in significant ground motion. These waves are generally more destructive in the immediate vicinity of the seismogenic source, however, their energy diminishes with distance from the epicenter, reducing their impact at greater distances. Landslides, rockfalls, and tectonic activity are the main natural sources of surface waves. Human activities, including walking, traffic, and explosions, also generate surface waves, and can be noticeable in urban or industrial areas.

The FO-DAS instrument is sensitive to vibrations in the longitudinal direction of the fiber optic cable. Typically, the fiber is deployed on the surface or lightly buried, running parallel to the ground. As a result, FO-DAS instrument is highly sensitive to surface waves (Zhu et al., 2023). For body waves, the sensitivity is generally lower, as it depends on the angle between the wave propagation direction and the cable orientation. However, specific configurations, such as downhole setups, can increase the sensitivity to body waves (Lellouch & et al., 2019; Yuan et al., 2020). Helical configurations can also improve FO-DAS sensitivity to the body waves (Ning & Sava, 2018; Al Hasani & Drijkoningen, 2023).

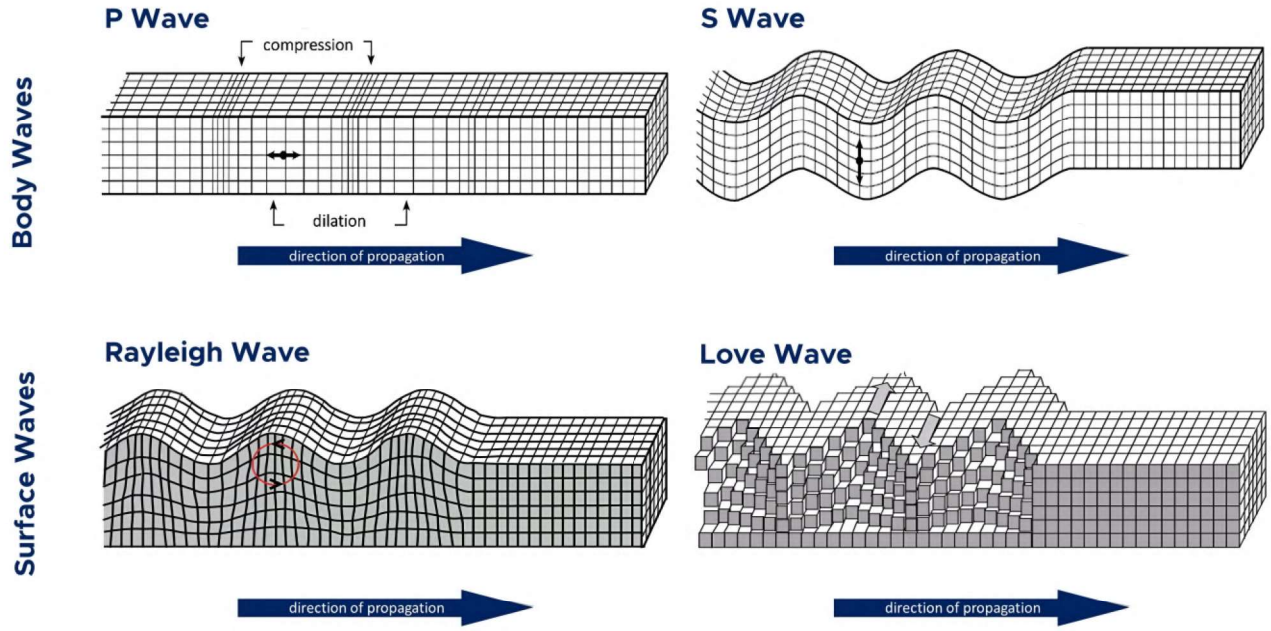


FIGURE 2.1: A seismic wave is an elastic wave generated by an impulse such as an earthquake or an explosion. Seismic waves may travel either along or near the earth surface (Rayleigh and Love waves) or through the earth interior (P and S waves) (from Wikimedia Commons, work by Luke Triton, distributed under CC by 4.0).

2.2 Seismogenic Sources

Seismic signals are generated by a wide range of natural and anthropogenic sources. These sources can be classified into four main categories: tectonic earthquakes, volcanic seismogenic sources, environmental seismogenic sources, and anthropogenic sources. Each of these sources produces distinct seismic signals, which can be detected and analyzed using seismological instruments like FO-DAS. Their frequency response is summarized in Figure 2.2. In this section, we provide an overview of these seismogenic sources and their associated seismic signals.

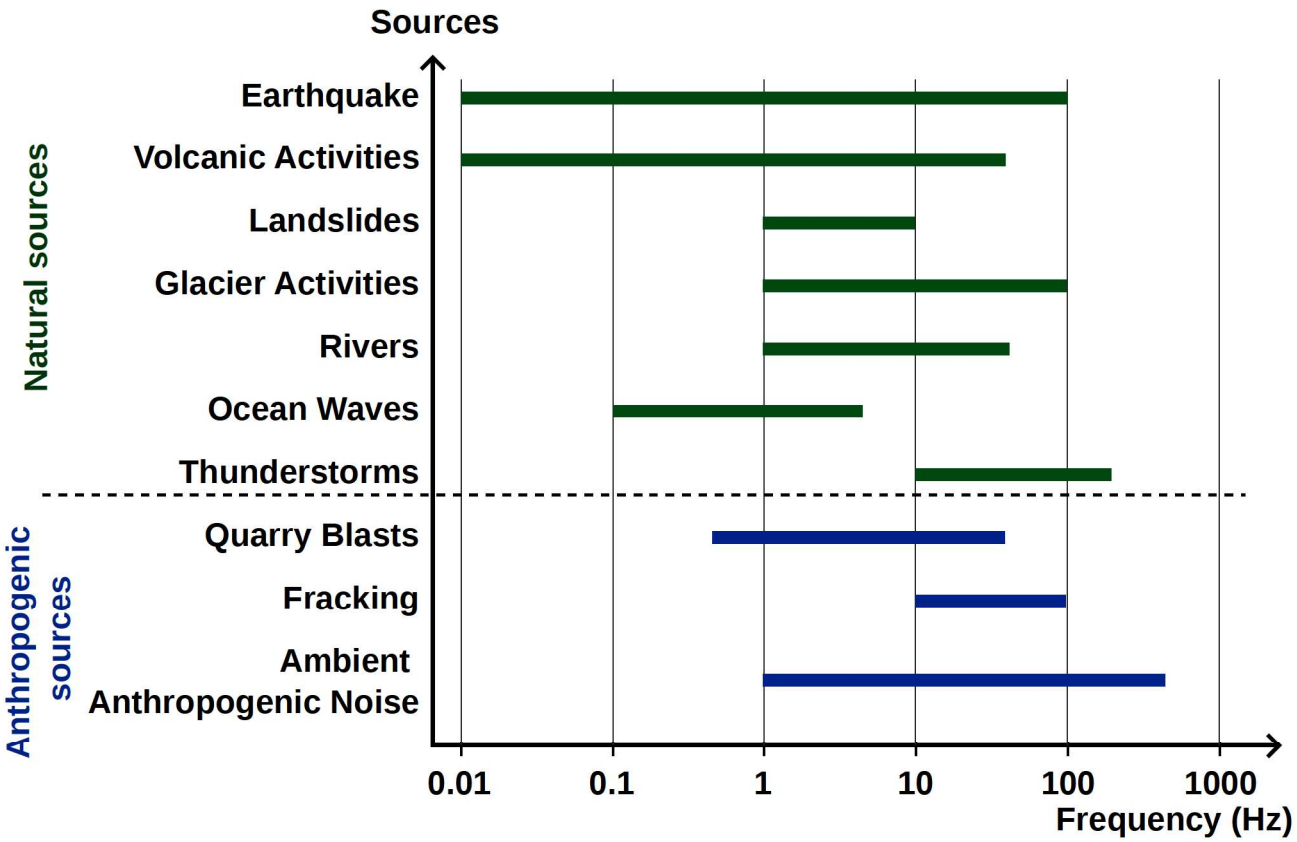


FIGURE 2.2: Comparison of frequency responses for several families of seismological instruments. Broadband seismometers, accelerometers, geophones and distributed acoustic sensing are present in the comparison.

2.2.1 Tectonic Earthquakes

Tectonic earthquakes occur due to the movement of tectonic plates that generates mechanical stress along fault lines where the plates meet. As the Earth crust moves, stress builds up in these faults over time. When the accumulated stress exceeds the fault ability to resist, the fault suddenly ruptures, releasing energy as seismic waves. This process, driven by the continuous movement of tectonic plates, is the main cause of natural earthquakes and has been studied for decades (Rice, 1979; Olson & Allen, 2005; Wesnousky, 2008; Wang et al., 2013; Romano et al., 2014). The occurrence frequency and intensity of tectonic earthquakes are influenced by several factors. Proximity to tectonic plate boundaries is a key factor, with subduction zones often experiencing high stress as one plate slides beneath another (Schellart & Rawlinson, 2013; Nishikawa & Ide, 2014). Other factors include the speed of plate movement (Conrad et al., 2004), the type and geometry of faults (e.g., normal, reverse, strike-slip) (Kagan, 1992; Schmedes et al., 2010; England, 2018), and the physical properties of the surrounding rock formations (Leary, 1997; Lockner & Beeler, 2002; Scholz, 2002).

Earthquakes are the most frequently observed seismic events. The frequency of seismic waves produced by earthquakes depends on the size and depth of the event. Smaller earthquakes generally produce higher-frequency waves (up to several tens of Hz), while larger, deeper earthquakes generate lower-frequency signals (below 1 Hz). Some of the most significant earthquakes, particularly those that occur at great depths, can be detected globally, with seismometers located thousands of kilometers away recording the seismic waves. These global recordings are known as teleseismogenic events (Engdahl et al., 1998).

Figure 2.3 shows DAS data from an earthquake with a magnitude of $M_w = 1.6$, which occurred 1.4 km from the fiber setup on September 16, 2022, at 11:20:10 UTC in the Hautes-Pyrénées. In Figure 2.3a, the Strain Rate (SR), measured in nstrain/s, clearly shows the arrival of the P-wave at

1 s, followed by the S-wave at 2.2 s. Figure 2.3b presents the trace from a single point on the fiber, while Figure 2.3c displays the corresponding spectrogram, with dominant frequencies below 40 Hz.

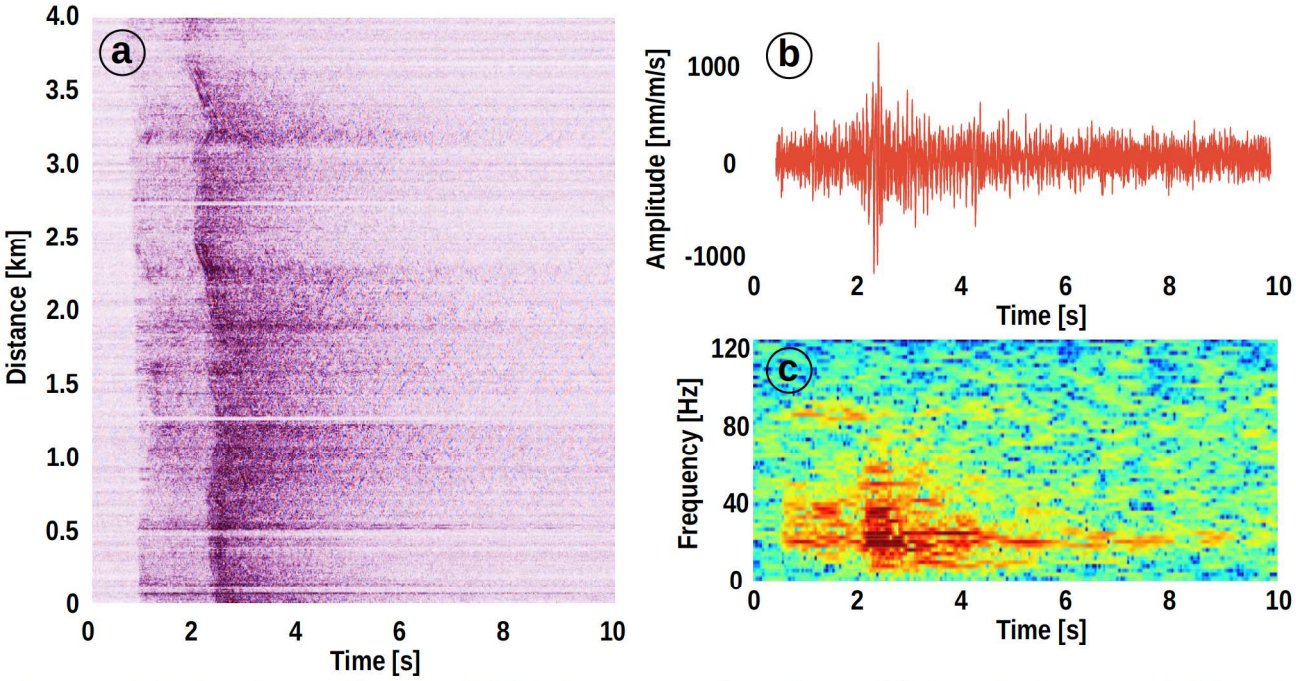


*FIGURE 2.3: Earthquake recorded using FO-DAS instrument. The strain rate **(a)** shows the P-wave arrival (at 1 s) followed by the S-wave (at 2.2 s). We represent a trace taken at one point on the fiber **(b)**, and its spectrogram **(c)**. The recorded event is an earthquake of magnitude Mw=1.6, that happened on on September 16, 2022 at 11:20:10 UTC in the Hautes-Pyrénées.*

Several studies have demonstrated the effectiveness of FO-DAS instrument in detecting various seismic events. FO-DAS instrument has been successfully used to detect earthquakes (Nayak et al., 2021; Li et al., 2021) , microseisms (Williams et al., 2019) and teleseisms (Yu et al., 2019). FO-DAS instrument has also proven useful in estimating earthquake magnitudes and predicting ground motion, as highlighted by Lior et al. (2023).

2.2.2 Volcanic Seismogenic Sources

The heterogeneity of volcanic systems results in a wide range of seismic signals. These signals include Volcano-Tectonic (VT) events, explosive volcanic events, hydroacoustic signals, and tremors, which have been extensively described in the literature (Minakami, 1974; Lahr et al., 1994; McNutt, 2005). Other events, such as landslides and rockfalls, which are not specific to volcanic activity, are also observed but will not be discussed here.

Volcano-Tectonic (VT) events are primarily caused by magma movement or tectonic shifts in the Earth crust. These events are often associated with magma intrusion or rock fracturing (Aki et al., 1977; Ferrick et al., 1982; Koyanagi et al., 1987). Also referred to as shear fractures, rock fracturing generates high-frequency signals typically ranging from 5 to 40 Hz. VT events are crucial for eruption forecasting, as they represent one of the earliest precursors to eruptions (White & McCausland, 2016). Explosive volcanic events are driven by the release of gas, steam, and magma, producing powerful shock waves. These events involve seismic signals with frequencies ranging from 1 to 10 Hz. Several factors influence the seismological features of these events, including the depth of the eruption, the coupling of energy between the ground and air (Dautermann et al., 2009;

De Angelis et al., 2012), and the seismic or acoustic efficiency (Johnson et al., 2003; Johnson & Aster, 2005). Hydroacoustic signals are generated when volcanic activity interacts with water, such as in volcanic lakes or underwater eruptions. These events produce sound waves that propagate through the water medium, making them detectable by specialized hydrophones. Hydroacoustic signals are crucial for monitoring volcanoes located beneath water, where traditional seismological methods may be less effective (Fox & Dziak, 1998). Tremors are low-frequency seismic signals linked to the continuous movement of magma close to the surface, which causes conduit vibrations in the surrounding rock. These signals generally have energy predominantly below 1 Hz. Certain tremors, like eruptive tremors, are often precursors to volcanic eruptions and are critical for volcanic hazard assessment (Scarpa et al., 1996; Sparks, 2003; Segall, 2013).

Several studies have demonstrated the effectiveness of FO-DAS instrument in monitoring volcanic activity. FO-DAS has been used to detect volcano-tectonic sources at shallow depths, such as at Azuma volcano in Japan (Nishimura et al., 2021), and to record microseisms and tremors, with nearly 400 daily events observed at Mt. Meager in British Columbia (Klaasen et al., 2021). FO-DAS instrument can also detect subtle volcanic events, such as fluid migration and degassing, both on inland volcanoes (Jousset et al., 2022) and underwater volcanoes (Caudron et al., 2024). Figure 2.4 shows the strain rate records and spectrum computation of a volcanic explosion (Figure 2.4a,b) and a tremor of a degassing event performed in Jousset et al. (2022). Métaxian et al. (2024) combined FO-DAS instrument with geophones and infrasound to track pyroclastic flows at Stromboli volcano.

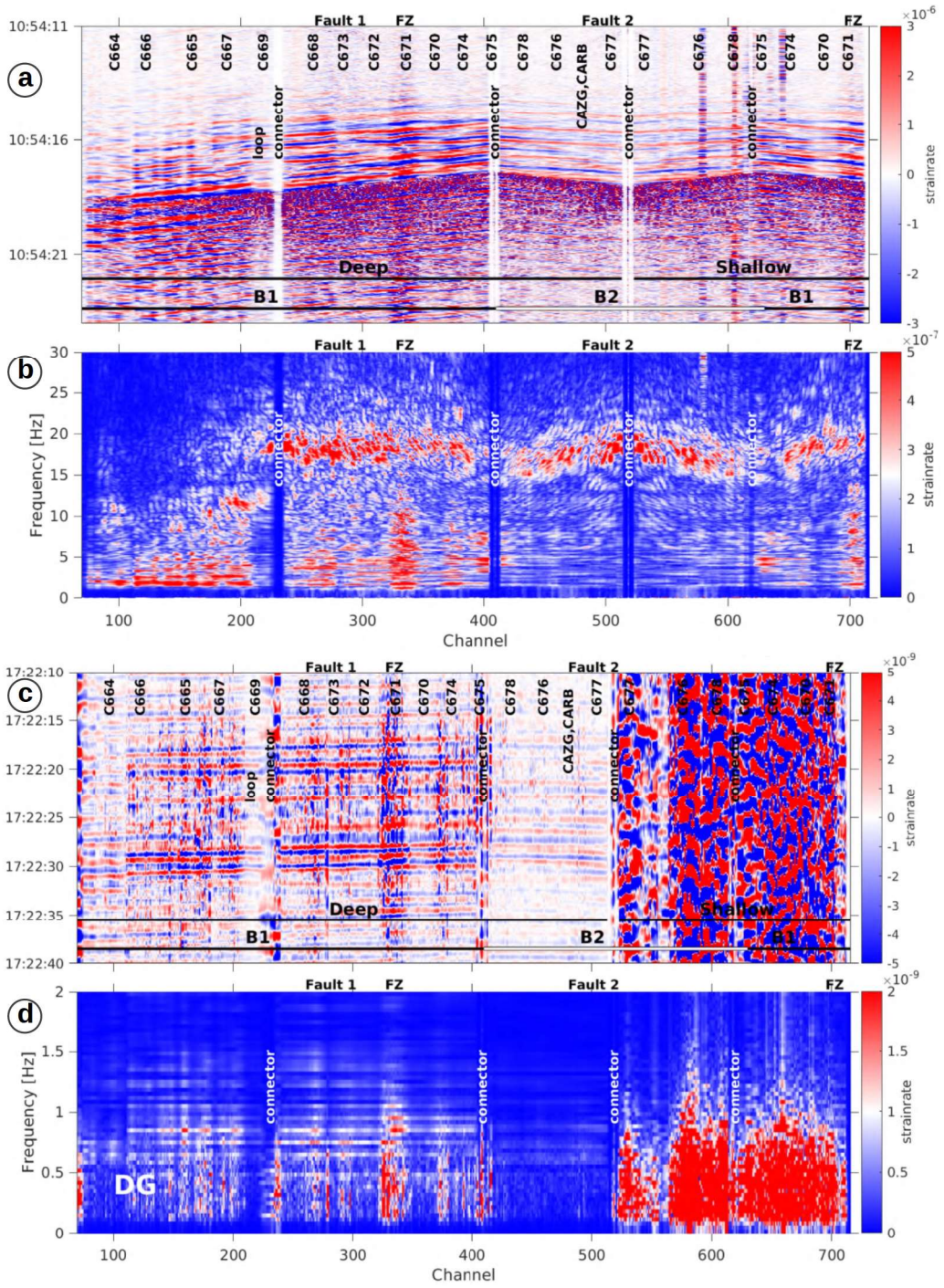


FIGURE 2.4: Volcanic events recorded using FO-DAS instrument. ***(a)*** *and* ***(b)*** *show the strain rate and the spectrum for each channel during an explosion at the New South-East Crater of Mount Etna on September 5, 2018, at 10:54:11 UTC.* ***(c)*** *and* ***(d)*** *display the strain rate and spectrum for each channel during a tremor associated with a degassing event. The setup consisted of a 1.3 km fiber optic cable with 710 FO-DAS channels (taken from Jousset et al. (2022)).*

2.2.3 Environmental Seismogenic Sources

Environmental seismogenic sources are signals generated by geological environmental phenomena, excluding tectonic earthquakes, and events associated with volcanoes. These sources include events such as landslides, glacier movements, and vibrations caused by environmental factors like weather or climate change. For studying large events, existing permanent or semi-permanent programmable seismic networks can be used (Aster & Winberry, 2017; Whiteley et al., 2019). For specific sites, such

as specific landslide, glacier, river or catchment, in-situ monitoring can be performed using portable geophone networks (Teja et al., 2014; Martinez et al., 2017; Provost et al., 2018). FO-DAS instrument is good for both approaches, as it offers continuous, real-time data with high spatial resolution.

2.2.3.1 Landslides

Landslides are mass movements driven by gravity that generate distinctive seismic signals. These events involve large-scale displacement of soil, rock, or debris down a slope, often covering vast areas. Landslides are complex, with materials sliding, flowing, and collapsing at different speeds (Iverson et al., 1997; Hungr et al., 2005). As a result, the seismic signals produced span a wide frequency range, typically dominated by low frequencies between 1 and 30 Hz (Suriñach et al., 2005; Yan et al., 2020), which can be lower than 0.1 Hz in the case of very large landslides (Hibert, Ekström, & Stark, 2017). The duration of landslide signals can vary, lasting from several seconds to minutes (Yan et al., 2020). Provost et al. (2018) identified three primary types of seismic signals associated with landslides: rockfalls, granular flows, and slopequakes (Figure 2.5). Rockfalls are characterized by high-frequency signals (10 to 100 Hz) with distinct, successive impacts visible on both the waveform and spectrogram, typically lasting around 20 seconds (Figure 2.5a). These signals reflect the rapid movement of rocks down a slope. Granular flows generate low and high frequency signals (1 to 100 Hz) with durations ranging from tens to thousands of seconds, indicating the slower movement of debris and soil (Figure 2.5b). Slopequakes produce low-frequency signals (1 to 30 Hz) with short durations (less than 10 seconds), originating from subsurface or deeper sources such as fractures or fluid migration (Figure 2.5c). These events are often triggered by external factors, such as seismogenic activity (Keefer, 1984, 2002; Valagussa et al., 2014), extreme weather conditions like heavy rainfall or freeze-thaw cycles (Take et al., 2004; Helmstetter & Garambois, 2010; Gariano & Guzzetti, 2016; Ran et al., 2018), and human activities like triggered explosions (Vilajosana et al., 2008; Hibert, Ekström, & Stark, 2014). As extreme weather conditions are one of the most frequent causes of landslides, it is interesting to combine seismological measurements with other physical parameters such as electrical resistivity (Hibert et al., 2012).

Although relatively few studies have focused on landslide monitoring using FO-DAS instrument, several applications have emerged. Michlmayr et al. (2017) investigated the use of FO-DAS instrument to detect seismic emissions as precursors to landslides, using a 3-meter setup with adjustable inclination and sprinklers to simulate precipitation on sandy soil. In real-world applications, FO-DAS instrument has been employed to monitor landslide dynamics with nanostrain-rate sensitivity (Ouellet et al., 2024) and track rockfall events (Xie et al., 2023).

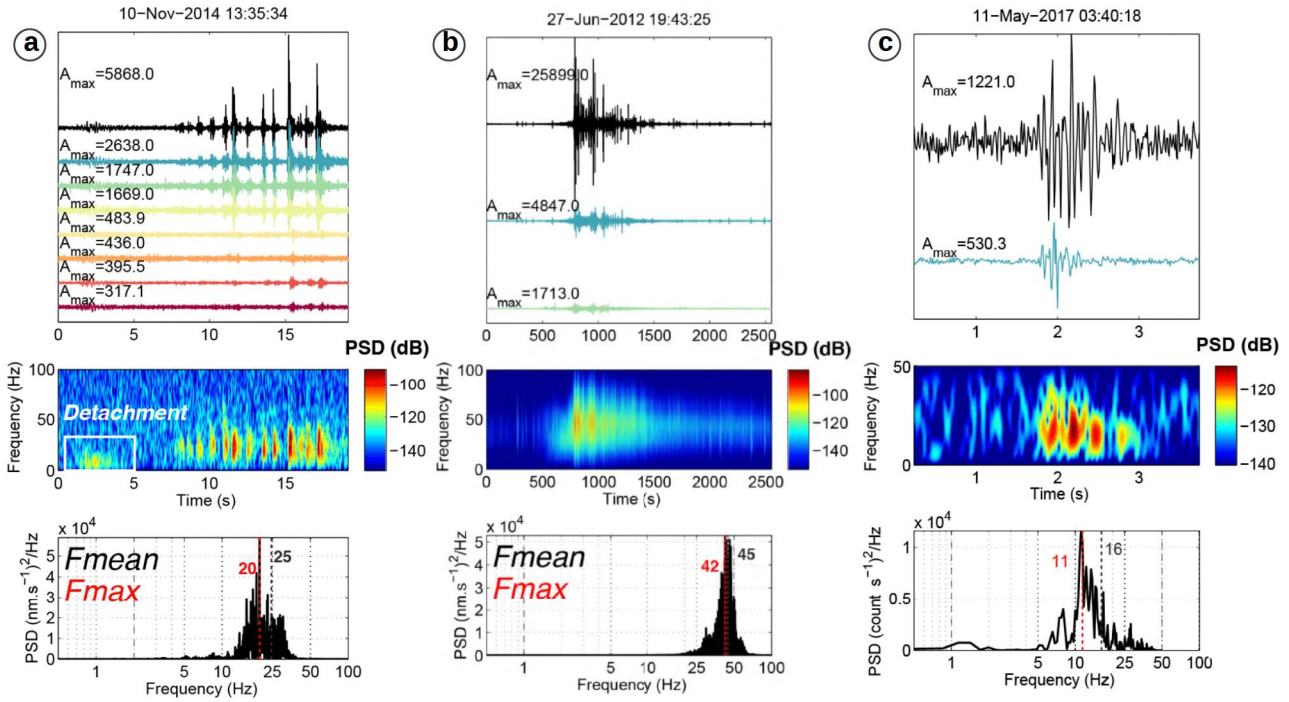


FIGURE 2.5: Landslides-linked events recorded using conventional seismometers. ***(a)*** *shows a rockfall occuring at Super-Sauze, France (Provost et al., 2017),* ***(b)*** *a granular flow occuring at Rebaixader torrent, Espana (Abancó et al., 2012; Arattano et al., 2014; Hürlimann et al., 2014),* ***(c)*** *a slopequake occuring at Pont Bourquin (taken from Provost et al. (2018)).*

2.2.3.2 Glaciers

Seismic signals generated in glaciers come from various sources, including icequakes (Métaxian et al., 2003; West et al., 2010; Röösli et al., 2014), glacier calving (Amundson et al., 2008; Bartholomaus et al., 2012; Canassy et al., 2012), and basal sliding (Lipovsky et al., 2019). Each of these events has distinct characteristics in terms of seismic signal frequency, duration, and amplitude. Icequakes are short-duration, impulsive events caused by the fracturing of ice within a glacier, typically triggered by stress changes due to movement or melting. These events have high-frequency content (2 to 50 Hz) (Figure 2.6a), reflecting the rapid release of energy during ice fracture (Roux et al., 2008; Richardson et al., 2010; Köhler et al., 2019). Glacier calving, when large chunks of ice break off from the glacier terminus and enter the water, generates seismic signals with both low and high-frequency components. The frequency content depends on the size and speed of the calving, with larger events producing more prominent low-frequency signals (Figure 2.6c) (Qamar, 1988; Amundson et al., 2008; Bartholomaus et al., 2012). Basal sliding, where a glacier slides over its bed due to water or sediment, produces low-frequency, continuous seismic signals (Figure 2.6b) that reflect the steady movement of the glacier (Lipovsky et al., 2019).

Conventional seismometers have been used to study glacier seismicity for decades, but FO-DAS instrument offers distinct advantages for glacier monitoring. Traditional seismic coverage in glaciers is often inadequate, especially in regions like Alpine glaciers (Walter et al., 2020), the Greenland ice sheet (Booth et al., 2020), and Antarctica (Hudson et al., 2021). FO-DAS instrument has been employed to detect microseismicity and analyze wave propagation in glaciated areas (Walter et al., 2020). Booth et al. (2020) also demonstrated how FO-DAS instrument can be used for subsurface imaging, mapping the structure of the Greenland ice sheet and identifying sedimentary layers beneath the glacier.

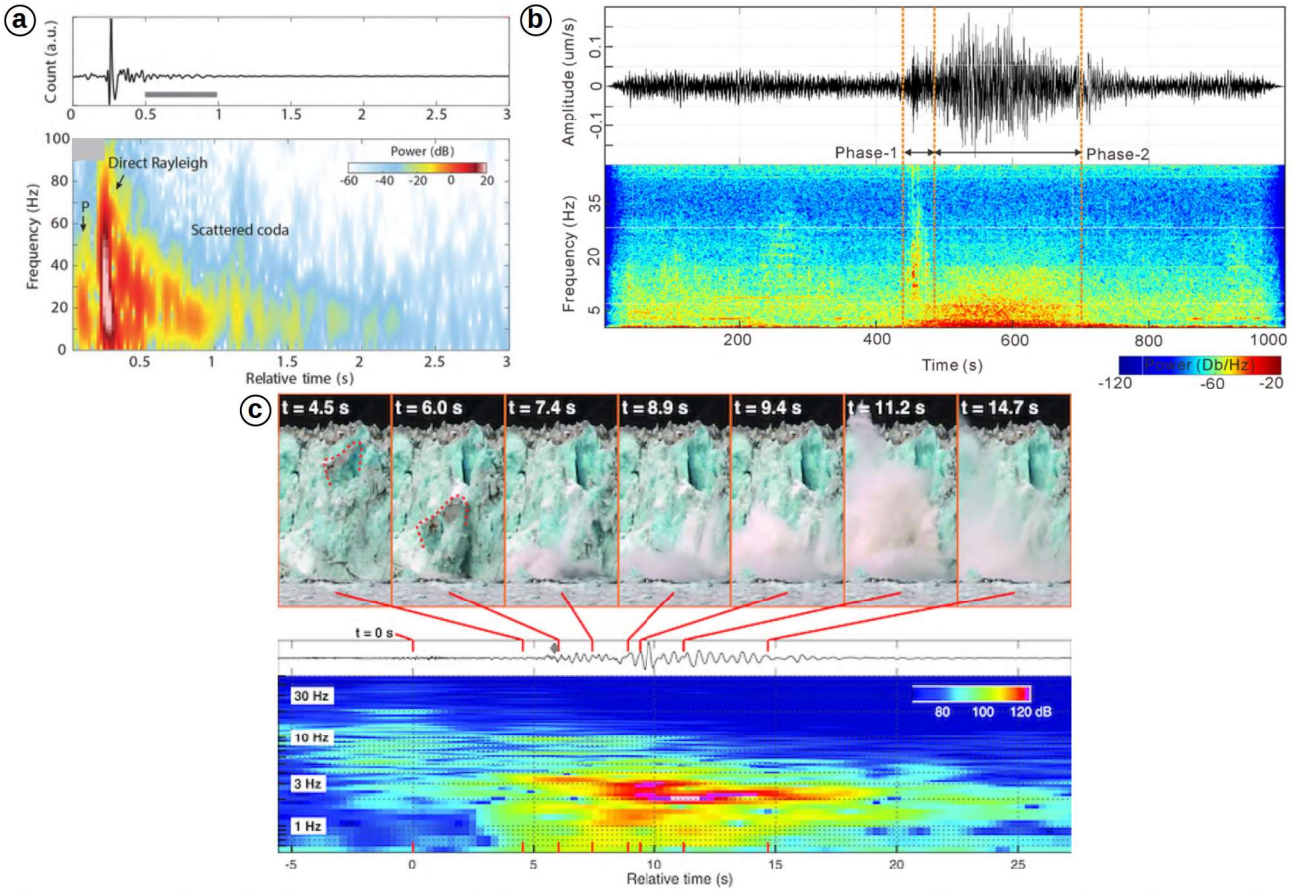


FIGURE 2.6: Glacier-linked events recorded using conventional seismometers. ***(a)*** *shows an icequake occuring at Gornergletscher, Swiss Alps,* ***(b)*** *a basal sliding occuring at Linzhi, Eastern Himalaya,* ***(c)*** *a glacier calving occuring at the terminus of Yahtse Glacier, Alaska (composition of figures from several articles:* ***(a)*** *Sergeant et al. (2020),* ***(b)*** *Li et al. (2024), and* ***(c)*** *Bartholomaus et al. (2012)).*

2.2.3.3 Rivers

Rivers are underground geological formations that store water, and monitoring them requires understanding their properties to assess how they respond to factors like rainfall, drought, water extraction, and contamination. These properties vary depending on the type of rock forming the aquifer, such as porous, or fractured rock (karstic aquifer) (Lee et al., 2006). They also affect how water enters the aquifer during rainfall or ice melt, how it moves through the aquifer, and how much water the aquifer releases (Molson & Frind, 2012). Seismic signals can help monitor these processes. Interest in using seismological methods for aquifer monitoring has grown since the 1990s (Govi et al., 1993), with early research focusing on exploring seismic data. These studies aimed to identify the physical processes linked to seismic signals, examining factors like frequency, amplitude, duration, and wave polarization, as well as interpreting the spatial and temporal patterns of these events (Burtin et al., 2008; Schmandt et al., 2013; Burtin et al., 2016). Some of the processes identified include bedload movement (Govi et al., 1993) and water discharge (Mejías et al., 2012).

We used a DAS interrogator to an installed fiber optic in a karstic cave at Fontenotte, Jura, France, for a period of 14 days, from January 28 to February 12, 2022. Figure 2.7 shows a seismic trace recorded by an underwater section of the fiber optic and the corresponding spectrogram. The envelope of the seismic trace closely follows the volume of water in the aquifer, with a clear increase in amplitude several hours after heavy rainfall on February 2 (day 5), 7 (day 10), and 10 (day 13). The spectrogram reveals a dominant frequency below 10 Hz, which is typical of the low-frequency seismic signals generated by water movement in the aquifer.

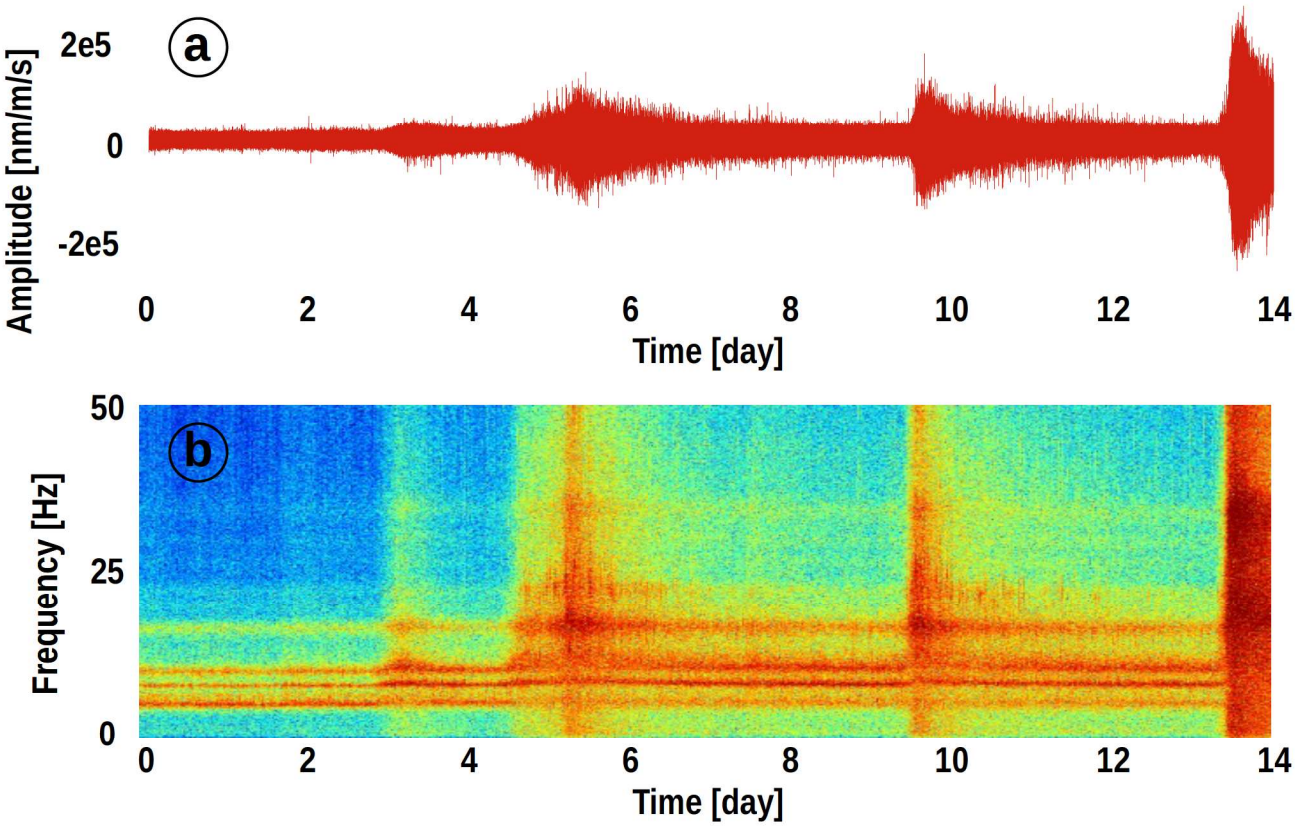


*FIGURE 2.7: Farstic river monitoring using FO-DAS instrument was conducted with a fiber optic cable deployed at Fontenotte cave, Jura, France. The analysis includes a trace recorded at an underwater point on the fiber optic cable over a 14-day period **(a)** and its corresponding spectrogram **(b)**.*

Several studies have explored the use of FO-DAS instrument for aquifer monitoring. Tribaldos and Ajo-Franklin (2021) demonstrated its ability to monitor aquifer dynamics in California using ambient seismic noise. Ajo-Franklin et al. (2019) applied FO-DAS instrument and ambient noise interferometry to map shallow structural profiles and groundwater depth by analyzing surface wave velocity, while Sobolevskaia et al. (2024) used it to predict water table changes in a shallow aquifer in Fairbanks, AK.

2.2.3.4 Meteorological Sources

Seismometers, traditionally used for earthquake detection, are also valuable for monitoring meteorological and climatic conditions such as ocean waves, tsunamis, and storms. Ocean waves, caused by wind interacting with the surface, produce low-amplitude seismic signals that help study the ocean interaction with the lithosphere and monitor atmospheric and oceanic phenomena (0.1 to 5 Hz) (Webb, 1998; Nishida et al., 2013). Tsunamis, generated by seismic events or underwater landslides, produce long-period seismic waves that can be detected to monitor and predict tsunami risks along coastlines (Ward, 1980; Tanioka & Sataka, 1996; Choi et al., 2003; Ward, 2001). Seismometers also capture seismic signals from thunderstorms, where airborne acoustic waves are converted into high-frequency seismic wave (10 Hz to 200 Hz) by interaction with the ground (Kappus & Vernon, 1991). These signals provide information about atmospheric conditions such as temperature, humidity, and wind during storms (Diaz et al., 2023).

Several studies have explored unique applications of FO-DAS instrument in meteorological and climatic condition monitoring. For example, Lin et al. (2024) used FO-DAS instrument to monitor ocean currents during Typhoon Muifa, recording significant changes in current direction and speed,

which enhanced the understanding of atmospheric-ocean interactions. FO-DAS instrument has also been applied to observe thunder-induced ground motions, with Zhu and Stensrud (2019) demonstrating a positive correlation between thunder-seismic power recorded by FO-DAS instrument and lightning current power. Figure 2.8 shows several thunder-induced seismic events recorded during the study.

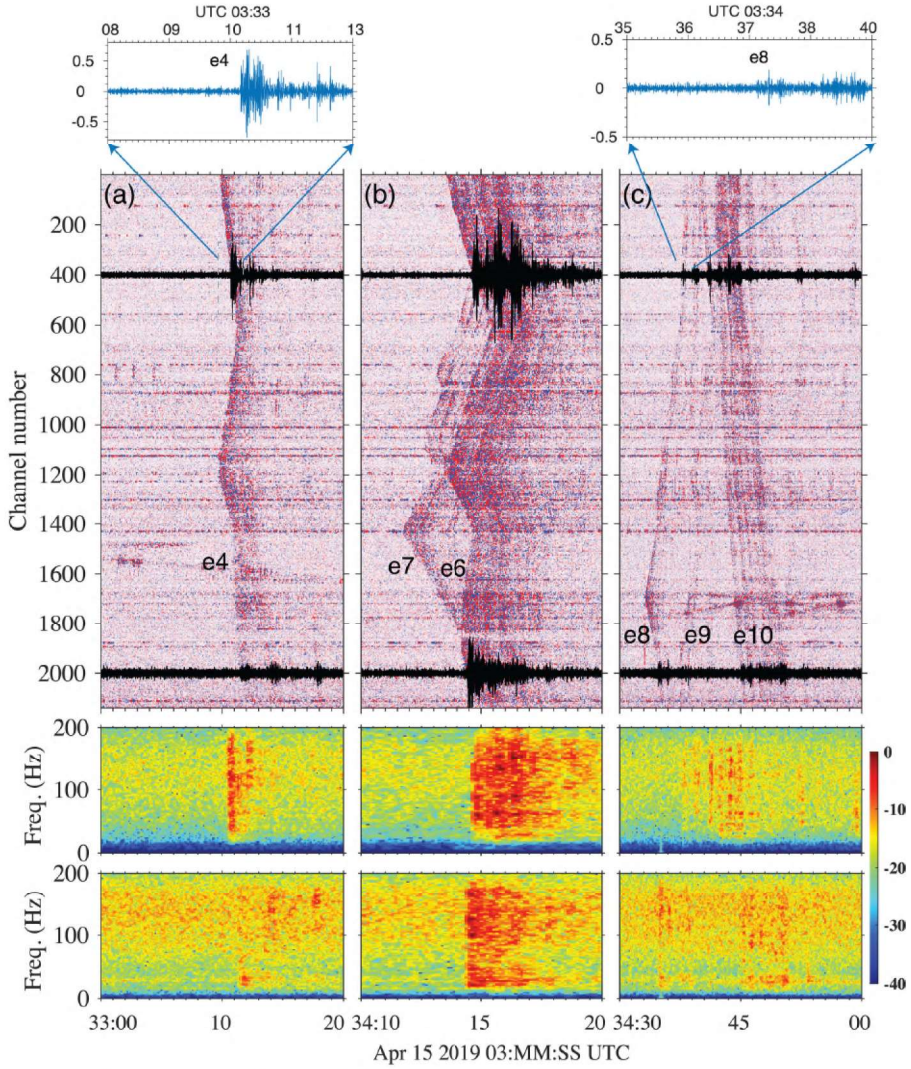


FIGURE 2.8: DAS array observations of several thunder-induced seismic events on 15 April 2019 in State College, PA. Two traces (Channels 400 and 2000) are overlaid. Lower panels show DAS spectrograms (dB) computed for Channel 400 and Channel 2000, respectively (taken from Zhu and Stensrud (2019)).

2.2.4 Anthropogenic Sources

2.2.4.1 Quarry Blasts

Quarry blasting involves controlled explosions used in mining to break rock for material extraction. These blasts generate shock waves that propagate through the ground and can be detected by seismological sensors. The seismic signals are typically short-duration, high-intensity pulses caused by the rapid release of energy from the explosives. The signals are characterized by energy concentrated in low to mid-range frequencies, typically between 0.5 and 50 Hz, with some high-frequency components that distinguish them from natural earthquakes (Allmann et al., 2008; Horasan et al.,

2009). Quarry blasts primarily produce surface waves. As a result, these events are generally only detectable near the epicenter.

Figure 2.9 shows a quarry blast recorded using a FO-DAS instrument. The recorded event is a quarry blast of magnitude Mw=1.1, which occurred on September 8, 2022, at 10:03:36 UTC in the Hautes-Pyrénées. The spectrogram shows a peak in frequency content around 10 Hz. The trace also exhibits a sharp peak at the time of the blast, followed by rapid attenuation, which is consistent with the impulsive nature of quarry blasts.

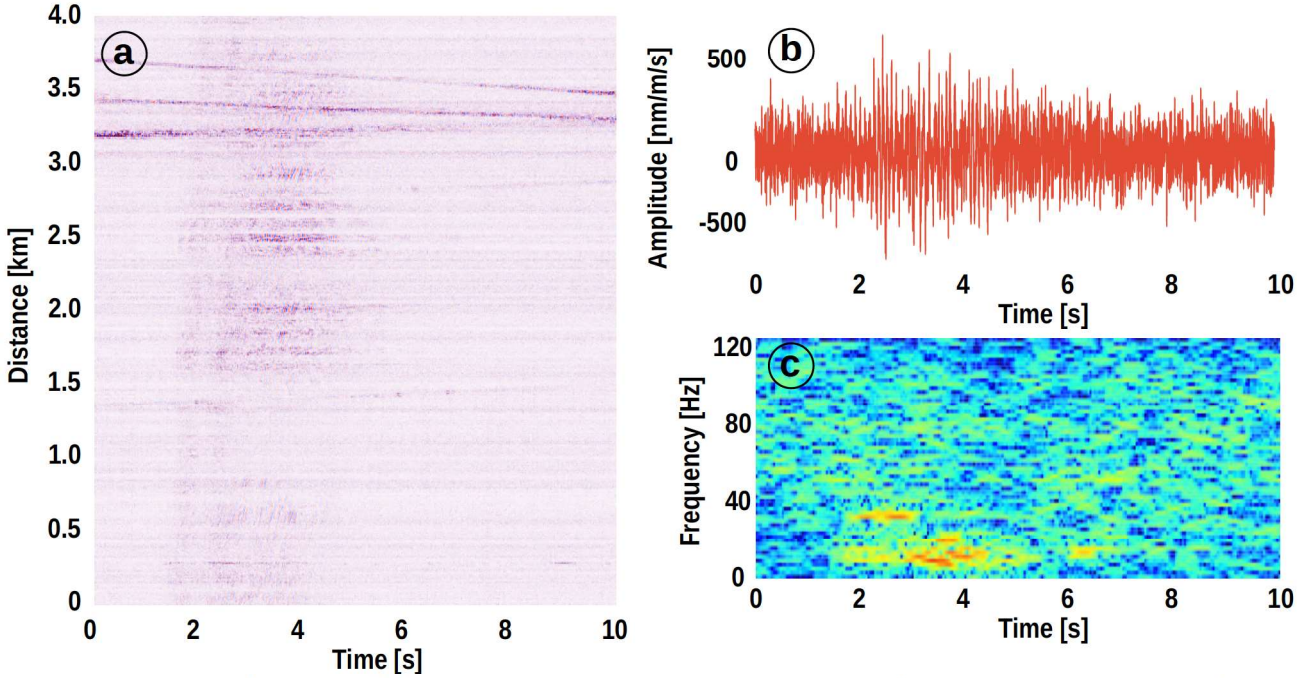


*FIGURE 2.9: Quarry blast recorded using FO-DAS instrument. We represent the strain rate **(a)**, a trace taken at one point on the fiber **(b)**, and its spectrogram **(c)**. The recorded event is a quarry blast of magnitude Mw=1.1, that happened on September 8, 2022 at 10:03:36 UTC in the Hautes-Pyrénées.*

Several studies have demonstrated the potential of FO-DAS instrument for quarry blasts recording in the context of subsurface imaging. Fang et al. (2020) successfully recorded seismic signals from quarry blasts at a distance of 13.3 km using the Fiber Optic Seismic Observatory in an urban setting. Shaheen et al. (2023) utilized a 10 km dark fiber link to record signals from a 25 kg explosive test.

2.2.4.2 Fracking

Hydraulic fracturing (fracking) is a technique used to extract oil and natural gas by injecting high-pressure fluid into rock formations, creating fractures that allow hydrocarbons to flow more easily. This process generates seismic signals due to the sudden release of stress from the rock fractures and the propagation of fluid pressure within the fractured rock. These signals are typically low-magnitude, high-frequency events (10-100 Hz, sometimes up to 200 Hz) that last only a few milliseconds to seconds. Their brief duration, high-frequency nature, and sharp waveform characteristics distinguish them from natural seismogenic events. In some cases, fracking can induce small earthquakes, especially near existing faults, raising concerns about environmental and safety risks. Monitoring these seismic signals is crucial for regulating fracking and minimizing its impact.

Several studies have explored the use of FO-DAS instrument for monitoring microseismicity and subsurface conditions during hydraulic fracturing. Karrenbach et al. (2019) demonstrated the use

of FO-DAS instrument for microseismic monitoring, alongside temperature and strain monitoring using another fiber optic-based sensor called Distributed Temperature and Strain Sensing (DTSS). Kavousi et al. (2017) also utilized FO-DAS instrument for subsurface imaging during hydraulic fracturing in the Marcellus Shale.

2.2.4.3 Ambient Anthropogenic Noise

Ambient anthropogenic noise refers to seismic signals generated by human activities, which can be divided into two main categories: non-mechanical and mechanical sources. Non-mechanical sources, such as walking, manual excavation, and livestock farming, typically produce lower-frequency signals. For example, walking generates frequencies between 1–40 Hz (Houston & McGaffigan, 2003), while manual excavation ranges from 1–10 Hz. Mechanical sources, on the other hand, generate higher-frequency, more energetic signals, often containing both fundamental frequency components due to motor rotation and their harmonics. Trains, for instance, generate frequencies between 1–100 Hz, depending on speed and Doppler effects (Mosleh et al., 2021). Similarly, vehicles passing by produce frequencies in the range of 30–100 Hz, with frequency shifts depending on speed. Helicopters and airplanes produce distinct frequency bands, with helicopters generating a fundamental around 20–30 Hz and airplanes around 60–100 Hz, both influenced by their movement. Wind turbines, depending on their speed, generate frequencies between 20–30 Hz, also affected by movement (Díaz et al., 2022). Mechanical sources like excavation machines and hydraulic hammers produce frequencies between 10–500 Hz (Cao et al., 2016), with electric hammers generating broadband, repetitive signals. These frequency ranges are key to distinguishing anthropogenic signals from natural seismogenic events (Díaz et al., 2022).

As an example, the Figure 2.10 shows a moving vehicle recorded using FO-DAS instrument. The event is recorded on September 8, 2022 in the Hautes-Pyrénées. The strain rate, trace, and spectrogram of the event are shown in the figure.

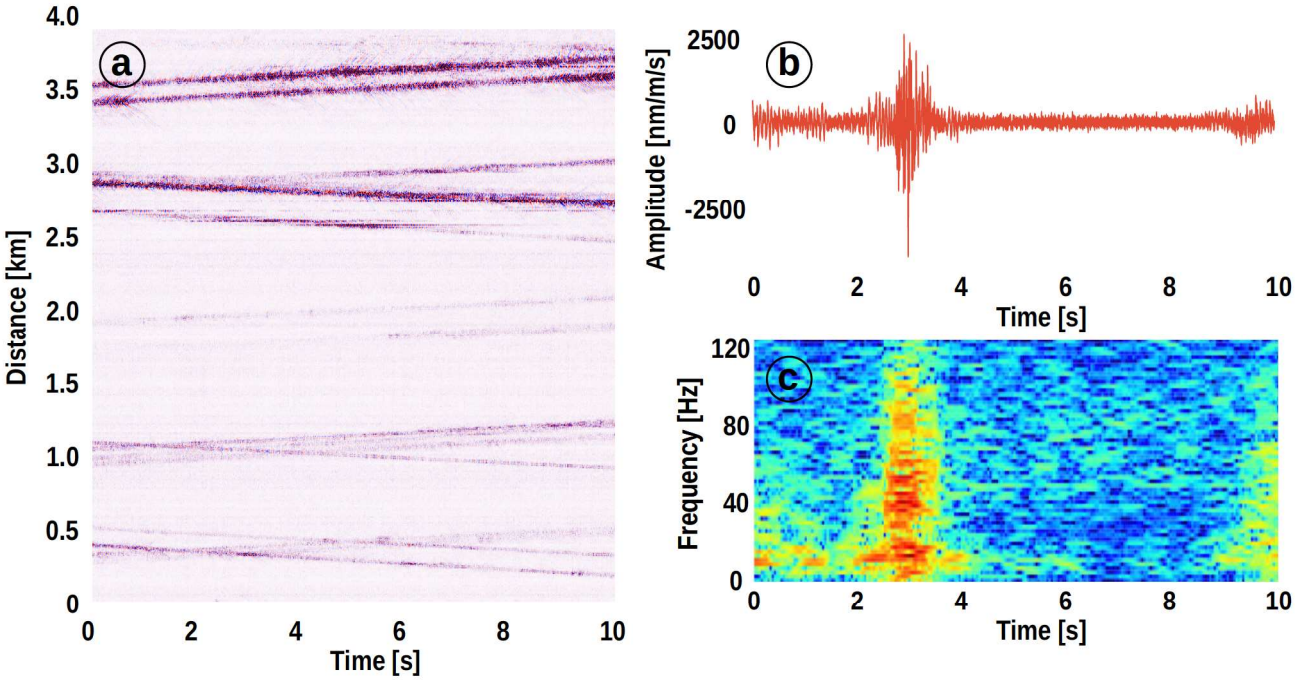


*FIGURE 2.10: Vehicles recorded using FO-DAS instrument. We represent the strain rate **(a)**, a trace taken at one point on the fiber **(b)**, and its spectrogram **(c)**. The recorded event contains several moving vehicles, measured on September 8, 2022 in the Hautes-Pyrénées.*

Several studies have explored the use of FO-DAS instrument for monitoring transportation and security-related activities. Liu et al. (2020) demonstrated its application in detecting and classifying vehicles, estimating speed, and managing traffic in real-time. FO-DAS instrument has also been used for tracking train position and speed (Peng et al., 2014; Kowarik et al., 2020), detecting perimeter intrusions for security purposes (Yu et al., 2015), and tracking pedestrian movements for crowd space identification (Cai et al., 2021). Additionally, FO-DAS instrument has been applied for passive imaging using common anthropogenic sources such as traffic noise (Dou et al., 2017) and other seismogenic activities (Lellouch et al., 2019; Nziengui-Bâ et al., 2023).

2.3 Detectability of Seismogenic Sources with DAS Instrument

The previous section outlined the various potential applications of FO-DAS instrument, including earthquake and volcano monitoring, as well as the tracking of environmental and anthropogenic events. However, using FO-DAS instrument for these purposes comes with its own set of challenges. Several studies have highlighted both the advantages (pros) and limitations (cons) of employing FO-DAS instruments for seismic event detection:

Pros

- **High Spatial Resolution and Sensitivity**
  FO-DAS instruments provides a high spatial resolution and a good seismic signal sensitivity, making them very good for detecting microseismic events. In earthquake monitoring, FO-DAS instrument has been used for detecting microseisms and teleseism signals (Williams et al., 2019). FO-DAS instrument is also effective in volcano monitoring for detecting microseisms and tremors (Klaasen et al., 2021; Jousset et al., 2022), in landslide early warning systems (Michlmayr et al., 2017; Ouellet et al., 2024), and in glacier monitoring to detect microseismicity (Walter et al., 2020). FO-DAS instrument has also been employed for aquifer monitoring in rivers (Ajo-Franklin et al., 2019).

- **Reuse of Telecom Fiber Optics in a Different Application Context**
  In modern civil construction, fiber optic cables are commonly installed along or within structures such as dams, bridges, and pipelines to facilitate rapid telecommunication over long distances. These cables often contain multiple fibers for redundancy, ensuring that if one fiber encounters an issue, others remain functional. One of these fibers can therefore be used for seismic monitoring. For earthquake monitoring, FO-DAS instruments have been employed to quickly detect aftershocks (Li et al., 2021). In volcanic regions, these instruments are also effective for monitoring pyroclastic flows (Métaxian et al., 2024). Additionally, fibers not originally intended for seismic purposes can be used for short-term measurements, such as monitoring thunder impacts within a brief half-hour period (Zhu & Stensrud, 2019).

- **Deployment or Reuse of Fiber Optics in Hard-to-Reach Areas**
  FO-DAS instrument is particularly valuable for monitoring in hard-to-reach or remote areas where traditional seismic sensors may be difficult to deploy. For example, submarine dark fiber has been used for earthquake monitoring (Williams et al., 2019), and for monitoring submarine volcanoes (Caudron et al., 2024). In glacier monitoring, FO-DAS instrument can cover long distances with a single device, reducing costs and simplifying maintenance (Walter et al., 2020).

Additionally, FO-DAS instruments have been successfully deployed for low-cost monitoring in hard-to-reach locations, such as aquifer studies in rivers (Tribaldos & Ajo-Franklin, 2021; Ajo-Franklin et al., 2019), fracking wells (Kavousi et al., 2017; Karrenbach et al., 2019), and ocean monitoring for meteorological conditions (Lin et al., 2024).

- **Resistance of Fiber Optic to Hard Conditions**
  FO-DAS instruments are resistant to harsh environmental conditions, making them well-suited for monitoring in extreme environments. For instance, FO-DAS instrument is used to track pyroclastic flows in volcanic regions (Métaxian et al., 2024) and to monitor the passage of typhoons and thunderstorms, where the conditions can be particularly challenging (Lin et al., 2024; Zhu & Stensrud, 2019).

Cons

- **Directional Sensitivity**
  FO-DAS instruments are sensitive to vibrations in only one direction, which means careful planning of the fiber trajectory is crucial to optimize data collection. For example, in glaciers, the fiber trajectory must be carefully planned to account for directional sensitivity (Hudson et al., 2021). Additionally, the fiber installation can be more complex in other applications, such as fracking, where the fiber must be arranged in a straight line or with multiple loops to maximize data capture (Lindsey et al., 2017; Martin et al., 2018; Huot et al., 2018), and in helical configurations, especially for Vertical Seismic Profiling (VSP) (Kuvshinov, 2016; Wuestefeld & Wilks, 2019).

- **Data Volume**
  The large amount of data generated by FO-DAS instruments, particularly during long-term monitoring, can be a significant challenge in terms of storage and processing. In glaciers, for example, the volume of data can overwhelm processing systems in case of long-term monitoring (Hudson et al., 2021). For aquifer monitoring, FO-DAS instrument has been used to record passive seismic data over extended periods, such as 7 months of data from a 27 km fiber section (Ajo-Franklin et al., 2019) and 5 months of data from a 23 km fiber stretch (Tribaldos & Ajo-Franklin, 2021).

- **Fiber Optic Coupling with Ground**
  The coupling of the fiber optic cable with the ground is crucial for accurate data collection. Poor coupling can lead to signal loss and reduced sensitivity, particularly in applications like microseismicity detection. For example, for glacier monitoring, the fiber must be in direct contact with the ice (Hudson et al., 2021).

- **Low Signal-to-Noise Ratio (SNR)**
  FO-DAS instruments can struggle with low SNR compared to the use of geophones , particularly in environments with weak seismic signals or when multiple sources produce similar signals. Early FO-DAS instruments faced significant SNR challenges when compared to traditional geophones, with differences in the range of 40-50 dB (Daley et al., 2013). In glaciers, distinguishing between seismic events like icequakes and basal sliding is difficult due to the low SNR in these environments (Hudson et al., 2021).

FO-DAS instruments have demonstrated significant potential for monitoring a wide variety of seismogenic sources, ranging from earthquakes and volcanic activity to environmental and anthropogenic events. However, challenges such as directional sensitivity, data volume, and low SNR still

persist. Overcoming these obstacles is essential to fully leverage the capabilities of FO-DAS instruments in seismic applications. In the following sections, we will delve into how advanced data processing techniques can address the noise issues in DAS data and explore how these instruments can be used to analyze and characterize seismic events effectively.

2.4 Data Processing Technique for DAS

Data processing techniques are essential for extracting information from raw data in seismology. These techniques serve two main purposes: identifying seismogenic sources and characterizing subsurface properties. The first focuses on detecting and monitoring areas where seismic events originate, aiming to understand seismic activity, its causes, and earthquake dynamics. The second involves creating detailed images of underground structures, such as rock layers, faults, and fluid reservoirs, for applications like resource exploration and geological studies. Despite their different focuses, both rely on similar data processing techniques, such as instrumental noise filtering, signal stacking, STA/LTA (Short-Term Average Over Long-Term Average), template matching, and Fourier transform-based tools.

2.4.1 Data Pre-Processing: Instrumental Noise Filtering

Instrumental noise in DAS data arises from external environmental factors that disrupt the normal functioning of the instrument, including electrical fluctuations and equipment vibrations (Hartog, 2017). Since FO-DAS instruments are opto-electronic and require an electrical outlet, power fluctuations like voltage spikes, surges, or dips can lead to instrument malfunctions and data loss. Using power conditioning equipment, like voltage stabilizers, helps stabilize the power supply. Equipment vibrations from nearby machinery or vehicles can also interfere with the aquisitions. A proper installation of the DAS interrogator using vibration isolation techniques and stable mounting structures help to minimize this effect. Instrumental noise is problematic in seismological applications where weak signals must be detected (Dou et al., 2017; Ajo-Franklin et al., 2019). The most common instrumental noise that affects DAS aquisition is common-mode noise, which occurs when sound and vibrations in the vicinity of the DAS interrogator simultaneously influence all data channels.

Several techniques exist for filtering instrumental noise, including digital filtering, instrumental noise deconvolution, and common-mode noise suppression. In the case of DAS aquisition, the most commonly used method is **common-mode noise suppression**. Since the instrumental noise often affects the entire fiber simultaneously at certain time during aquisition, this technique involves subtracting the common noise from all data channels. The suppression of common-mode frequency components is computed by averaging over all channels and at each time the SR, then subtracting it to the measured SR (Hartog, 2017). This method is particularly effective in reducing background noise caused by mechanical or thermal vibrations and is quick to implement, making it a popular choice for DAS data processing. An example of common-mode noise removal is shown in Figure 2.11, where we compare the strain rate measured from FO-DAS instrument before and after common-mode removal.

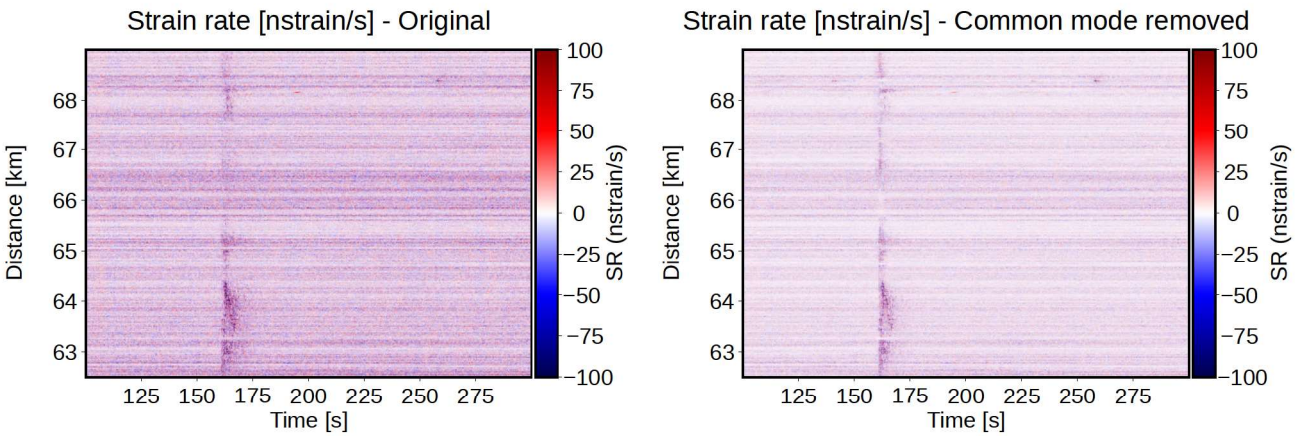


FIGURE 2.11: Example of Common Mode Removal, on a seismic signal recorded using FO-DAS instrument. The figure shows the strain rate directly mesured from FO-DAS instrument (left) and the strain rate after common mode removal (right). The recorded event is a small earthquake of magnitude Mw=1.3, that happened on September 20, 2022 at 04:39:26 UTC in the Hautes-Pyrénées.

2.4.2 Data Exploration Methods

The second type of undesirable noise, referred to as background noise, arises from unwanted signals caused by environmental factors (e.g. rain, storms) or unrelated sources (e.g. moving vehicles in the case of earthquake monitoring). It poses particular challenges for detecting low-amplitude seismogenic events, such as small earthquakes, tremors, or other subtle ground motions. To address this, several tools exist for conventional seismological monitoring and DAS applications. We present in this section temporal analysis methods such as Short-Term Average to Long-Term Average (STA/LTA), stacking technique, filtering methods, and template-matching method.

- **STA/LTA** technique is widely used in conventional seismology for detecting seismogenic events in noisy data (Allen, 1982; Roberts et al., 1989; Withers et al., 1998; Trnkoczy, 2009). This method involves comparing the average signal over a short temporal window to the average signal over a much longer temporal window. The key idea behind STA/LTA is that when a seismogenic event occurs, it generates a transient signal that causes a sudden increase in the short-term average, making it stand out from the long-term average, which represents the background noise. The result is a STA/LTA ratio above 1. Similarly, when the earthquake coda finishes propagating through the sensor, the ratio temporally decrease below 1. As STA/LTA emphasizes the relative change in signal strength over time, the technique is particularly useful for detecting low-amplitude events that may be masked by ambient noise. An example of STA/LTA applied to a seismic signal recorded using FO-DAS instrument is shown in Figure 2.12.

- **Signal stacking** technique involves collecting multiple seismic traces of the same event from different seismological stations in case of use of geophones or from different channels in case of DAS aquisition and averaging them together to create a stacked signal (Schimmel & Paulssen, 1997; Flanagan & Shearer, 1998). Random noise, which is neither spatially nor temporally coherent, tends to cancel out when averaged, while the true seismic signal, being consistent across multiple traces, adds constructively. By stacking signals, the reliability of event detection is improved, and the seismogenic source can be characterized more accurately, even in noisy conditions. This technique is particularly useful for detecting weak or low-amplitude events, such as small earthquakes or environmental signals, which might otherwise be masked by background noise.

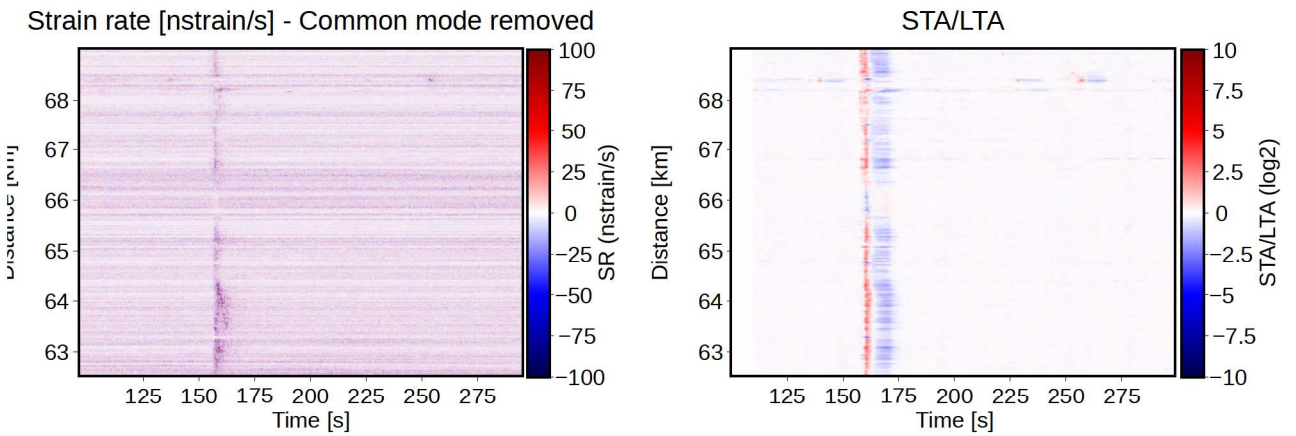


FIGURE 2.12: Example of STA/LTA Computation, on a seismic signal recorded using FO-DAS instrument. The figure shows the strain rate processed using common mode removal (left) and its STA/LTA (right). The recorded event is a small earthquake of magnitude Mw=1.3, that happened on September 20, 2022 at 04:39:26 UTC in the Hautes-Pyrénées.

- **Filtering** techniques aim to improve the quality of seismic data by isolating specific frequency ranges to eliminate unwanted noise (Dziewonski et al., 1969). A low-pass filter allows frequencies below a certain cutoff value to pass through, effectively removing high-frequency noise that is typically associated with high-frequency environmental disturbances, such as wind or anthropogenic activities. Conversely, a high-pass filter removes low-frequency signals, which are often caused by long-period instrumental noise or slow temperature variations in the fiber. This technique can also be used prior STA/LTA or stacking, as it helps to enhance the signal-to-noise ratio of each individual seismic traces (Withers et al., 1998). Figure 2.13 shows an example of a band-pass filter [5,30] Hz applied to a seismic signal recorded using FO-DAS instrument.

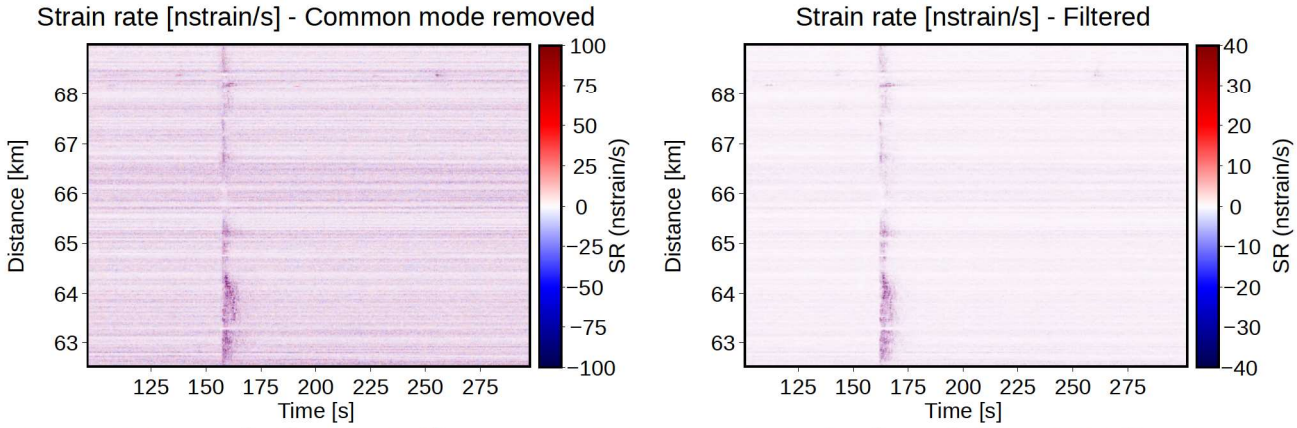


FIGURE 2.13: Example of strain rate filtering computation, on a seismic signal recorded using FO-DAS instrument. The strain rate is filtered using a band-pass filter [5,30] Hz. The figure shows the strain rate processed using common mode removal (left) and its STA/LTA (right). The recorded event is a small earthquake of magnitude Mw=1.3, that happened on September 20, 2022 at 04:39:26 UTC in the Hautes-Pyrénées.

- **Template matching** techniques involve comparing a seismic signal with a reference seismic trace template in order to identify similar patterns. The reference template can be the signal itself, in which case the technique is called auto-correlation, or it can be another signal, in which case it is referred to as cross-correlation. Auto-correlation involves computing the correlation of a seismic trace with itself over various time lags. Auto-correlation is useful for identifying repeating patterns in the data, such as tremors, or for assessing the asymmetry of the signal. In the case of an earthquake, a coda wave persist after the main shock caused by

body waves and surface waves arrivals, often characterized by lower frequencies and longer durations (Aki & Chouet, 1975; Su et al., 1996).

Frequency-based event detection methods leverage the spectral properties of the signals to distinguish seismogenic events from background noise and other sources of interference. Among the most commonly used methods are the spectrum and representations computed from the spectrum like spectrogram or energy band.

- The **spectrum** is a representation of the seismic signal that shows how its power is spread across different frequencies. The spectrum helps to identify key frequency components, harmonic components and patterns that characterize specific seismogenic sources, and can be used to distinguish between various events like earthquakes, quarry blasts, or noise. The spectrum can be computed directly using the Fourier transform, and it can be normalized by the frequency bin size to yield the Power Spectral Density (PSD). Figure 2.14 shows an example of the PSD applied to a seismic signal recorded using FO-DAS instrument, with the spectrum computed for each position along the fiber.

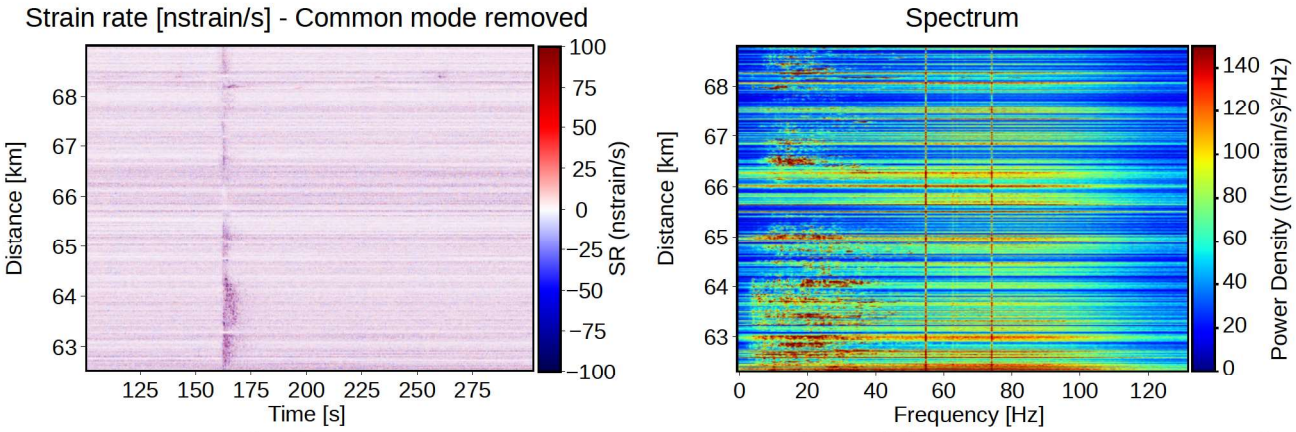


FIGURE 2.14: Example of spectrum computation, on a seismic signal recorded using FO-DAS instrument. The figure shows the strain rate processed using common mode removal (left) and the module of the spectrum (right). The recorded event is a small earthquake of magnitude Mw=1.3, that happened on September 20, 2022 at 04:39:26 UTC in the Hautes-Pyrénées.

- The **spectrogram** is a time-frequency representation that shows how the frequency content of a signal changes over time, making it useful for detecting transient signals, such as earthquakes or impulsive events, and tracking their evolution. It helps identify the onset, duration, and frequency content of seismogenic events, providing valuable information about the nature of the source. Examples of spectrogram associated with seismic signals recorded using FO-DAS instrument were shown in previous section for an earthquake (Figure 2.3c), a quarry blast (Figure 2.9c), moving vehicles (Figure 2.10c) and charge-discharge of an aquifer (Figure 2.7c).

- The **Energy Band (EB)** of a seismic signal refers to the integration of the spectrum computed for a fixed window size over a specific frequency range. The EB analysis helps in characterizing the signal behavior, detecting specific sources of energy, and distinguishing between various seismic phenomena based on their frequency content. Figure 2.15 shows an example of EB computation applied to a seismic signal recorded using FO-DAS instrument.

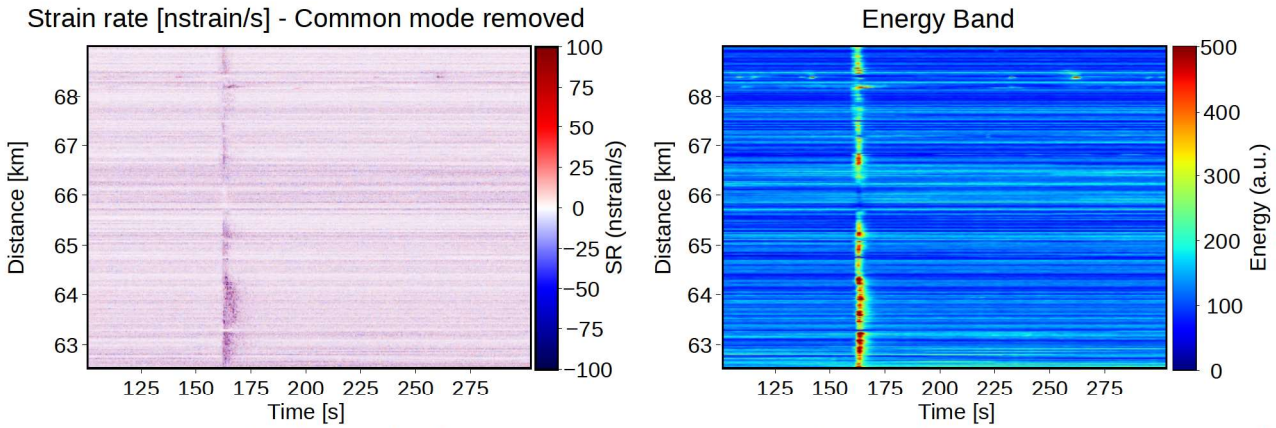


FIGURE 2.15: Example of energy band computation, on a seismic signal recorded using FO-DAS instrument. The chosen frequency band is [5,100] Hz. The figure shows the strain rate processed using common mode removal (left) and its STA/LTA (right). The recorded event is a small earthquake of magnitude Mw=1.3, that happened on September 20, 2022 at 04:39:26 UTC in the Hautes-Pyrénées.

Data pre-processing techniques, like common-mode noise removal, remove unwanted instrument-induced noise. These computationally efficient methods improve seismic data quality without the need for user-defined parameters, providing a straightforward solution for noise reduction. Conventional data exploration tools, such as STA/LTA, stacking, filtering, template matching, spectrum analysis, spectrograms, and energy bands, help detect and characterize seismogenic events. However, these tools are often insufficient on their own and typically require combination with other techniques to achieve reliable and accurate results. Additionally, they often demand fine-tuning of parameters, such as selecting the optimal frequency bands for energy band representation. Artificial Intelligence (AI) can address these challenges by integrating multiple conventional data exploration tools and allowing the use of various user-defined parameters, a concept known as "features" in AI. This enables a more flexible, automated, and efficient approach to seismic data analysis, improving accuracy and minimizing the need for manual intervention.

2.5 Artificial Intelligence for Seismic Data Processing

This section is divided into two parts: the first provides an overview of key machine learning concepts, terminology, and techniques that are particularly relevant for seismic data analysis. These foundational principles form the basis for understanding how AI models can be trained and used for different purposes such as detection, classification or dataset exploration. The second section focuses on the specific applications of AI in processing seismic data measured with conventional seismometers and FO-DAS instrument.

2.5.1 Background Information on Artificial Intelligence

2.5.1.1 Introduction to Artificial Intelligence: Concepts of Features and Latent Space

Artificial Intelligence (AI) focuses on creating statistical models that enable computers to learn from datasets. AI algorithms identify patterns within the data, learn from them, and make predictions or decisions based on that learning. This process relies on an intermediate representation called the latent space, which is constructed from tens to thousands features depending to the complexity of the problem. These features can either be human-engineered, in which case the technique

is called **Machine Learning (ML)**, or automatically learned from the data, in which case we refer to the technique as **Deep Learning (DL)**. For example, ML techniques are often used to predict house prices, with typical features such as the number of bedrooms, square footage, and location. DL, a revolutionary method in AI (LeCun et al., 2015), provides an automatic solution to the challenge of constructing latent spaces. The core idea behind DL is the use of deep neural networks—models with multiple layers of neurons. These networks typically consist of an input layer (which presents the raw data), several hidden layers (which transform the data), and an output layer (which generates the final result depending on the learning task). The hidden layers perform transformations on the data, progressively calculating increasingly relevant features for both the data representation and the task, effectively constructing a meaningful latent space. However, DL faces challenges, such as the large amount of data required for training and the difficulty in interpreting the resulting latent space.

2.5.1.2 Types of Artificial Intelligence Algorithms

AI tackles several learning tasks, which generally fall into four categories: prediction, exploration, interaction, and self-improvement (Figure 2.16).

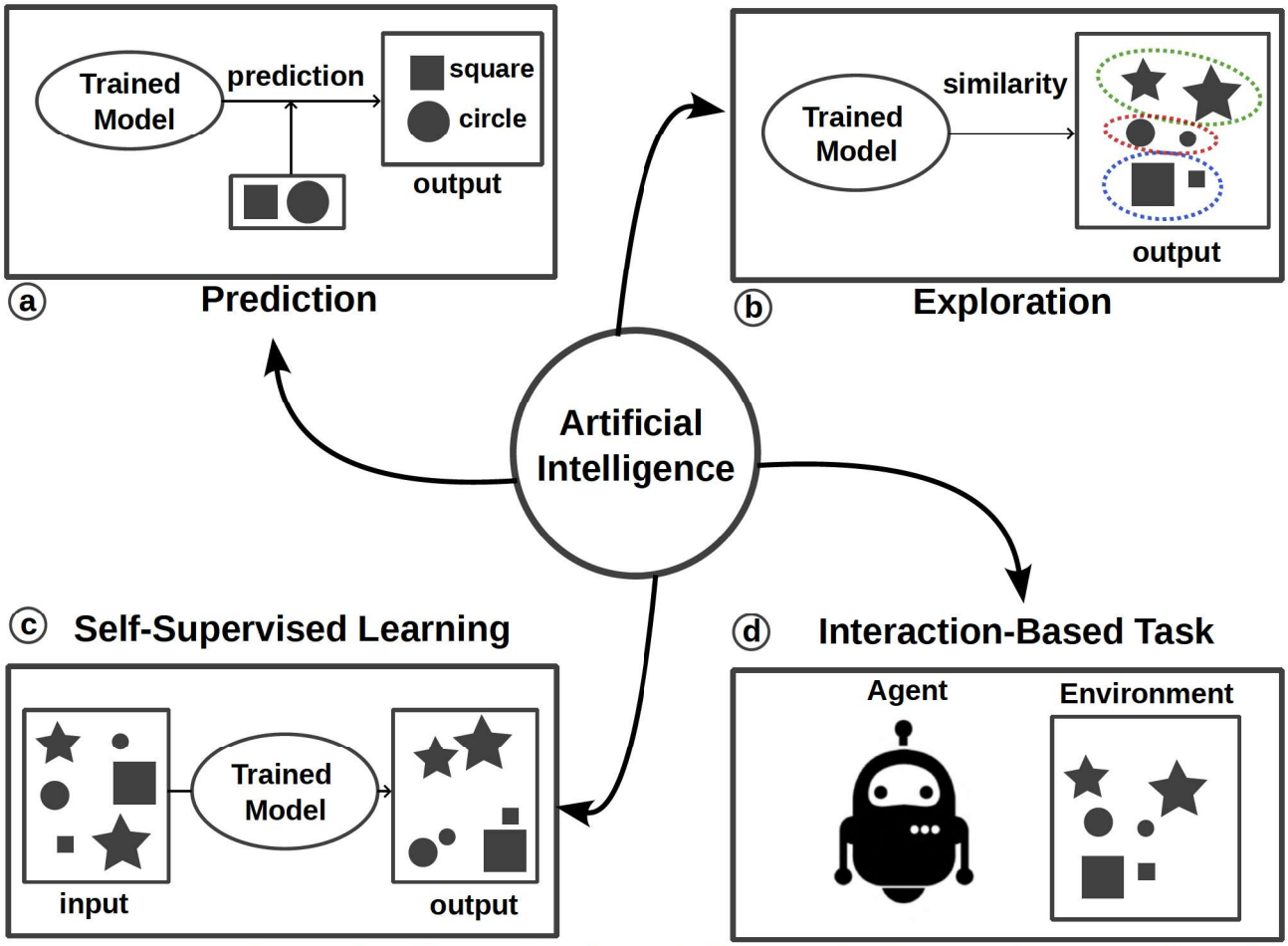


*Figure 2.16: Type of Artificial Intelligence algorithms. The four main families aims **(a)** prediction, **(b)** exploration, **(c)** self-supervised learning, and **(d)** interaction-based task.*

- In predictive tasks, algorithms learn from labeled data to make decisions about new, unseen data. **Supervised learning** is commonly used for prediction task, like regression (predict a value) or classification (predict a label) (Figure 2.17a). Several common algorithms include

linear regression (Seber & Lee, 2012) for regression task, and Support Vector Machines (SVM) (Burges, 1998; Hearst et al., 1998) and Random Forest (RF) (Ho, 1995; Breiman, 2001) for classification task. SVM and RF can also perform regression tasks using variation of these algorithm (Drucker et al., 1996; Ho, 1998; Meinshausen & Ridgeway, 2006). **Semi-supervised learning** combines labeled and unlabeled data for learning (Zhu, 2005; Zhu & Goldberg, 2022) (Figure 2.17b), while **transfer learning** leverages pre-trained models to adapt knowledge from one domain to another, reducing the need for labeled data (Pan & Yang, 2009) (Figure 2.17c).

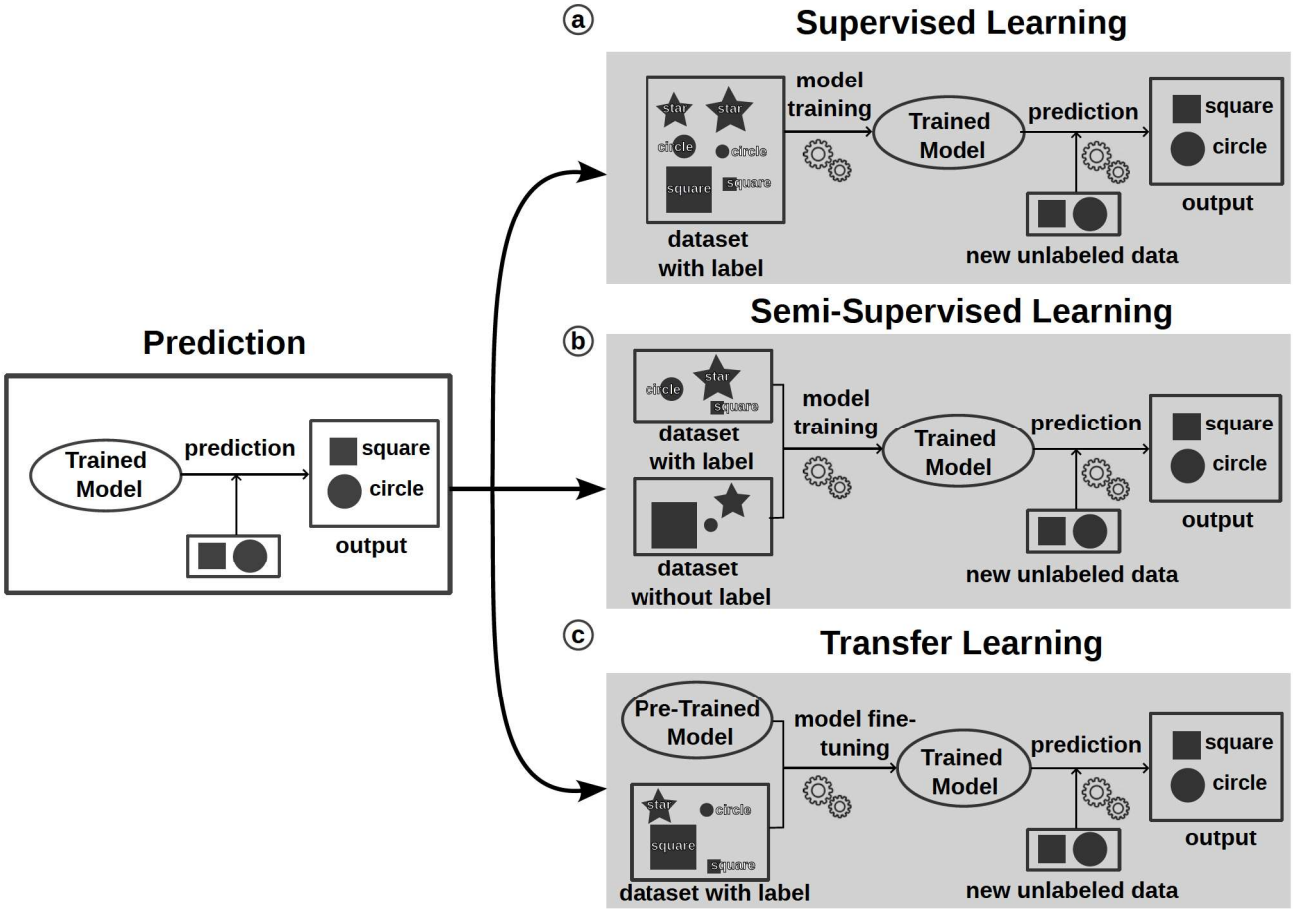


*FIGURE 2.17: Artificial Intelligence algorithms for prediction. Prediction can be performed using **(a)** supervised learning, **(b)** semi-supervised learning, and **(c)** transfer learning.*

- For exploration, **unsupervised learning** methods identify hidden patterns in data without explicit labels (Figure 2.18a). Clustering techniques are widely used for grouping data based on similarities or differences, with methods like exclusive clustering (e.g., K-means) (Forgy, 1965; Hartigan & Wong, 1979), fuzzy clustering (Gath & Geva, 1989), hierarchical clustering (e.g., agglomerative clustering) (Johnson, 1967; Murtagh & Legendre, 2014), and probabilistic clustering (e.g., Expectation-Maximization) (Dempster et al., 1977; Kriegel et al., 2011). Density-based clustering, such as DBSCAN, forms clusters based on data point density, separating outliers as noise (Ester et al., 1996). **Dimensionality reduction** methods, such as Principal Component Analysis (PCA) (Wold et al., 1987; Jolliffe & Cadima, 2016), t-SNE (Van der Maaten & Hinton, 2008), and autoencoders (Hinton & Salakhutdinov, 2006), simplify the data latent space representation by reducing the number of dimensions of latent space while preserving essential features, aiding in data exploration and visualization (Figure 2.18b).

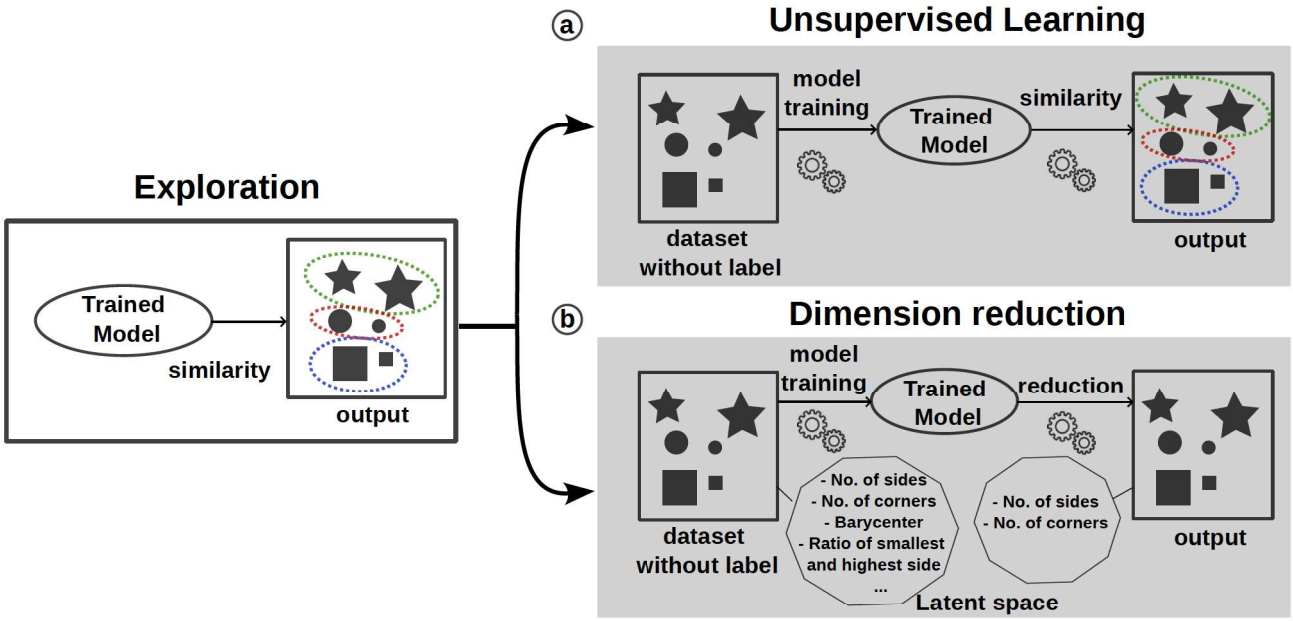


FIGURE 2.18: Artificial Intelligence algorithms for exploration. Exploration can be performed using ***(a)*** *unsupervised learning, and* ***(b)*** *dimension reduction.*

- **Self-supervised learning** is an emerging approach where models create pseudo-labels (Figure 2.19). This technique is mainly used for natural language processing, with examples like wav2vec, a self-supervised algorithm for speech recognition (Schneider et al., 2019), and BERT (Bidirectional Encoder Representations from Transformers) which enhances the understanding of search query context (Devlin, 2018). It is also applied in feature learning, as seen in algorithms like BYOL (Bootstrap Your Own Latent) (Grill et al., 2020).

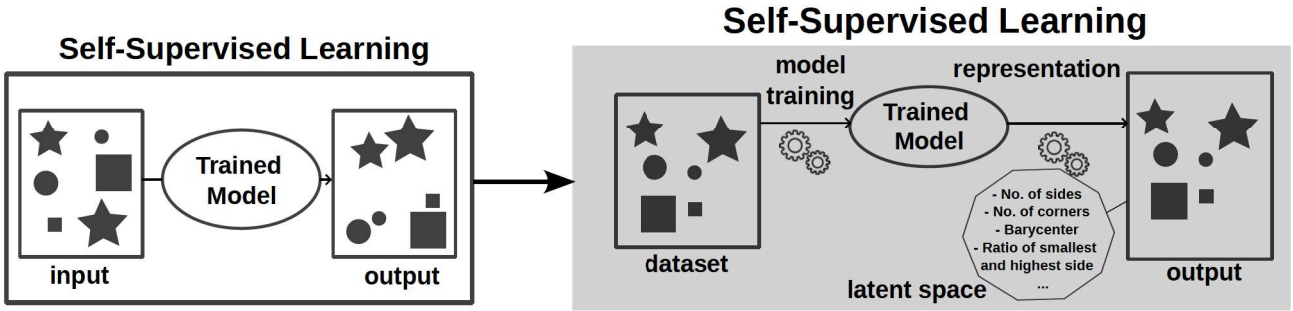


FIGURE 2.19: Artificial Intelligence algorithm for self-supervised learning.

- Interaction-based tasks are addressed using **reinforcement learning**, where models learn by interacting with their environment and receiving feedback in the form of rewards or penalties (Kaelbling et al., 1996) (Figure 2.16d). Reinforcement learning has broad applications, including in game playing (Mnih, 2013), robotics (Kober et al., 2013), and autonomous systems (Kiran et al., 2021), where the model continuously refines its actions based on trial and error.

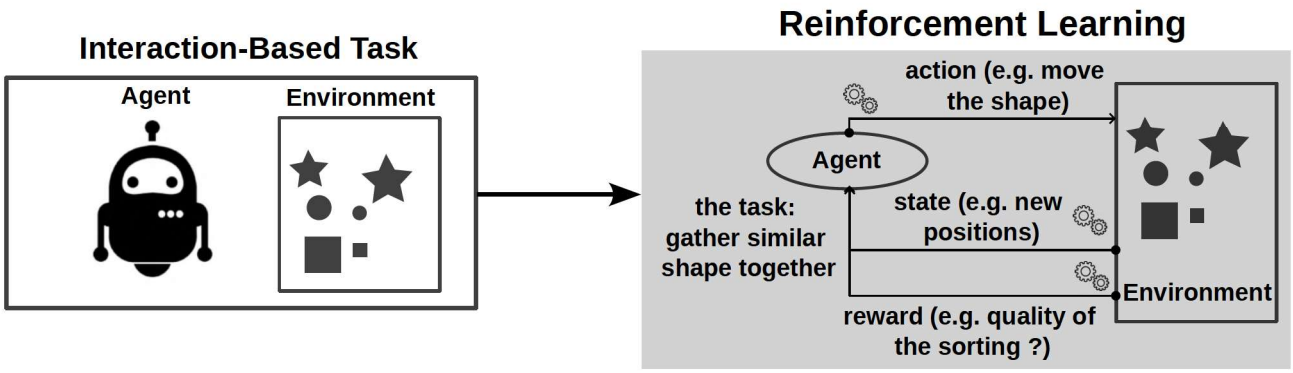


FIGURE 2.20: Artificial Intelligence algorithm families for interaction-based task. Interaction-based task can be performed using reinforcement learning.

2.5.1.3 Dataset Splitting and Class Imbalancing Management

Splitting the dataset into training and test sets is a crucial step in the development of AI model and evaluation process. The training set is used to teach the model by allowing it to learn patterns, relationships, and features within the data. The test set is reserved for evaluating the model performance on unseen data. This separation is essential to ensure that the model generalizes well beyond the data it was trained on, rather than simply memorizing the training data (a phenomenon known as overfitting) (Cawley & Talbot, 2010). Several strategies exist for dataset splitting: for huge datasets, **random splitting** is common, where the dataset is randomly divided into training and test sets, typically in a 50/50, 70/30 or 80/20 ratio. Stratified random splitting is a variation that ensures the distribution of target classes is proportional in both training and test sets, making it suitable for imbalanced datasets (Särndal et al., 2003). For reduced dataset, **k-fold cross-validation** is a common technique that involves splitting the dataset into k equally sized data subsets, training the model on k-1 data subsets, and evaluating its performance on the remaining data subset (Stone, 1974). This process is repeated over the number of fold k, with each subset serving as the test set once. Stratified cross-validation is variation of k-fold cross-validation that ensures each fold has a proportional distribution of target classes compared to the initial dataset (Zeng & Martinez, 2000; Krstajic et al., 2014). Leave-one-out cross-validation is a special case of k-fold cross-validation where the number of folds equals the number of samples, making it good for small datasets by maximizing data usage (Wong, 2015; Vehtari et al., 2017).

In the case of highly imbalanced dataset, where one class is significantly more prevalent than others, class rebalancing techniques can be used to rebalance the training set. These techniques include **oversampling**, where the minority class is augmented to balance the class distribution, using random sample duplication, selective sample duplication (e.g. ROSE) (Menardi & Torelli, 2014) or synthetic data generation (e.g. SMOTE, ADASYN) (Chawla et al., 2002; He et al., 2008); **undersampling**, where the majority class is reduced to match the minority class, using random sample removal or selective sample removal (e.g. NearMiss (Mani & Zhang, 2003), Tomek's link (Tomek, 1976), Edited Nearest Neighbour (Wilson, 1972)); and algorithm using a **mix of both oversampling and undersampling** (Batista et al., 2003, 2004). Other methods include cost-sensitive learning, where misclassification costs are adjusted to account for class imbalance (Sun et al., 2007), and ensemble methods, which combine multiple models to improve performance (Chen & He, 2011). Each class rebalancing technique has its advantages and drawbacks. Oversampling can lead to overfitting, as it may duplicate instances or generate synthetic data that does not fully capture the minority class diversity. Undersampling reduces the majority class, but it can result in the loss of important information, especially when the imbalance is large. Cost-sensitive learning adjusts misclassification costs, focusing more on the minority class, but requires careful tuning of cost values. Ensemble methods combine multiple models for better performance but can be computationally expensive

and still suffer from bias if the individual models are not balanced.

2.5.1.4 Evaluation Metrics

To evaluate the performance of AI models in prediction tasks, various metrics are used depending on the specific task. For classification tasks, metrics are derived from the **confusion matrix**, which summarizes the model performance by comparing predicted and actual class labels.

TABLE 2.1: Example of confusion matrix for classification

	Predicted class: Positive	Predicted class: Negative
Actual class: Positive	True positive (TP)	False negative (FN)
Actual class: Negative	False positive (FP)	True negative (TN)

Table 2.1 shows the confusion matrix for a binary classification problem. Common evaluation metrics include **accuracy**, **precision**, **recall**, and the **F1 score** (Powers, 2020). Their formulas are given respectively in Equation 2.1, 2.2, 2.3 and 2.4.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

$$F1\_\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

Accuracy measures the proportion of correctly classified instances, while **precision** quantifies the ratio of true positive predictions to the total number of positive predictions made by the model. **Recall**, also known as sensitivity, calculates the proportion of true positive predictions relative to the total number of actual positive instances. The **F1 score** is the harmonic mean of precision and recall, offering a balanced evaluation of model performance. Lastly, **AUC-ROC** assesses the model ability to distinguish between classes, with higher values indicating better performance. For regression tasks, metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared are used to quantify the model predictive accuracy. MSE measures the average squared difference between predicted and actual values, while MAE calculates the average absolute difference (Willmott & Matsuura, 2005). R-squared, also known as the coefficient of determination, assesses the proportion of variance in the dependent variable explained by the model.

This section introduced key AI concepts that are essential for understanding the rest of the thesis. The concepts of features and latent spaces explain how data is represented and how AI models learn from it. These ideas are particularly relevant to ML algorithms, where we define our own features. In the next section, we will explore existing literature, focusing on application involving seismic data measured with conventional seismometers and FO-DAS instruments. Additionally, understanding the different AI algorithms helps us identify tasks and choose the appropriate algorithms for classification, regression, or data exploration. These concepts will be applied in the following chapters,

especially for classification (Chapters 3 and 4) and exploration (Chapter 5). Finally, topics such as dataset splitting, class imbalance management, and evaluation metrics are vital for training and assessing AI models, especially when dealing with imbalanced datasets. These methods will be used to evaluate the built models in the subsequent chapters.

2.5.2 Artificial Intelligence for Conventional Seismometer and DAS Data

2.5.2.1 Machine Learning Features for Conventional Seismometer and DAS Data

Temporal, Frequency, and Time-Frequency Based Features

The application of signal-based methods, such as FFT, energy analysis, and bin division, is widely used in seismic signal processing for feature extraction. These methods can be categorized into temporal, frequency, and time-frequency domain approaches, each suited to specific types of signals. **Temporal domain** methods are suitable for non-stationary signals and include features built on the raw trace or its envelope such as statistical metrics (e.g. mean, median, standard deviation, kurtosis, skewness, autoregressive coefficients), signal energy, zero crossing rate, or correlation. **Frequency domain** methods are used for stationary or quasi-stationary signals. These methods examine the shape of the signal spectrum, its envelope (e.g. harmonics, spectral skewness), and various statistical metrics, such as the maximum, mean, and quartiles. **Time-frequency domain** methods analyze signals by combining time and frequency information, with the spectrogram being a common example. Mel-Frequency Cepstral Coefficients (MFCC) is another representation derived from the spectrogram, primarly used to imitate the human auditory perception using the Mel frequency scale. Temporal, frequency, and time-frequency features are crucial for machine learning applications in analyzing natural seismogenic sources measured using conventional seismometers, as demonstrated in studies by Hibert, Mangeney, et al. (2014); Hibert, Provost, et al. (2017); Hibert et al. (2019); Maggi et al. (2017); Provost et al. (2017); Chmiel et al. (2021); Domel et al. (2023). For DAS data, temporal, frequency, and time-frequency features have also been used together for various tasks such as intrusion detection, train tracking, and microseismicity detection. For example, Cao et al. (2015) used FO-DAS instrument for intrusion detection, applying FFT to extract spectral features like energy, low-frequency energy, peak value, and mean value for each channel. Similarly, Papp et al. (2016) created a real-time algorithm for tracking train positions using FFT to separate train signals from background noise, dividing the frequency range below 1000 Hz into 10 bins for classification with PCA and SVM. Fukushima et al. (2022) applied 2D-FFT (f-k filtering) to extract surface wave from DAS records. In another application, Stajanca et al. (2018) used FFT on a helicoidal fiber optic to detect small leaks in pipelines, as small as 0.1% of the flow rate.

Wavelet Transform, Empirical Mode Decomposition and Morphological Feature

Following the signal-based methods, more advanced feature extraction techniques based on signal decomposition have gained attention for their ability to capture multi-scale and non-linear features from seismic data. These methods are known as Wavelet Transform (WT) (Meyer, 1992; Daubechies, 1992; Akansu & Haddad, 2001), Empirical Mode Decomposition (EMD) (Huang et al., 1998; Rilling et al., 2003), or Morphological Feature Extraction (MFE) (Thiran & Macq, 1996).

- For seismic signal processing, **WT** provides a time-frequency representation that allows the analysis of both high- and low-frequency components across different time scales. Unlike spectrograms, which rely on a fixed Fourier transform (decomposing the signal into sinusoidal

TABLE 2.2: Comparison of Spectrogram, Wavelet Transform (WT), and Empirical Mode Decomposition (EMD) for seismic event recognition.

Transformation	Basis Functions	Resolution
Spectrogram	Sinusoidal	Fixed
WT	Scaled and shifted wavelets	Variable
EMD	Data-driven functions	Variable

components), WT offers the flexibility to choose the transformation function based on the specific needs of the analysis (Chakraborty & Okaya, 1995; Gaci, 2013; Liu et al., 2015).

Using conventional seismometer, Lapins et al. (2020) evaluate the WT for analyzing volcano-seismic signals. They show that WT provides better time-frequency resolution than Fourier transforms, especially for detecting very-long-period (VLP) signals and distinguishing volcanic activity from ambient noise. Figure 2.21 compare the representation of volcano-tectonic event using a linear-scaled spectrogram, log-scaled spectrogram and WT scalogram representation for volcano-tectonic event. (Atterholt et al., 2022) uses WT as a tool for improving DAS data processing, isolating earthquake signals from noise and improving aftershock detection by over 30%. (Wu et al., 2015) also show that WT can improve signal quality in DAS, achieving a 35 dB improvement by separating useful signals from noise. (Liu et al., 2020) extracts ML features using WT for vehicles detection purpose, achieving a detection accuracy of more than 80% and a classification accuracy of more than 70%.

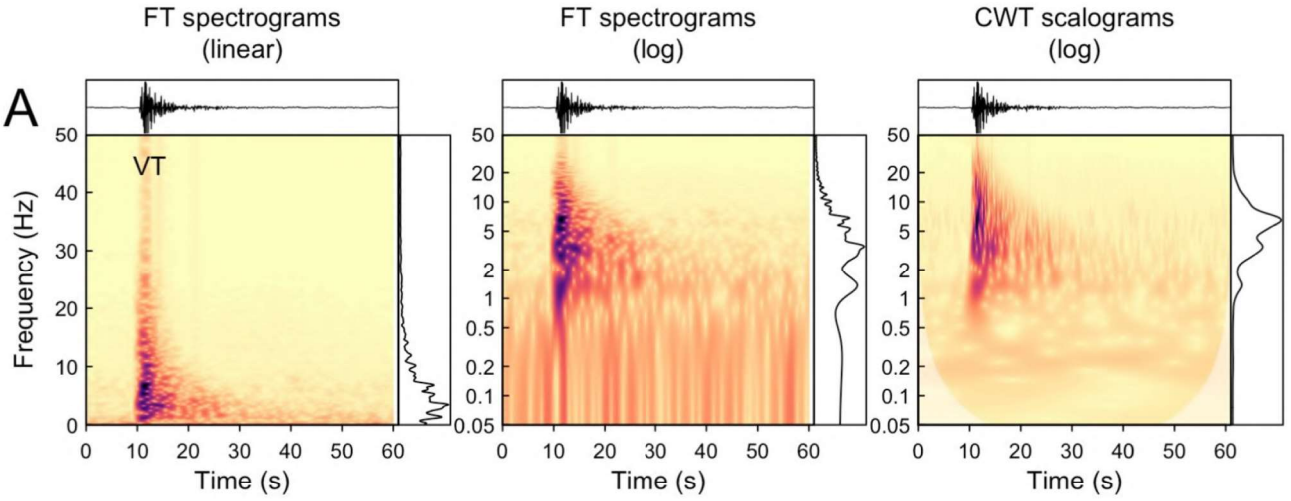


FIGURE 2.21: Linear-scaled spectrograms (left), log-scaled spectrograms (middle) and Wavelet Transform scalograms (right) for typical volcano-seismic events, and whose events were published in McNutt and Roman (2015) (taken from Lapins et al. (2020)).

- **EMD** uses a data-driven transformation function that adapt the differnet time scale (Battista et al., 2007; Han & van der Baan, 2013; Gaci, 2016), unlike WT which uses a fixed Fourier-based transformation.

  Han and van der Baan (2013) applied EMD for seismic signal denoising in sedimentary basin monitoring. In his study, he first demonstrated the method on a synthetic example, showing how a complex signal can be effectively decomposed into its intrinsic mode functions (IMFs) (Figure 2.22). Wang et al. (2020) extracted features from the EMD representation to identify five types of disturbance events: watering, knocking, climbing, pressing, and false disturbances. They selected 11 features from the data and applied XGBoost for classification, achieving over 90% accuracy for each event class.

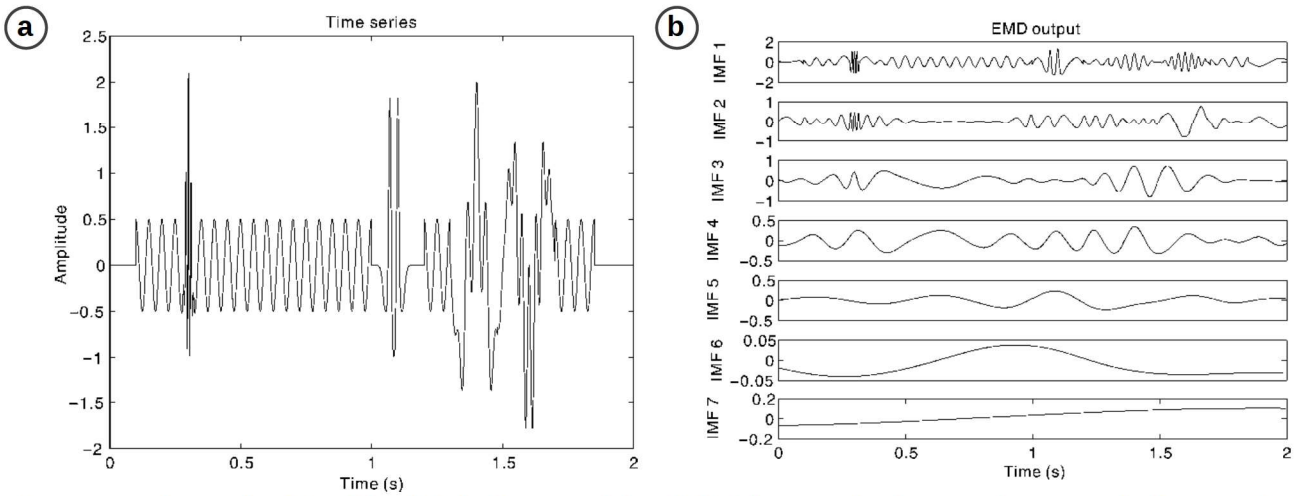


FIGURE 2.22: Example of Empirical Mode Decomposition (EMD) for a synthetic example. (a) Synthetic example: background 20 Hz cosine wave, superposed 100 Hz Morlet atom at 0.3 s, two 30 Hz Ricker wavelets at 1.07 and 1.1 s, and there are three different frequency components between 1.3 and 1.7 s. (b) EMD output displaying mode mixing. IMF1 extracts the high-frequency Morlet atom and some low-frequency components. IMF2 and IMF3 also mix different signal components (taken from Han and van der Baan (2013)).

- **MFE** relies on morphological operators to extract features related to the shape and structure of the signal. These operators are based on the principles of mathematical morphology, which focuses on the manipulation of the geometrical structure of data to alter the initial data.

  Pérez et al. (2021) propose a novel method for classifying long-period (LP) and Volcano-Tectonic (VT) seismic events at Cotopaxi volcano by combining mathematical morphology and similarity criteria techniques. Their approach, using edge maps from spectrograms, achieved high accuracy (93.34% and 96.88%) and faster execution times compared to state-of-the-art methods (Figure 2.23). Sun et al. (2015) classified three types of events using MFE: walking, digging, and vehicles passing. They applied segmentation to separate the event region from the background. Features such as region interval, roundness, pixel count, and amplitude were then selected. SVM was used for classification, achieving 95% accuracy.

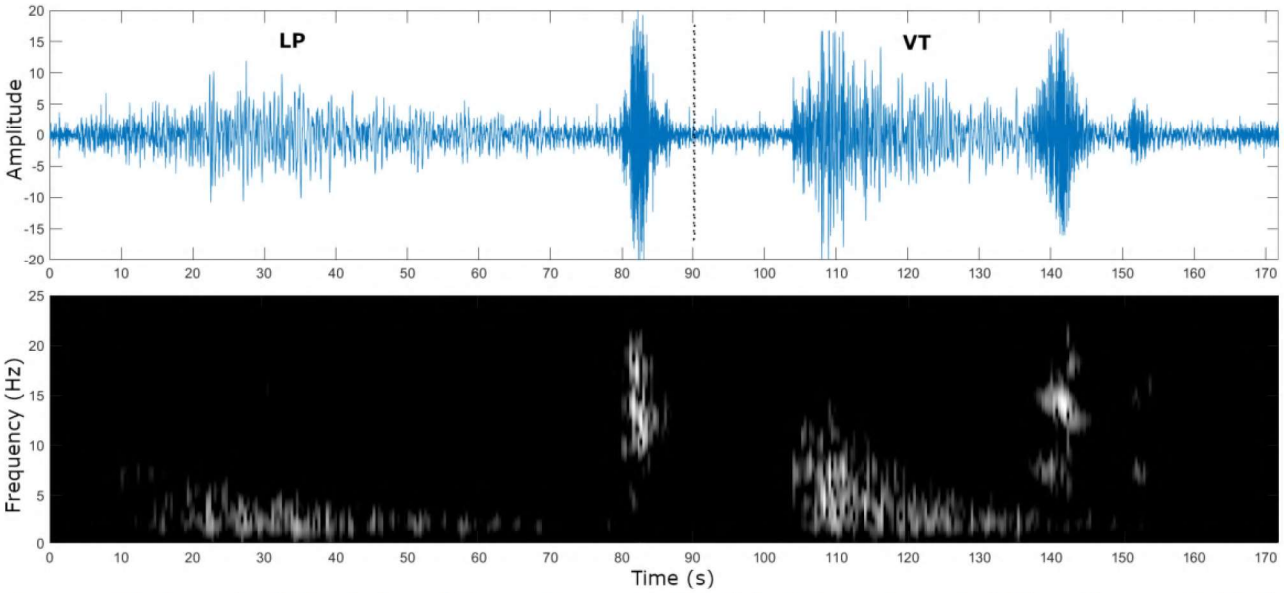


FIGURE 2.23: Example of Morphological Feature Extraction (MFE) for Volcano-Tectonic (VT) and long-period (LP) seismic event recognition. The figure shows the trace recording and the binarized spectrogram of a VT and LP recorded at the Cotopaxi volcano in Ecuador (taken from Pérez et al. (2021)).

2.5.2.2 Application of Artificial Intelligence For Seismic Data Pre-Processing

Data Denoising

AI can help to reduce the noise level for seismic data analysis. Recent advancements include the use of Convolutional Neural Networks (CNN), as seen in the work by Yang et al. (2023), where a DL model with dense and residual connections was developed to attenuate noise and improve the signal-to-noise ratio in DAS data. This method demonstrated a remarkable ability to recover weak signals hidden within strong noise, outperforming traditional approaches. Furthermore, self-supervised learning techniques, such as those explored by Van den Ende et al. (2021), enable denoising without the need for labeled data, leveraging the inherent structure of seismic signals. In addition, innovative methods like the one proposed by Lapins et al. (2024), which utilizes two fiber-optic cables to denoise data without requiring clean reference signals, are pushing the boundaries of what DAS technology can achieve in seismological monitoring.

Event Picking

AI is used in seismic data pre-processing to automate the picking of seismic phases, as P-waves, S-waves, and surface waves (see Section 2.1.2). Accurately identifying these arriving times is essential for calculating key event parameters, including the epicenter, magnitude, and depth of the seismic event. Traditional methods rely on manual picking by experts, which can be time-consuming and subjective. Ross and Ben-Zion (2014) presents a method for automatic seismic phase picking that sequentially applies different data processing algorithms. Since the detectors are applied in a fixed order, this approach is not refered as ML. The approach combines STA/LTA, kurtosis, and skewness detectors to accurately detect both P and S-wave, and detect S-wave arrivals using polarization analysis and filters to eliminate P-wave energy. Zhu and Beroza (2019) developed PhaseNet, a supervised DL model designed to automatically pick seismic phases from conventional seismometer data. By using three-component seismic waveforms as input, PhaseNet outputs probability distributions for P and S wave arrivals. Trained on over 700 000 waveform samples, it outperforms existing methods, achieving precision of up to 0.85 and recall of up to 0.75 for both P and S waves. An adapted version for DAS data is also discussed in Zhu et al. (2023). Chen (2020) proposes an unsupervised ML method for microseismic event picking, using fuzzy clustering to effectively separate waveform points from noise. This approach outperforms traditional STA/LTA methods, showing improved robustness in detecting events even with moderate-to-strong background noise.

2.5.2.3 Application of Artificial Intelligence For Seismic Data Classification and Exploration

In Figure 2.16, we present the main types of AI algorithms, which are designed to address prediction, exploration, interaction-based tasks, or self-supervised learning tasks such as feature creation. In seismology, AI are primarily applied to prediction and exploration tasks.

Prediction

AI is being used in seismology with DAS data in prediction field to improve the detection and analysis of seismogenic events. For example, DL methods have been used to better detect earthquakes with DAS data (Hernández et al., 2021). In microseismicity, techniques like 2D FFT and CNN are

being used to classify seismogenic events, outperforming traditional methods (Stork et al., 2020; Binder & Tura, 2020). AI is also used to filter out traffic noise from seismic data for interferometry (Huot et al., 2018) and to detect intrusions on railways using LSTM networks (Li et al., 2020). In smart city issues, AI is applied to detect and count footsteps (Jakkampudi et al., 2020) and classify vehicles (Liu et al., 2020).

Exploration

AI can be used to categorize seismic signals, discover new patterns in seismogenic events, and identify rare occurrences that traditional signal processing methods may miss. The use of DAS data in exploration is still limited but growing rapidly, while conventional seismometers have been more widely used. For volcanoes, unsupervised learning techniques are used to identify clusters of volcanic activity phases (Hammer et al., 2012). These methods help with challenges such as the lack of labeled data during eruptions. Clustering algorithms are also used to study tremor patterns and their link to volcanic features like magma viscosity and storage depth (Unglert & Jellinek, 2017). DL methods track volcanic eruptions by analyzing seismic data (Zali et al., 2024). Clustering algorithms classified bubble events based on their temporal and spectral features, providing information about volcanic processes and gas infiltration patterns (Caudron et al., 2024). Self-supervised approaches identify different types of volcanic tremors by using features from seismic records and spectrograms (Rimpot, Hibert, Retailleau, et al., 2024). In earthquake monitoring, methods like Fingerprint And Similarity Thresholding (FAST) technique detects similar seismogenic events, helping to monitor earthquakes (Yoon et al., 2015). Clustering methods group earthquake events by similar characteristics, which aids in understanding seismicity and faulting processes (Cesca, 2020). For landslides, clustering techniques identify different seismic behaviors in landslide areas (Seydoux et al., 2020). Slowly evolving clusters, often linked to ambient seismic noise, dominate the data. Localized clusters are associated with ocean-radiated microseismic energy, revealing specific sources and frequencies. Sparse clusters, tied to rare seismogenic events, may help detect potential landslide precursors. In DAS applications for exploration, there are fewer studies, but volcano monitoring has seen progress. For instance, an experiment at Laacher See volcano used FO-DAS instrument to detect and analyze acoustic signals from bubble emissions.

2.6 Research Questions and Hypotheses

In exploring the state of the art, FO-DAS instrument has emerged as a transformative tool for seismic monitoring, offering advantages such as high spatial resolution, cost efficiency, ease of deployment, and resilience in harsh environmental conditions. These benefits make FO-DAS instrument highly promising for applications ranging from earthquake detection and landslide monitoring to intrusion detection and railway surveillance. However, DAS data also introduces significant challenges. These challenges include the massive data volumes it generates, the elevated background noise levels, and its sensitivity to diverse noise sources, which can complicate its use in specific monitoring contexts. Addressing these challenges requires innovative solutions to denoise, classify, and efficiently process DAS data while maintaining the interpretability and reliability of results.

Several solutions can be considered after reviewing existing research to address the challenges. For background noise, signal processing tools can reduce instrumental noise, improving data quality. To select sources, methods like waveform and frequency analysis can help identify specific events.

These techniques can be automatically combined together using AI algorithms. To manage data volume, we can choose which sources to record using a trained AI algorithm, allowing for a real-time monitoring system to efficiently handle the data.

As seen in this Chapter, there are two main trends for AI applications in seismic monitoring: Machine Learning (ML) and Deep Learning (DL). While DL models have shown impressive performance in various applications, **we choose to use ML techniques** (hypothesis **H0**) in our work for several reasons:

- ML models are easier to interpret because they use features derived from seismology, making it clearer how the data relates to the predictions.
- ML is more robust, even when there is little data available, whereas DL typically requires large datasets to perform well.
- ML has lower computational costs, which makes it more efficient in terms of both time and resources.
- ML models are less sensitive to changes in model parameters, allowing us to focus more on the quality of the data and the features for model performance analysis.

2.6.1 Research Questions

In the light of these challenges, the possible solutions and the use of ML rather than DL, we have identified several research questions to guide our work:

1. **Real-time monitoring**: The most effective ML algorithms typically classify complete seismic signals of events, from the beginning to the end of the detection. However, in real-time monitoring, waiting for an event to finish (e.g., several minutes for detecting a vehicle moving along an fiber optic path) is not feasible. Given our initial hypothesis, we want to prioritize features derived from conventional seismology. **How can existing processing methods be adapted to handle real-time data while considering the distributed spatial nature of DAS data in the processing chain?**
2. **Signal classification**: ML algorithms based on seismic features tend to be more interpretable but do not account for the distributed spatial nature of DAS data. **How can we integrate the unique characteristics of DAS data into ML algorithms, such as analyzing the spatial waveform of seismic data?**
3. **Data labeling**: For long-term monitoring systems, extended recordings are needed to account for factors like seasonal effects and environmental variations over time. In these cases, manually labeling the data is highly time-consuming and labor-intensive. **How can training datasets for ML models be created more efficiently, especially when dealing with large volumes of continuous data that span across seasons or even years?**

2.6.2 Hypotheses

To address these research questions, we have formulated several hypotheses:

- **H0**: We prefer to use ML techniques over DL for the interpretability of features.
- **H1**: We have chosen to adopt a data stream approach, building our processing chains around this concept.
- **H2**: While we are focusing on real-time applications, we prioritize questions about the benefits of using new features in the classification process rather than minimizing the computation time of these features. We also allow for data processing on High-Performance Computing (HPC) systems.
- **H3**: For long-term labeling, we aim to avoid manual event-by-event labeling. However, we allow for intervention on a few event or cluster of events in the labeling process to ensure the quality of the training data.

2.6.3 Solutions

To address these research questions and consider the hypotheses, we develop multiple solutions. A summary of the datasets used to validate these solutions is provided in Table 2.3.

- **Solution 1: Real-time Event Detection and Classification Pipeline**
  We develop a complete DAS data processing pipeline for real-time event detection and classification. This uses ML techniques adapted from conventional seismology. The pipeline includes spatial data coherency processing, using Markov Random Fields to account for spatial dependencies. We test the processing chain on a dataset from the FEBUS Optics test center, and containing anthropogenic events (see Table 2.3). This is covered in Chapter 3.
- **Solution 2: Integration of Spatial Features in ML process**
  In this step, we incorporate the spatial features of DAS data and the relationships between virtual point sensors into the ML model. We apply this method to a DAS dataset acquired in the field. The dataset consists of data from a 91-km fiber optic in the Hautes-Pyrénées, containing both natural seismogenic events and anthropogenic events (see Table 2.3). This is discussed in Chapter 4.
- **Solution 3: Streamlining Dataset Labelizing**
  We design a method to streamline the creation of training datasets for DAS acquisition. This involves clustering similar DAS data, reducing the need for manual event-by-event labellization. We test this method on a dataset acquired in the Hautes-Pyrénées and at Viella landslide, which includes natural seismogenic events, daily farm anthropogenic events, and other anthropogenic events (see Table 2.3). This is presented in Chapter 5.

Table 2.3: Summary table of the acquisition parameters for the test center, Hautes-Pyrénées, and Viella landslide datasets. The upper section of the table presents the acquisition configuration, while the lower section lists the observed seismogenic sources and their usage in the chapters.

	FEBUS Test Center	**Hautes-Pyrénées**	**Viella Landslide**
Chapter	3	4, 5	5
Cable Type	Loose, Tight, Telecom	Telecom	Tight
Fiber Length (m)	24x 22	91 000	800 (begin) 214 (end)
Duration	120.6 min	21 days	44 days
Used Duration	120.6 min	19 times 10-min	44 days
Time Sampling (Hz)	200	200	400
Space Sampling (m)	0.8	4.8	2.4
Gauge Length (m)	5	10	10
Data Volume (GB)	5	206	3000
Event types	Construct. noise Leakages	Earthquake Quarry Blast Oth. Anthropogenic	Earthquake Farming Oth. Anthropogenic
Usage	Classification	Classification Clustering	Clustering

Chapter 3

Classifying Fiber Optic Data Using Human-Engineered Features from Conventional Seismometers

3.1 Introduction

The processing and analysis of data from Distributed Acoustic Sensing (DAS) systems present several challenges. Compared to traditional seismometer data, DAS data has a higher noise level and includes a spatial dimension due to its distributed nature along an fiber optic. DAS also generates large amounts of data, making data storage bulky. These unique characteristics require a specific processing chain for DAS data. A key element of this chain is the use of a suitable latent space for machine learning (ML) models.

In this chapter, we present our approach, which uses features that have been successful in classifying seismic signals from conventional seismometers (Hibert, Mangeney, et al., 2014; Hibert, Provost, et al., 2017; Hibert et al., 2019; Maggi et al., 2017; Provost et al., 2017; Chmiel et al., 2021; Domel et al., 2023). These features are applied in this chapter directly to DAS data flow. We also introduce a post-processing step using a Markov random field algorithm to enhance classification. This step leverages the spatial nature of the signals to maintain classification coherency and reduce the impact of noise. We demonstrate how this processing chain works on data collected at the FEBUS Optics test center from a controlled environment, specifically a 20-meter length trench. This chapter establishes the foundation for a processing chain to classify DAS data using latent spaces derived from conventional seismology features.

This chapter consists in one published paper (section 3.2):

Huynh, C., C. Hibert, C. Jestin, J.-P. Malet, P. Clément, and V. Lanticq (2022). *Real-Time Classification of Anthropogenic Seismic Sources from Distributed Acoustic Sensing Data: Application for Pipeline Monitoring*, Seismol. Res. Lett. 93, 2570—2583, doi: 10.1785/0220220078.

3.2 Paper: Real-Time Classification of Anthropogenic Seismic Sources from Distributed Acoustic Sensing Data: Application for Pipeline Monitoring (Seismological Research Letter, 2022)

3.2.1 Abstract

Distributed Acoustic Sensing (DAS) is an innovative method to record seismic waves using an fiber optic as a network of sensors. Current DAS devices can monitor up to 50 km of fiber optic and the use of optical repeaters can raise even more this length, while allowing a spatial discretization of the order of a meter. Handling such amounts of data is a challenge in terms of data management and data analysis (such as event source identification), more specifically for monitoring applications such as infrastructures or natural hazards. In this work, we propose a processing chain for real time classification of anthropogenic sources using a combination of Random Forest (RF) and Markov Random Field (MRF). To develop the method, we choose to focus on the application of pipeline monitoring. The algorithm is therefore trained to recognize six classes of seismic sources: pedestrian, impact, backhoe, compactor, leak and noise. All the sources were triggered and recorded on our own test bench under controlled conditions. The average sensitivity of our processing chain reaches 83% with the use of only RF and achieves 87% in combination with MRF. Classification maps show that the MRF approach can increase the average sensitivity by removing isolated signals. In addition to this improvement in sensitivity, this new approach also permits to identify synchronous events taking place at nearby positions, which is difficult with classical methods.

3.2.2 Introduction

Distributed Acoustic Sensing (DAS) is a new technology that turns a telecommunication fiber optic into a large array of equally spaced seismic sensors. Such a high density of sensors is ideal for monitoring of large infrastructures, such as oil/gas pipeline (Tejedor et al., 2021a), train tracks and roads (Li & et al., 2020; Wiesmeyr et al., 2020; Yuan et al., 2021) or even submarine power cables (Hicke et al., 2017b). It is also a powerful solution for geophysical applications, e.g., seismological monitoring (Lindsey et al., 2017; Lellouch et al., 2020; Nayak et al., 2021; Zeng et al., 2022), Ocean-Solid Earth interactions (Sladen et al., 2019) or natural hazards (e.g. volcanoes (Nishimura et al., 2021); water reservoir and rivers (Zhu et al., 2021)). Using a single opto-electronic device called interrogator, the system can monitor up to 50 km of fiber length. Once a LASER pulse is sent by the interrogator into the fiber optic, a small part of the light is continuously backscattered due to natural asperities present in the fiber optic. Knowing the speed of the light in the fiber optic, the position where the light has been backscattered can be calculated. The phase difference of the backscattered light between two points spaced by a distance named gauge length (GL) is proportional to the strain along the fiber. Then, a passing seismic wave with enough energy to excite the fiber will be detected by the DAS system. DAS data are dense, with a spatial resolution under the meter and an acquisition frequency which can reach 100 kHz for the interrogation of 1 km-fiber-length. The complexity of seismic signals and the high amount of data make the use of automation mandatory. For continuous monitoring, real-time processing is critical.

In the field of DAS signal classification and pattern recognition, two main approaches can be distinguished: temporal seismic signal classification and image-based classification. The first approach aims to use an already existing classification method for seismic and seismological signals measured with microphones or geophones. This approach is based on the analysis of relevant features like the

energy in several frequency bands (Wiesmeyr et al., 2020; Tejedor et al., 2021a), correlation-based features, or cepstral coefficients (Bublin, 2021). However, it only focuses on a single position on the fiber and rarely considers the classification of neighboring positions. The second approach considers the strain rate recorded by the DAS as an image and uses artificial neural networks (e.g. convolutional neural network (Huot et al., 2018; Jakkampudi et al., 2020; Li & et al., 2020; Peng et al., 2020); or residual neural network (Dumont et al., 2020)). However, these methods work mostly for already acquired signals and their use in real-time is quite complex. Further, the choice of window sizes depends on the sources to be detected and most of the developments were adapted for events of close temporal duration (Huot et al., 2018; Bai et al., 2019).

In this work, we propose a source classification method based on the first approach, as our goal is to propose a processing chain able to process data in real-time. To solve the issue of position independence, a post-processing step using Markov Random Field theory is added after the classification. The method has been tested on on-site recorded DAS signals and the implementation is also compatible for data acquired in real-time. The system is trained to identify several types of simulated seismic sources corresponding to threats and non-threats to pipeline integrity (pedestrians, falling objects, impacts, compactor, excavator, water or air leaks). We propose to use a combination of the Random Forest classifier along with 53 curated seismic signals features as it has proven its efficiency in many fields of seismological data analysis (Hibert, Mangeney, et al., 2014; Hibert, Provost, et al., 2017; Hibert et al., 2019; Maggi et al., 2017; Provost et al., 2017; Chmiel et al., 2021; Wenner et al., 2021). This article has three main objectives. The first one is to test the proposed feature set combined with Random Forest algorithm for data classification for the studied event types. The second is to study the contribution of spatial clustering using Markov Random Field for classification improvement. The last aim is to test the whole processing chain for a real-time classification.

3.2.3 Dataset

The data were acquired on a controlled test bench developed and installed by FEBUS Optics (Pau, South-West of France). The facility was built in February 2020 in order to develop, improve and benchmark the distributed sensing solutions developed by the company and to provide ideal experimental conditions to simulate various events, such as third-party intrusions, landslides or water/air leakages. The test site is composed of a buried pipeline of 22 m with several configurations of fiber optic cables (Figure 3.1): single/multi-mode fibers, loose/tight fibers in the cable, with/without cable sleeves, and cable located at various distances from the pipe. The goal is to be able to study the influence of the nature of the fiber, of the type of fiber cable, of the coupling between the cable and the ground and of the position of the fiber optic cable. The trench is also equipped with different-size nozzles along the pipe for leakage simulation.

The device used to interrogate the fiber optic cables is the DAS system FEBUS A1-R. The acquired data is called strain rate (SR, in nanostrain/s) and corresponds to seismic aquisition along the fiber. For each position on the fiber corresponds a temporal seismic aquisition called trace. For our simulation, strain rate is measured and filtered by a low-pass filter with a cut-off frequency of 100 Hz and downsampled. The gauge length is chosen equal to 5 m with a spatial sampling of 80cm. For real-time applications, one interesting property is that the FEBUS A1-R measures the strain rate in flow: during launched aquisitions, 4 s of strain rate is continuously acquired and stacked into the same file. The size of the blocks is editable, which makes it convenient for real-time processing. The dataset consists in a series of six controlled seismic sources (Figure 3.2): “pedestrian walk”, “impacts”, “backhoe”, “compactor”, “water/air leakages” and “noise”. The different seismic sources constituting our dataset were acquired using all available fiber configurations (Figure 3.1), which allows us to have for each event multiple measurements called sub-event. Figure 3.3 gives a qualitative view of

the signals acquired along the different fibers in terms of energy over a frequency band 0-100 Hz and emitted by pedestrian and backhoe. Each sub-event can be considered as an ensemble of traces measured at several positions on the same portion of the fiber optic cable, and each trace is further divided using a 4 s window with an overlapping of 50%. These new entities are called signal units and are used by the machine learning algorithm presented in this paper. Table 3.1 gives an overview of the number of events, sub-event and signal units per class. The protocols for creating the sources for all classes are detailed in Appendix A1.

TABLE 3.1: Event occurrences for each considered class in the database.

Event Type Name	Number of Events*	Number of Subevents†	Number of Signal Unit‡
Pedestrian walk	15	244	254,129
Impacts or falling object	48	1017	344,609
Backhoe	10	195	220,313
Compactor	5	114	57,516
Leakages	11	239	548,185

* One event gathers all signals emitted by the same seismic source.
† One subevent gathers all signals measured along the same fiber configuration.
‡ One signal unit consists of a 4 s windowing of a signal taken at a point of the fiber.

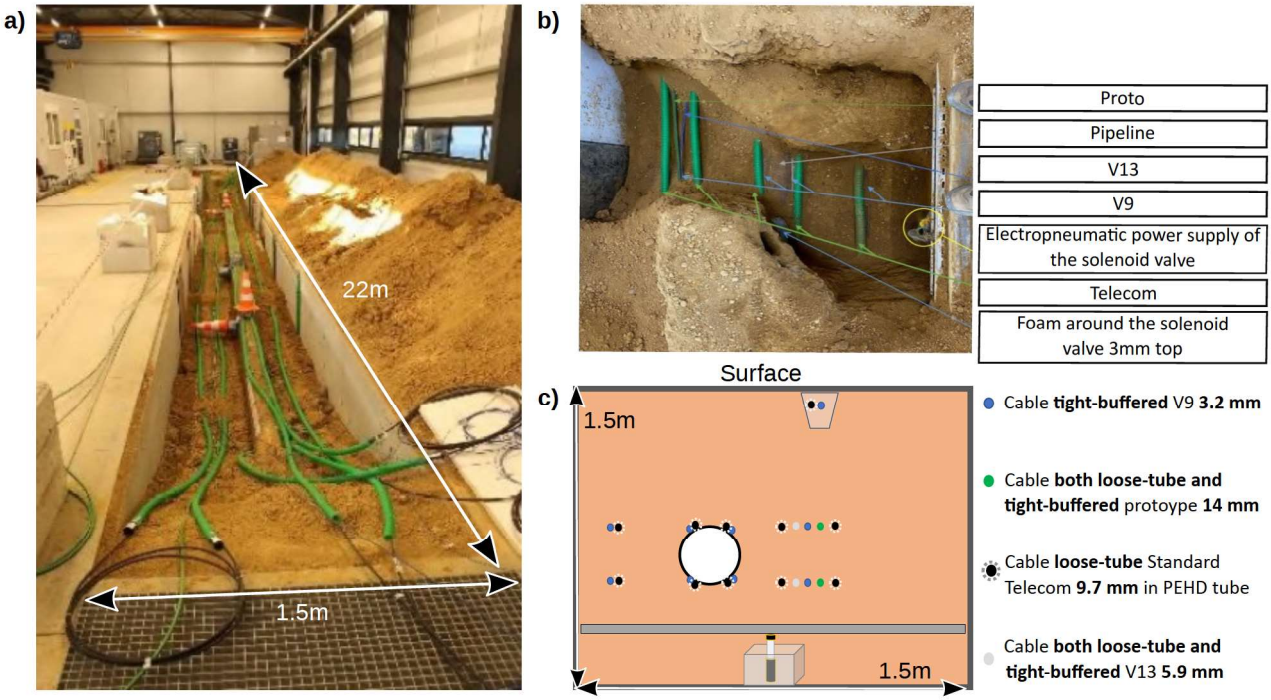


*FIGURE 3.1: Experimental setup for seismic source acquisition. **(a)** An overview, **(b)** a top view, and **(c)** a sketch of a sectional view with detailed fiber optic cables.*

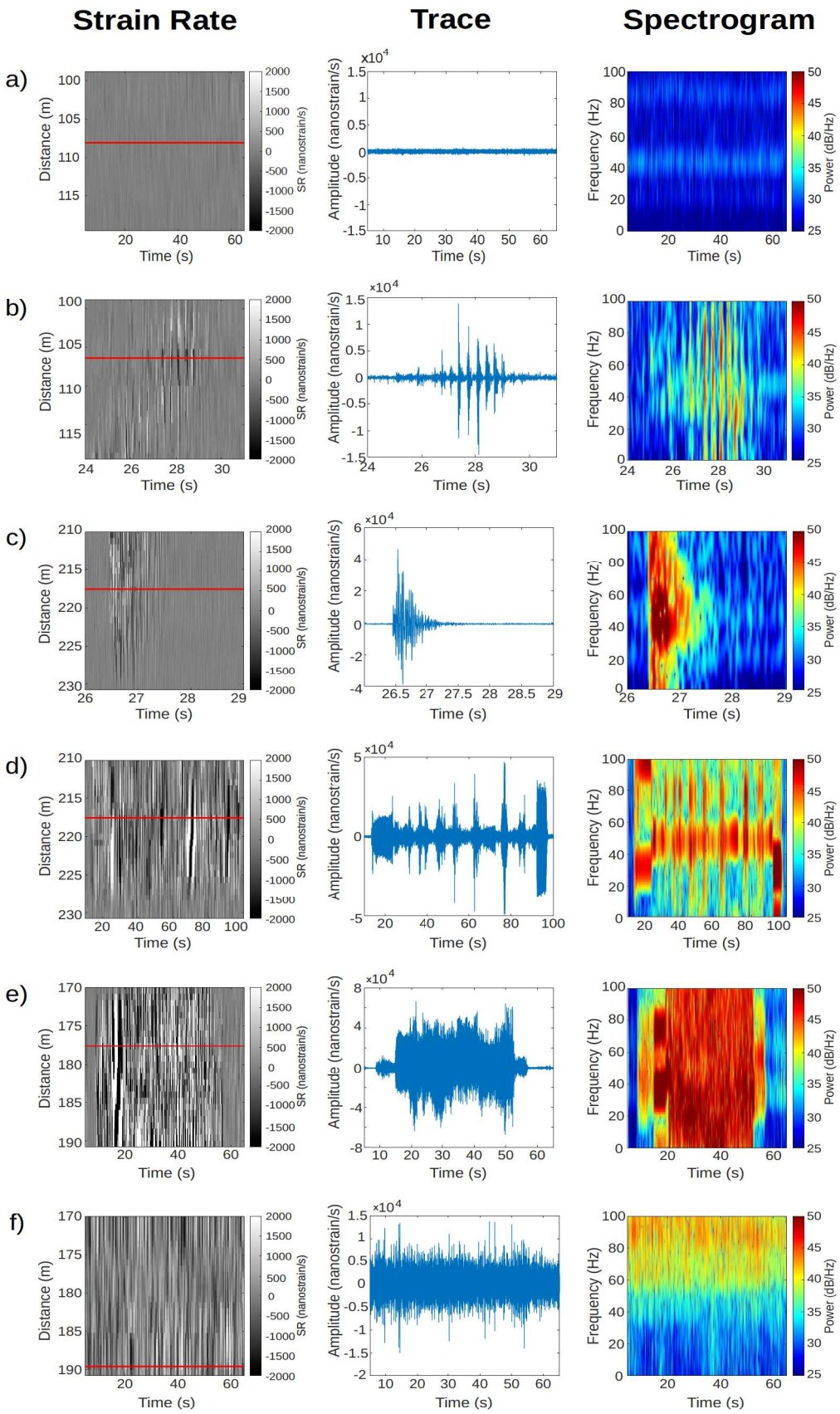


FIGURE 3.2: Examples of controlled seismic sources represented by their strain rate, their strain rate along the red line and their spectrograms. The sources are, respectively ***(a)*** *noise,* ***(b)*** *pedestrian walk,* ***(c)*** *impact,* ***(d)*** *backhoe,* ***(e)*** *compactor, and* ***(f)*** *water and air leakages.*

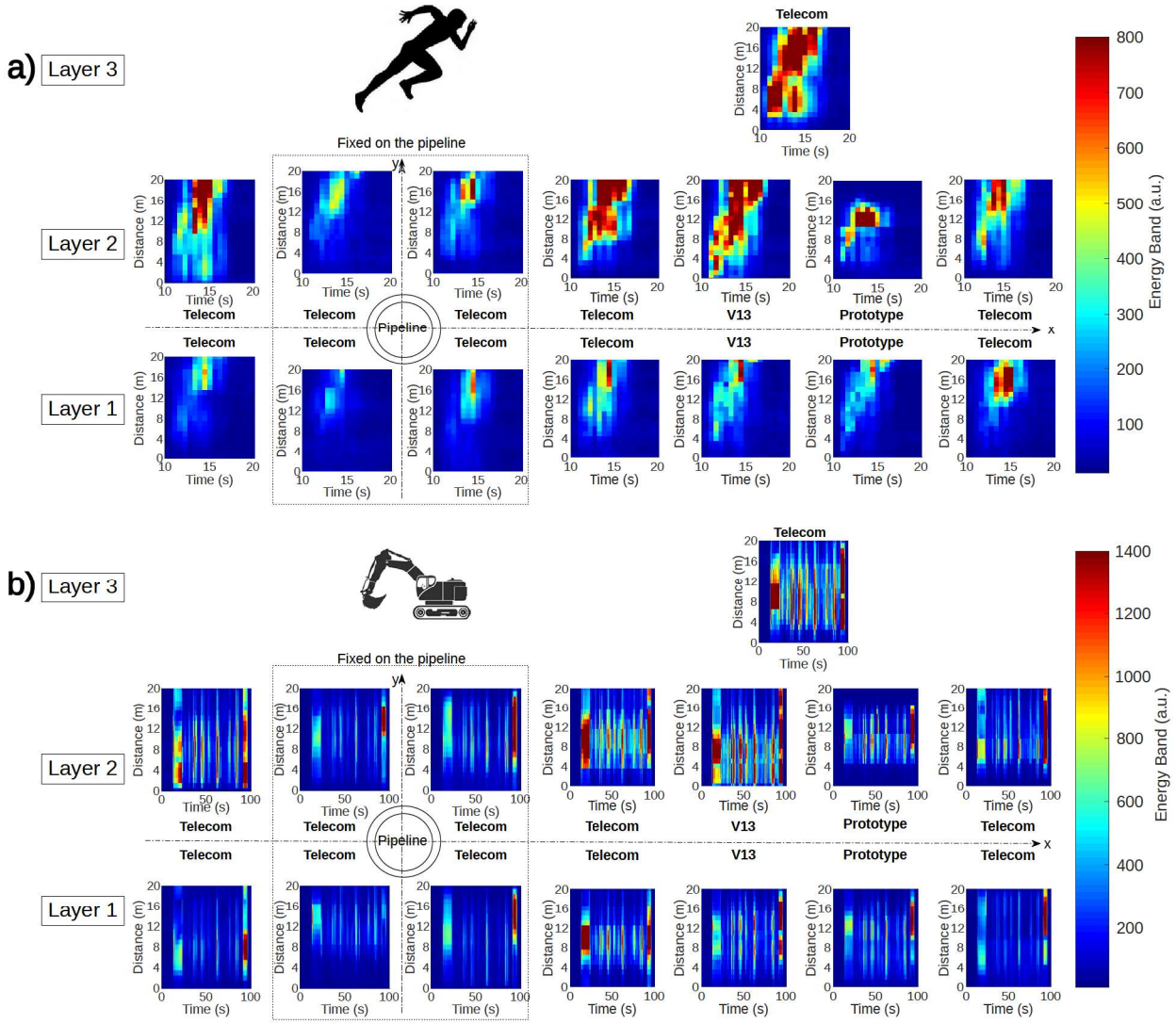


*FIGURE 3.3: Energy response of different cable configurations of **(a)** the movement of a pedestrian and **(b)** the work of an excavator.*

3.2.4 Methodology for Signal and Source Classification

3.2.4.1 Features to Describe the Seismic Signal

53 features (detailed in Appendix A2) are calculated to describe the studied seismic signals. These features have been selected from previous works related to seismological signals classification (Bessason et al., 2007; Provost et al., 2017; Hibert et al., 2019). They can be grouped into three main categories: features related to the waveform, features related to the signal frequency content and features related to the pseudo-spectrograms. Figure 3.2 provides a waveform (called trace) and spectrogram representation for each considered class. Features gathered under the denomination waveform are related to the signal in the temporal domain or its envelope, on which can be computed its energy, its kurtosis, or the autocorrelation function. Spectral features focus on the signal frequency distribution like the most powerful frequency, the number of peaks in the spectral representation or the energy contained between two frequencies. Pseudo-spectrogram features add to the spectral representation the time dimension, which allows study of the spectral evolution of the signal. The designed features for this third category are inspired from the feature built in the two previous categories of feature, as it is possible to create a new time-domain function by picking the value of max, mean, or median frequency. Spectral and pseudo-spectrogram features are computed using the Discrete Fourier Transform (DFT).

3.2.4.2 Source Detection and Classification with a Random Forest Model

The machine learning classification is based on the Random Forest method (RF) applied to the set of computed features described above. RF is an ensemble learning method based on the use of a high number of independent decision trees (Breiman, 2001). Referring to the denotation "supervised classifier", Random Forest needs to be trained, and uses for that purpose a well-localized and well-labeled event dataset. Each decision tree is then trained independently using for each tree a randomly chosen subsample of the training dataset (tree bagging) and a randomly chosen subsample of features (feature bagging). Once trained the model can be used to identify new data. The returned result of the algorithm is the class that gathers the majority of votes, knowing that each tree votes one time. RF works well, especially when a lot of features should be handled: RF is able to evaluate the importance of each feature for the classification task and then limit the use of less important ones. Another ability of the RF is to provide a score of classification for each possible class, based on the percentage of vote collected from the ensemble of trees (Figure 3.4).

The dataset is splitted in a training and a test sets. The training set is used to train the algorithm to identify several types of events. The goal of the test set is to quantify the performances of the classification algorithm, using tools like confusion matrix and metrics like sensitivity. Confusion matrix compares predicted class with real class and summarizes it inside a matrix, while sensitivity provides for each real class the proportion of correctly classified data. To avoid that the same event has traces in the training and test set, the dataset is splitted by event in the proportion of 50% of events for training and 50% for test. As the recording duration of same class events are close, the same proportion is automatically kept in terms of signal unit.

For the training of our algorithm, the dataset is balanced using an oversampling algorithm called SMOTE (Chawla et al., 2002). SMOTE method is useful when the headcount of one class is at least ten times smaller compared to the dominant one, as this is the case for class compactor (Table 3.1). The major contribution of dataset balancing is to classify the signal units only according to their likelihood and not in combination with the probability of occurrence.

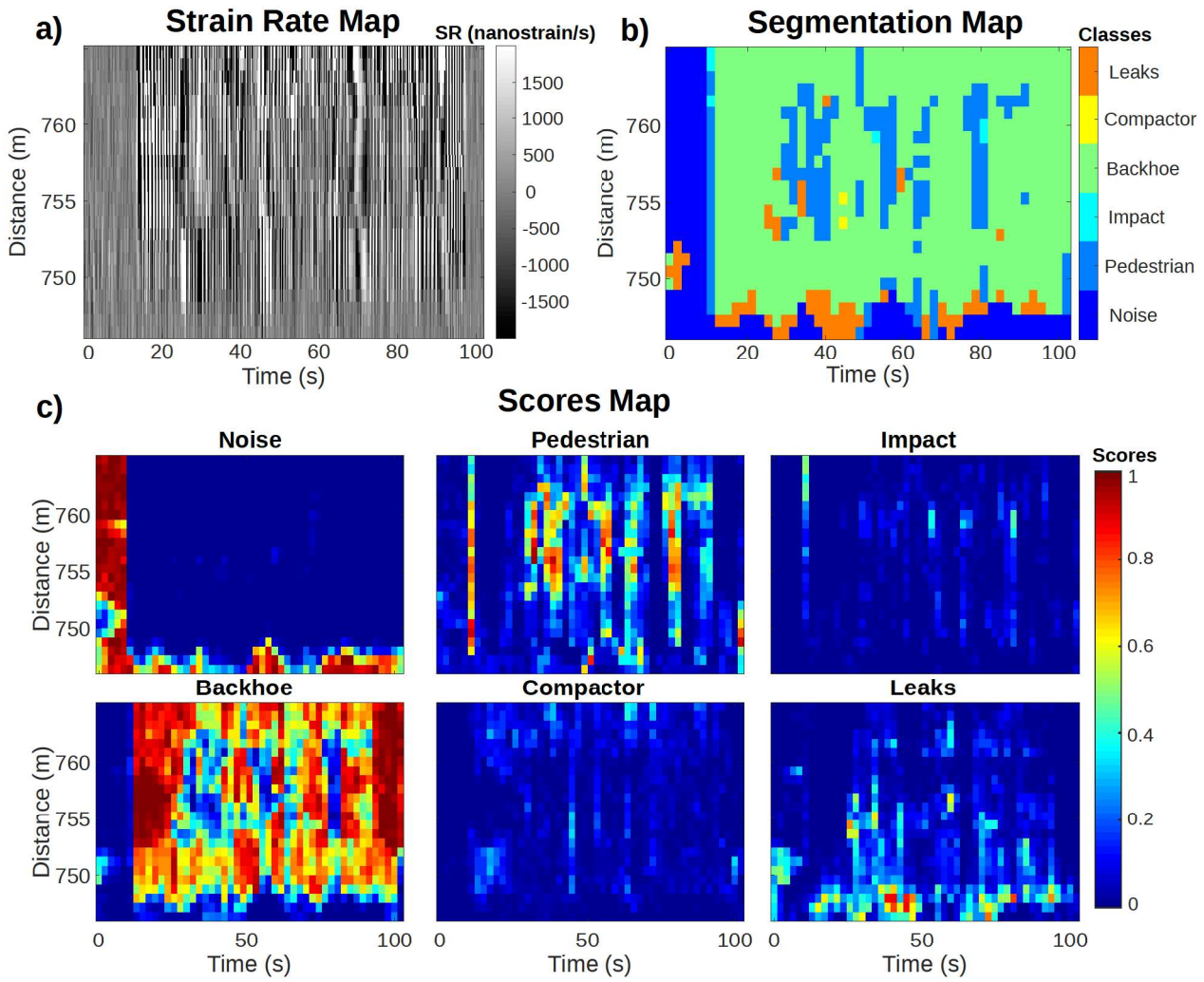


*FIGURE 3.4: Methodological workflow from the input data (strain rate) **(a)** to the source classification (segmentation map) **(b)** and the performance of the classification (score map) **(c)**.*

3.2.4.3 Markov Random Field

After getting the classification results from RF, a post-classification algorithm is performed using Markov Random Field (MRF). MRF is a spatial clustering algorithm developed for image processing (Lu et al., 2019), based on the use of a segmentation map combined with a probability of belonging to each possible class for each point (pixel) on the map (Cross and Jain, 1983). MRF relies on the consideration of two quantities for each pixel: the classification of neighbors pixels and the likelihood of the classification. To gather the two previous quantities, MRF defines an intermediate quantity called Energy E.

Two energies can therefore be defined for each class: E_{neigh} is the energy associated with the neighborhood, and E_{likeli} for the energy related to the probability of a pixel to belong to the class. The used convention is:

$$E_{\text{neigh}}(x) = \sum_{S \in V} \mathbb{1}_{\overline{x}} \left(\text{Class}(S)\right) \tag{3.1}$$

where V denotes the neighborhood, S a site belonging to the neighborhood, x the studied class, $\mathbb{1}$ the indicator function. In the case where the neighborhood is defined by the 8 nearest neighbors,

$E_{neigh} \in [0, 8]$ and is 0 if all neighbors are of class x, is 8 if none of the neighbors are of class x ; and:

$$E_{\text{likeli}}(x) = -\ln\left(L(\theta \mid x)\right) \tag{3.2}$$

where x denotes the class under study, θ the parameter vector describing the corresponding site observation and L the likelihood function. The combination of these two quantities is weighted by a parameter named $potential$:

$$E(x) = Potential \cdot E_{\text{neigh}}(x) + E_{\text{likeli}}(x) \tag{3.3}$$

The algorithm then iteratively updates the segmentation map by finding, at each iteration, the class x which minimizes the energy E.

In our method, MRF uses the classification and score maps provided for each signal unit and for each possible class by the Random Forest algorithm. The algorithm works iteratively: the classification map is updated considering the class of the neighborhood and the classification score associated with each signal unit. For our Markov Random Field model, parameters are set using the training part of the dataset by computing the evolution of the normalized accuracy (average of the proportion of well classified data per class) when the value of potential or the number of iterations vary. The study shows that in our case, a potential value of 0.61 is effective and 4 iterations of the algorithm is enough to achieve convergence (Figure 3.5). Using these parameters, the mean gain of sensitivity is evaluated to 4% on the training set. A third parameter, dependent on the size of the considered buffer, can be introduced for flow classification. This parameter enables one to choose the size of the considered map on which to apply MRF. Figure 3.5 shows that the processing with the buffer is fully efficient starting from four consecutive classification in time direction.

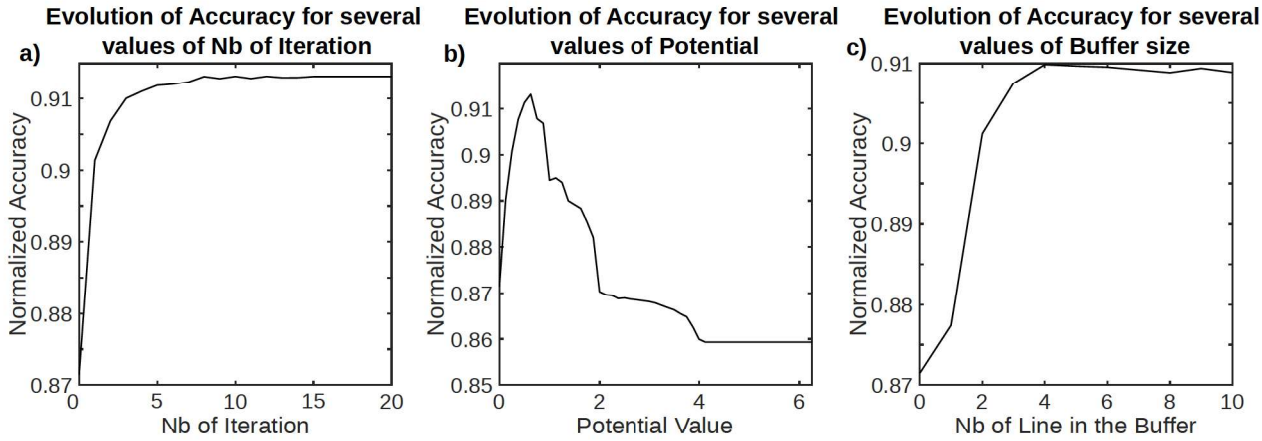


FIGURE 3.5: Influence of Markov Random Field parameters on the mean of the sensitivity of all classes (called normalized accuracy in the graphs). Evolution of the normalized accuracy in function of ***(a)*** nb_of_iter *parameter,* ***(b)*** $potential$ *parameter, and* ***(c)*** *number of time steps in the buffer. The optimal parameter which maximizes the accuracy is:* nb_of_iter *= 4,* $potential$ *= 0.61, and* $buffer\_size$ *= 4.*

3.2.4.4 Implementation of the Processing Chain for a Real-Time Processing

The previous sections are introducing the different bricks used in our process and fixing the different parameters for best results. In real-time application, the DAS signal is continuously acquired and is composed of a 4 s strain rate refreshed every 2 s. In our application, features are computed on

the currently acquired strain rate and directly injected into the Random Forest algorithm. Once the classification estimation and the associated scores are achieved, the results are stored into a buffer where the three last obtained results are also stored. Markov Random Field then post-processes the buffer content and reviews the small classification map saved in the buffer and updates the final segmentation map. We propose in Figure 3.6 a more graphical view of the processing chain described above.

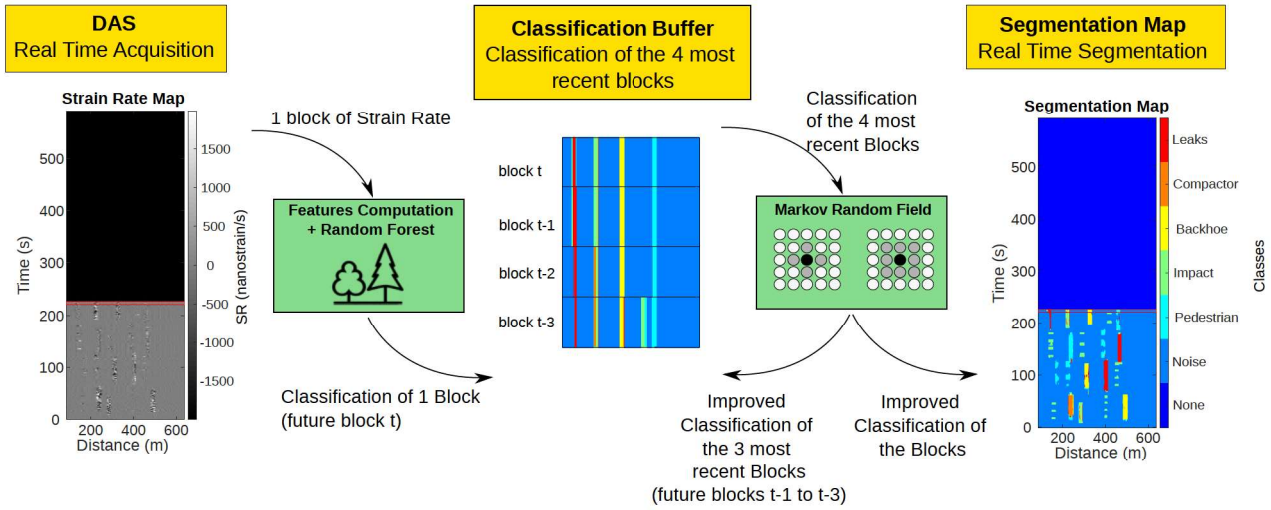


FIGURE 3.6: Overview of the implementation of the processing chain.

3.2.5 Results

The results can be examined using the scores calculated for each signal unit and for each class (Figure 3.7). Considering all signal units that actually belong to the same class, we average the scores obtained by the Random Forest. For example, for the "pedestrian walk" signals, the average score is 0.6, meaning that 60% of the trees in the RF model have voted for this class. At the same time, 16% of them vote for "Impact". The issue with this representation is that vote distribution cannot be considered, as only the mean appears. Figure 3.7 also contains a histogram representation of the scores of the true class for all signals which belong to each class (scores on the diagonal of the score matrix, 1 histogram per class). For classes entitled "backhoe", "compactor" and "leakages", the distribution is centered around 1, unlike for classes "pedestrian walk" and "impact" for which the score is more widespread.

The sensitivity of all classes with the single run of the Random Forest algorithms reaches on average 83.0%. The confusion matrix in Figure 3.8a compares for each signal unit the reference interpretation and the interpretation from the classification system. The percentages inside the matrices exhibit the distribution of signal unit classification for each available class. For example, if we consider only data of the class pedestrian, 77.0% of data are correctly classified, whereas 9.4% are classified as object falling, 6.0% as noise and 3.7% as leaks. In the case of background noise, only 0.4% of them are classified as a signal.

The use of Markov Random Field (MRF) improves the sensitivity of all classes on average to 87.1%, corresponding to an increase of 4.1% as estimated using the test set. The confusion matrix Figure 3.8b, achieved with MRF, shows progress for all classes. In our study, the highest gain of sensitivity by using MRF is for the "pedestrian walk" class (+8.3%), the "backhoe" class (+7.2%), the "compactor" class (+4.7%), the "Impact" class (+3.1%), the "water/air leakage" class (+1.2%) and finally the "noise" class (-0.1%).

Figure 3.9 presents the feature importance evaluated by the Random Forest algorithm. In our case, the most important feature concerns the skewness of the signal waveform (n°6), followed by the mean distance between the 1st quartile and the median of all Discrete Fourier Transforms (DFTs) as a function of time (n°51), the energy of the signal filtered in 50-75 Hz (n°15), the number of peaks in the DFT (n°29), and the number of peaks in the curve of the temporal evolution of the DFTs maximum frequency (n°47). The less important feature is the Ratio between ascending and descending time (n°3).

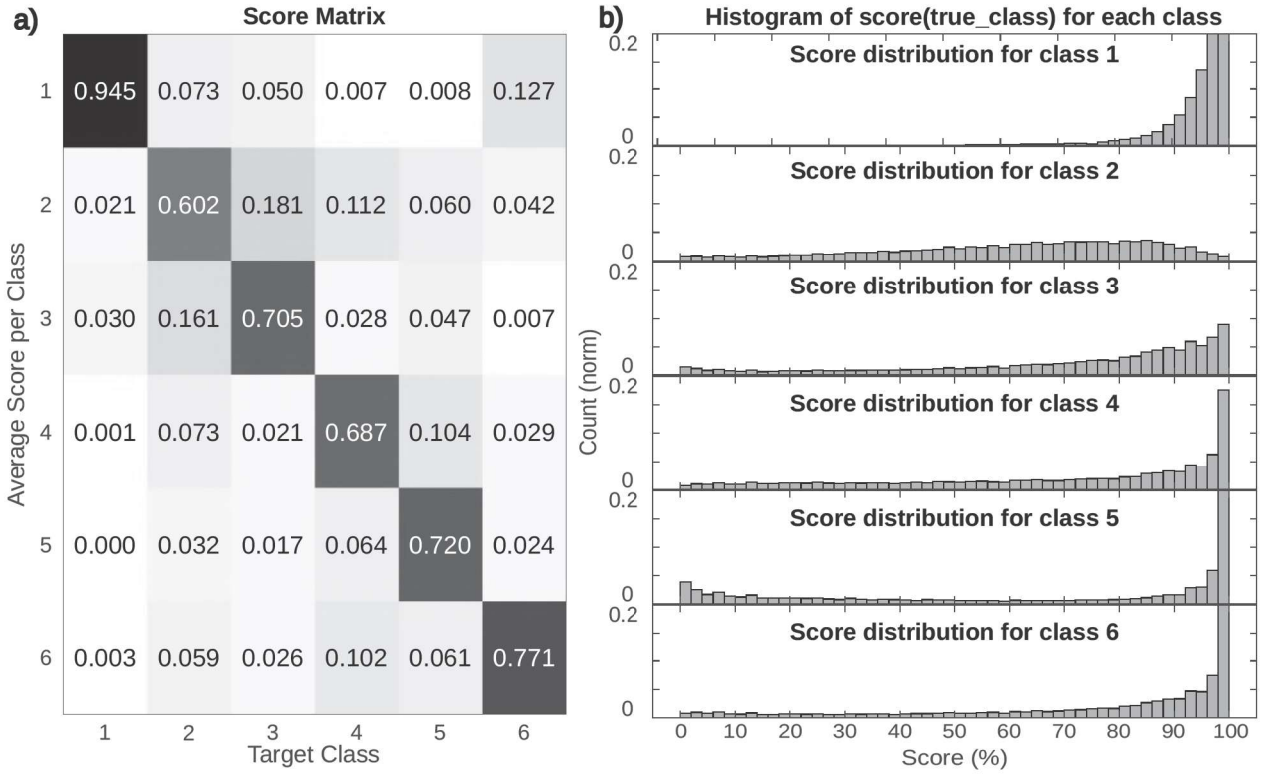


FIGURE 3.7: ***(a)*** *Score matrix showing for each true class the mean score after classification using Random Forest.* ***(b)*** *The histogram plots the distribution of score values for all true classes. Correspondence number class: (1) noise, (2) pedestrian walk, (3) impact, (4) backhoe, (5) compactor, and (6) water/air leakage.*

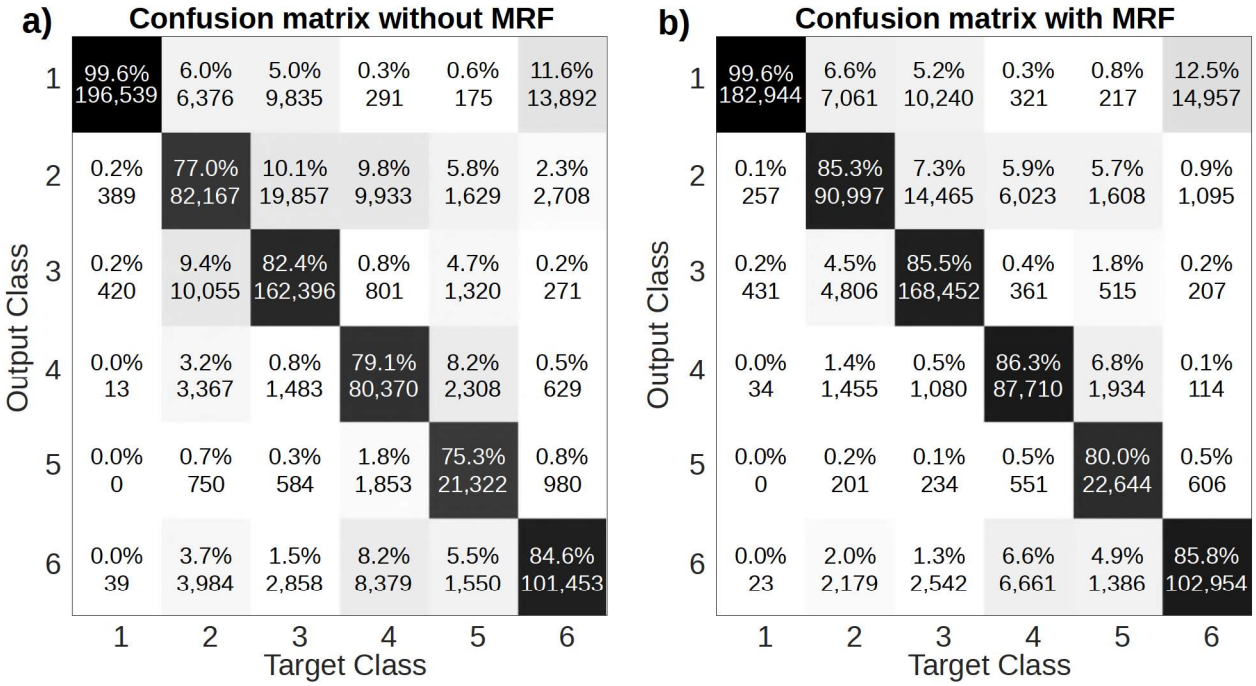


FIGURE 3.8: Comparison of confusion matrices ***(a)*** *with and* ***(b)*** *without the use of Markov Random Field algorithm. Correspondence number-class: (1) noise, (2) pedestrian walk, (3) impact, (4) backhoe, (5) compactor, and (6) water and air leakage.*

3.2.6 Discussion and Conclusion

We proposed a processing chain to efficiently classify DAS strain rate signals. Its improvements were achieved using a spatial clustering algorithm. The efficiency of the RF algorithm combined with our set of features are demonstrated, as the accuracy at the output of machine learning reaches 83% with a mean score higher than 0.7 for most of the classes. RF also highlights that waveform and spectrogram features also have their own importance even in the flux processing approach, as 3 features belonging to waveform or spectrogram families are more significant than the most significant spectral feature family and overall the waveform and pseudo-spectrogram have better scores. The addition of MRF raises the accuracy to 87%. Its main contribution is more visible on the segmentation map, as the aim is to avoid isolated misclassified points on the classification map. Two videos (Video S1, Video S2) have therefore been done and show the progressive segmentation of a strain rate in the process of being acquired (Electronic Supplement). In these videos, we can point out that the algorithm reacts quickly when an event occurs and is able to deal with spatially or temporally close events. This is thanks to the classification per trace process and the choice of a small enough window. Intuitively, a small window should not be suitable for long events, but we find experimentally that the classification results with a 20 s window is comparable. One explanation is that the algorithm considers the spectral and spectrogram feature and most long events have a regular frequency behavior over time.

Several classes have worse performance. In the case of pedestrian and impact classes, we can notice mutual confusion on the score and confusion matrix (Figure 3.7, Figure 3.8). The waveform of footsteps sound may look like a series of close impacts, and a solution could be to add more features to help the algorithm discriminating among these two classes. The compactor class is also confused with a backhoe signal. One possible explanation is that these classes are related to events produced by mechanical engines, and the confusion may therefore come from the motor noise itself.

These results are quite encouraging and open the way to many perspectives. In the short term,

one of the objectives will be to build and implement new features related to the spatial characteristics of the events and the displacement of the events, as this can help to improve the accuracy of classification results by integrating new information into the machine learning model. One way to do this would be to build on the features already used to temporally characterize the signal, and transfer them to the spatial domain, i.e. consider a signal along the space dimension. Although previous work involving the use of seismometers has shown that the crafted features are more affected by source physics than by signal propagation effects (Hibert, Provost, et al., 2017; Hibert et al., 2019; Maggi et al., 2017; Provost et al., 2017; Wenner et al., 2021), in-situ data should be integrated to enhance our model, as the behavior of the seismic signal may differ according to the environmental parameters: propagation medium (dirt, sand, stone, etc), humidity, temperature, etc. Other sources can also be implemented in our model and the resulting model can be evaluated using similar tools as introduced in this paper. A field test during a short duration time of the developed processing chain may also be considered.

In the medium term, a study of the influence of time-frequency features on the classification could be performed, in order to understand the seismic behavior of the different sources and to be able to extend the method towards semi-supervised or even unsupervised classification methods. For this task, the feature importance measurement provided by RF is a good starting point.

In the long term, the use of machine learning based processing chain can also help to automatically increase the catalog of events recorded along the fiber. This would overcome one of the major difficulties associated with deploying DAS for continuous monitoring over long periods of time, which is the amount of data to be processed and stored. A processing chain that allows to process in-situ the data and to store only the interesting events would make this long-term monitoring possible, which would allow to exploit the full capacity of the fiber optic for the monitoring of infrastructures, but also of geological objects such as landslides, volcanoes or tectonic faults.

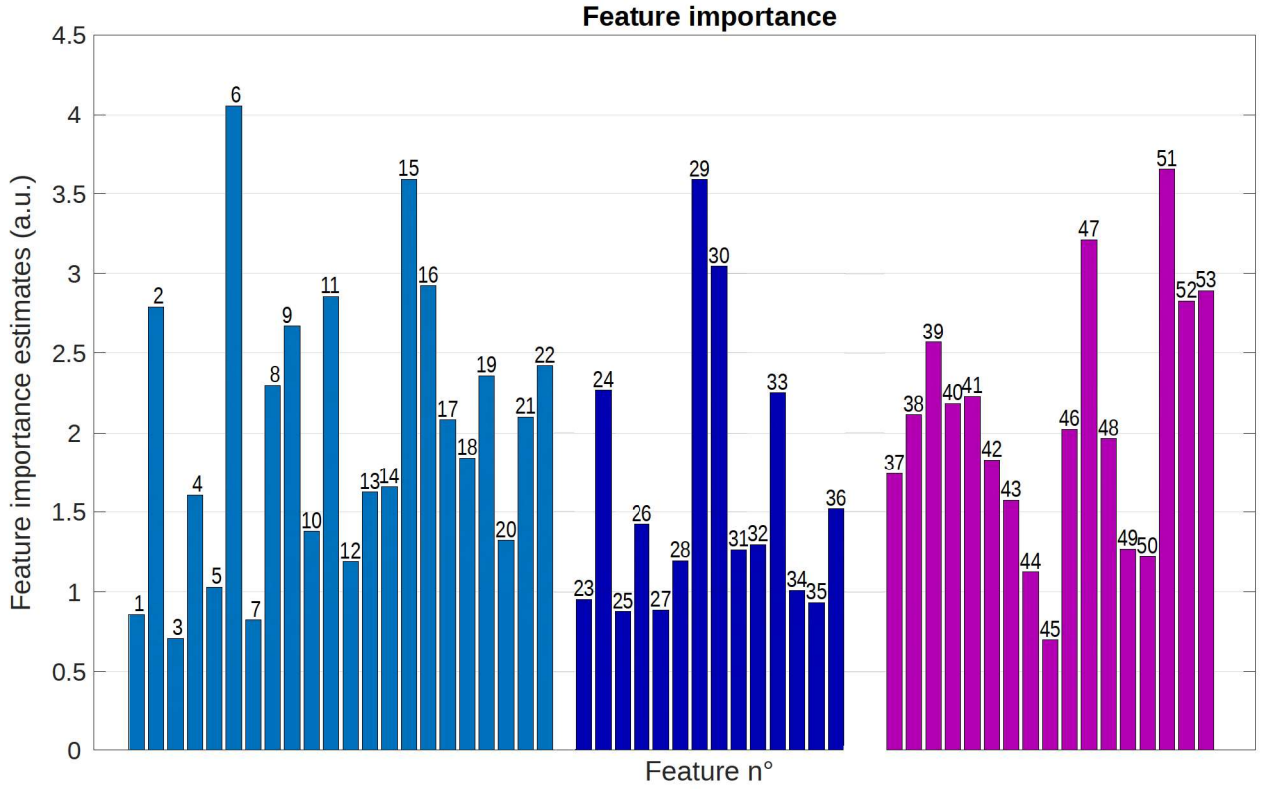


FIGURE 3.9: Feature importance evaluated with the Random Forest algorithm.

3.3 Chapter Summary

In this chapter, we propose a processing chain that combines human-engineered features, primarily used to classify natural seismogenic events measured with conventional seismometers (Provost et al., 2017), with a Markov random field algorithm to account for spatial redundancy by considering neighboring data points (Cross & Jain, 1983). The processing chain is designed to operate in a continuous data stream, eliminating the need for a separate detection algorithm. All tests were conducted at the FEBUS Optics test center. The approach applied to the FEBUS Optics test center database yielded promising results, achieving a classification accuracy of 87% for distinguishing between six event classes: pedestrian, impact, backhoe, compactor, and water or air leakages. Building on the successful results and the publication of this article, we implemented the system for embedded processing within the FEBUS A1-R DAS instrument at the FEBUS Optics test center, in collaboration with FEBUS Optics engineers. Additionally, we used a portable workstation for data processing. One of the challenges in embedded applications is the time and computational power required to compute the features. To address this challenge, we used the feature importance measure provided by Random Forest. This allowed us to reduce the feature list to a more manageable set of about a dozen features. During client demonstrations, we were able to showcase that the system could autonomously detect events with both accuracy and relevance, without the need for human intervention.

However, the study suffers from several limitations. The tests were conducted in a controlled environment with minimal external disturbances and a limited trench length of only 22 m. These factors raise questions about how well the test-bench simulation reflect field conditions, where various noise sources would be present. From a computational perspective, while spatial information was incorporated into the processing chain and decision-making process, it was not fully integrated within the machine learning algorithm. The current machine learning approach does not account for spatial patterns, which presents a limitation in more complex, field environments. Although the dataset allowed for testing the consideration of spatial redundancy in the signals, it was not ideal for exploring spatial patterns due to the limited scope of the experiment.

The next chapter will address these limitations. We will introduce a new dataset, acquired using 91 km fiber-optic cable in a real-world field conditions. The processing chain will also be applied to a different purpose: detecting earthquakes and quarry blasts, despite noise primarily of anthropogenic origin. Additional features will be added to complement the current feature set, allowing for more effective use of spatial information.

Chapter 4

Integrating the Specificities of DAS Data into the Classification Process

4.1 Introduction

In conventional seismology, event labeling relies on analyzing both the waveform and spectrogram to identify key features of seismogenic activity. When working with Distributed Acoustic Sensing (DAS) data, spatial information adds another layer that helps distinguish between different types of events. For example, when visualizing the DAS strain rate (SR) time-distance representation, moving seismogenic sources appear as oblique lines. By measuring the slope of these lines, we can deduce the apparent velocity of the signal source (Figure 4.1). This method is useful in applications like monitoring earthquake apparent velocity (Lellouch et al., 2019; Lior et al., 2021), volcanic events (Jousset et al., 2022), monitoring fluid velocity (Paleja et al., 2015) or monitoring road and rail traffic (Liu et al., 2020; Wiesmeyr et al., 2020; van den Ende et al., 2021; Wiesmeyr et al., 2021; Zhang et al., 2022). Additionally, the spatial footprint of seismic waves provides information about the source. Surface events, such as most anthropogenic seismogenic activities, typically have a small spatial footprint, from several tens meters for vehicles (van den Ende et al., 2021) to several thousands of meters for quarry blasts (Fang et al., 2020). In contrast, events occurring deeper beneath the surface, like earthquakes, can extend over several tens of kilometers.

In the previous chapter, we used spatial redundancy to improve the temporal classification process through Markov Random Field in the post-processing step. However, we did not account for the spatial seismic shape or trace-by-trace comparison. The dataset was acquired under controlled conditions with a limited fiber length, which differs significantly from real-world field conditions. In this chapter, we address two challenges by integrating spatial information-based features and trace-by-trace comparison, applying these methods to process a real-world field conditions, full-scale DAS dataset. These new features are added to the machine learning process to improve event classification. As in the previous chapter, the processing chain is split into two steps: classification using XGBoost and post-processing with a Markov Random Field (MRF). To assess the performance of this approach, we use data collected from a 91-km fiber-optic cable deployed in the French Hautes-Pyrénées. The dataset includes various seismogenic events, such as natural earthquakes, quarry blasts, and anthropogenic activities, allowing us to test and validate our method.

This chapter consists in one published paper (section 4.2):

Huynh, C., Hibert, C., Jestin, C., Malet, J. P., and Lanticq, V. (2025). *A real scale application of a novel*

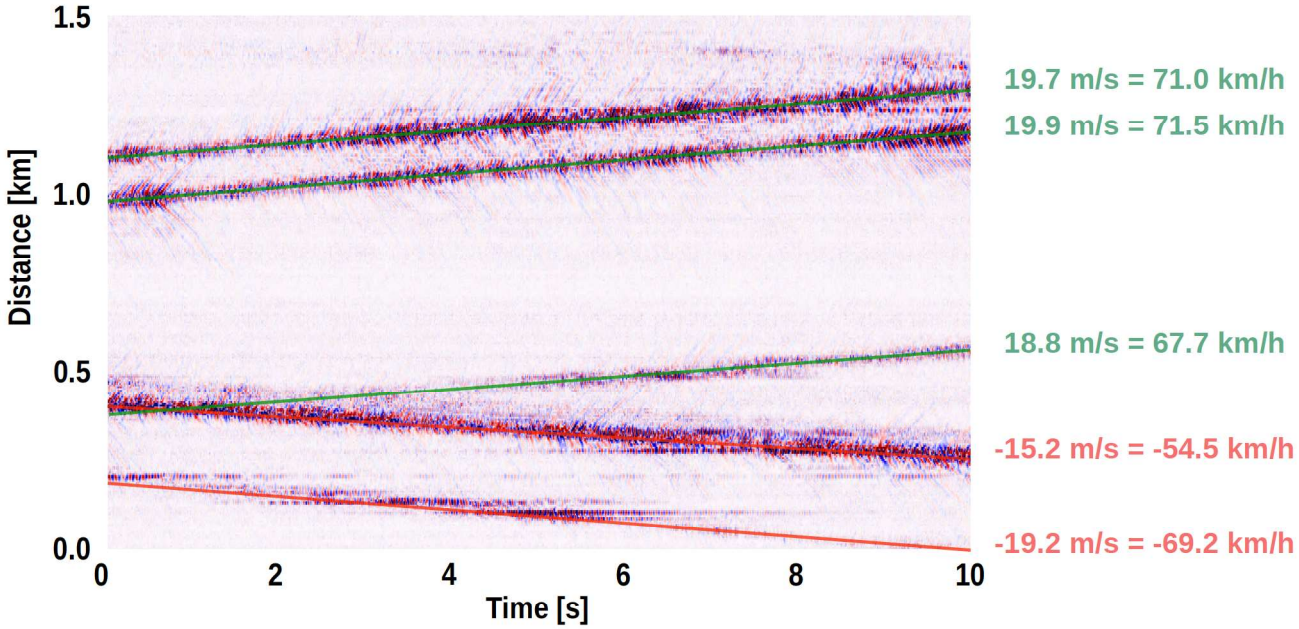


FIGURE 4.1: Plot of moving vehicles with their measured speed.

set of spatial and similarity features for detection and classification of natural seismic sources from distributed acoustic sensing data. Geophysical Journal International, 240(1), 462-482.

4.2 Paper: A Real Scale Application of a Novel Set of Spatial and Similarity Features for Detection and Classification of Natural Seismic Sources from Distributed Acoustic Sensing Data (Geophysical Journal International, 2025)

4.2.1 Abstract

Distributed Acoustic Sensing (DAS) turns a fiber optic into a very dense network of equally-distributed seismic sensors. We focused on the high-density sampling of the seismic wavefield, expressed in strain rates, measured by DAS. Classical approaches used to identify seismic signals rely on the recorded features at one station, but it is difficult to include spatial information in case of dense seismic station networks. This work aims at introducing new spatial and similarity features for seismic event classification suitable to analyze DAS observations. We propose a processing chain based on the XGBoost algorithm and the use of specifically designed spatio-temporal and similarity features for the event classification, and Markov Random Field for the spatial clustering. The methodology is designated to be applied on a continuous stream of DAS observations. We tested our processing chain to detect earthquakes and quarry blasts recorded in the region by permanent seismic networks and included in the RENASS catalog. These events are part of a strain-rate seismic survey carried out during a 3 weeks campaign of DAS measurements along à 91 km fiber optic cable deployed in the central Pyrenees mountains (France). Despite the high anthropogenic activities along the fiber optic path, the proposed method succeeded in detecting earthquakes of magnitude >0.4 and quarry blasts of magnitude >1.0 while limiting the number of false alarms. This performance is particularly noteworthy for low-magnitude events, where detection is accomplished despite a lower signal-to-noise ratio compared to traditional seismometers. The methodology opens the door to real time detection and classification of seismic events measured with long-distance fiber optic systems.

4.2.2 Introduction

The Pyrenees mountains, located in Western Europe between France and Spain, are the result of the collision of the Iberian microplate with the Eurasian plate at a rate of 0.85 mm/year (Fernandes et al., 2007). Considered as the second most active seismic zone in France after the Alps, the area has been monitored for the past 25 years by a geographically well-distributed seismological network, which has led to a better understanding of the tectonic processes in the region (Souriau & Pauchet, 1998; Rigo et al., 2005; Lacan & Ortuño, 2012; Sylvander et al., 2022). Nowadays, monitoring is carried out by a permanent network of about thirty seismic stations managed by the "Réseau de Surveillance Sismique des Pyrénées" of the French Observatoire Midi-Pyrénées. Seismic events are continuously detected and cataloged by the French facilities “Bureau Central et Sismologique Français" and "REseau NAtional de Surveillance Sismique" (BCSF-RENASS). Within a 50 km radius around Lourdes, located in the Central Pyrenees, six permanent stations are available, making the use of new families of sensors like Distributed Fiber Optic Sensing (DFOS) an opportunity to create a more exhaustive catalog of seismic events for the territory.

Using a laser pulse propagating along a fiber optic cable, Distributed Fiber Optics Sensors (DFOS) are able to measure various changes in the neighborhood of the fiber. The intrinsic presence of asperities on the fiber core results in light back-scattering effects that hold information about different physical properties. Among the DFOS family, Distributed Acoustic Sensing (DAS), based on Rayleigh back-scattering and phase shift, is sensitive to seismic acoustic waves crossing the inter-

rogated fiber optic cable. Spatial sampling, corresponding to the spacing between each point of measurement along the fiber, can be set down to a few tens of centimeters and acquisition rates can be set up to hundreds of kHz. Previous studies have demonstrated the contribution of spatial analysis for the analysis of DAS records in order to take into account the spatial character of the measurement. Spatial filtering, such as F-K filtering (Hudson et al., 2021; Fukushima et al., 2022), or curvelet decomposition (Atterholt et al., 2022) can be applied for data denoising. Further, spatial redundancy makes the DAS system suitable to capture low-magnitude events in delimited geographical areas, and a high diversity of low magnitude environmental sources (landslides, avalanches, floodings, ...) difficult to detect with permanent seismic stations but crucial to reach a better understanding of the regional seismicity of a territory. Further, DAS have proven to yield interesting new results in different contexts and for many different applications, for marine geophysics(Sladen et al., 2019; Spica et al., 2022), seismic imaging (Lindsey et al., 2017; Zeng et al., 2022; Young et al., 2022), volcanology (Jousset et al., 2022) and water reservoir monitoring (Zhu et al., 2021).

The dense measurement network and the high sensitivity of the DAS allow recording a large number and variety of seismic sources, making manual identification difficult and time-consuming. Classical automated detection and identification, based on short-time-average over long-time-average method (STA/LTA) or energy threshold, are complex to implement: the high diversity of events occurring along a fiber, as well as the increase of noise level with increasing distance, can generate high energy seismic signals making difficult the choice of suitable detection parameters. For natural event detection, instrumental noise and anthropogenic events should be filtered (Nayak et al., 2021; Chen et al., 2023).

Artificial Intelligence (AI) techniques are the most promising solution for DAS data automated classification. Two main families of AI algorithms for classification purposes exist: feature-based algorithms, also named machine learning algorithms, and deep-learning-based algorithms. Deep learning algorithms rely on the structure of deep neural networks to build their own discriminating features for the classification and detection of certain types of events such as micro-earthquakes (Binder & Tura, 2020) or footsteps (Jakkampudi et al., 2020). It is also possible to tackle topics difficult to solve with features, such as signal denoising (Ende et al., 2021; Zhao et al., 2021). Deep learning approaches require a large dataset for proper model training, and do not allow an access to a deep understanding of the algorithm decision criteria. They also require appropriate data representation, both in terms of the choice of input data and the scale at which they are represented. In contrast, feature-based methods rely on the use of human engineered seismic signal features which have been defined in previous studies during different studies of seismic event discrimination. Temporal-related features had been widely used in seismic data analysis (Hibert, Mangeney, et al., 2014; Hibert, Provost, et al., 2017; Hibert et al., 2019; Maggi et al., 2017; Provost et al., 2017; Chmiel et al., 2021; Domel et al., 2023). In addition to waveform analysis, features are often derived from signal transformation into the frequency domain using Fast Fourier Transform (Wiesmeyr et al., 2020; Tejedor et al., 2021a), wavelet decomposition (Wang et al., 2019), or the use of the mel-frequency cepstral coefficient (Bublin, 2021). Because of their ability to quantify individual contributions of each feature used for the classification, various feature-based AI algorithms are suited to build a must-have feature list adapted to the events we want to identify. This is the case for Random Forest and XGBoost machine learning algorithms (Breiman, 2001; Chen & Guestrin, 2016).

Huynh et al. (2022) proposed a methodology enabling to classify DAS data streams using seismic signal derived features that present interesting results when implemented for a 22 m-long trench for classifying anthropogenic event sources. In this article, we aim to apply a similar workflow for detecting and identifying earthquakes and quarry blasts under real-world field conditions. Expanding our preliminary processing chain to work a the regional scale is challenging due to 1) the variety of sources in an area spanning dozens of kilometers; 2) the spatial extension of the array, which lead to

the recording of numerous synchronous seismic sources; and 3) the background noise level which can be very different at each part of the buried fiber optics (when going through dense habitation area, along roads, along rivers, etc.). We propose in this study a new processing chain, which use new seismic signals features designed specifically to benefit from the dense spatial distribution of DAS data, to overcome those difficulties.

4.2.3 Fiber Optic DAS Data

Data acquisition has been achieved along a 91 km long fiber optic deployed between Lannemezan and Luz-Saint-Sauveur in the Bigorre area, located in central French Hautes-Pyrénées. The fiber is deployed close to several urban centers (Bagnères-de-Bigorre, Lourdes, Argelès-Gazost) and close to several quarries (for rock extraction) in activity. The cable also follows national and departmental roads (Figure 4.2). Seismic data measured by the DAS, named Strain Rate (SR) and expressed in nanostrain per second (nstrain/s) are extracted from the optical raw data by setting the optical-acoustic processing parameters prior to acquisition, known as gauge length and derivation time (Hartog, 2017). Setting these parameters is important because they have a direct impact on the measurement quality: the derivation time affects the maximum observable frequency, while the gauge length affects the spatial resolution and then the minimum detectable wavelengths. To fit with the observation of natural seismic events, SR has been recorded using a gauge length fixed to 10 m and a cut-off frequency of 200 Hz. Because of instrumental response biases, common mode frequency removal has to be carried out prior to data exploitation. Thus, average removal is performed on the strain rate raw data for each channel of measurement.

With a fixed gauge length of 10 m and assuming a 0.2 dB/km attenuation rate of the optical laser pulse along the fiber optic cable, the maximum distance range achievable is 50 km. To extend beyond this range while maintaining a favorable signal-to-noise ratio, additional optical devices can be used in combination with an interrogator such as repeaters, dual-channel interrogators, or long-range systems. The FEBUS A1-R DAS has been installed in the town of Lannemezan, supplemented by a FEBUS range extension module connected at the middle of the fiber, in Lourdes. Recording has been carried out continuously over 3 weeks between August 30 and September 20, 2022. The final acoustic DAS data are recorded with a spatial sampling of 4.8 m. This equates to 18,958 point sensors, called channels, equally spaced along the 91-km fiber optic. The DAS data size consists of 40 TB.

A first observation of the acquired data indicates the presence of a large number of anthropogenic events in the monitored area along the entire length of the fiber optic cable: we observe the presence of a day-night cycle with an increase in anthropogenic activity during the day, and clearly identify the presence of seismic signal corresponding to vehicles circulating on roads (Figure 4.3). However, in the absence of appropriate detection techniques for DAS systems, the identification of natural earthquakes and other environmental natural and anthropogenic seismogenic events is difficult.

We use the BCSF-RENASS catalog to build a reference catalog of seismic events. BCSF-RENASS provides an open-access catalog of natural and anthropogenic events (earthquakes, quarry blasts, induced seismicity, explosions) recorded by permanent seismic stations deployed in France and in neighboring countries. By cross-referencing the seismological waveforms between the stations, events are detected and localized automatically. Human operators then validate the detection, the location, and determine the natural (for example earthquakes, landslides, rockfalls) or anthropogenic (for example quarry blasts, explosion) origin of the event.

32 events (8 quarry blasts and 24 earthquakes of magnitude Mw estimated between 0.3 and 2.4)

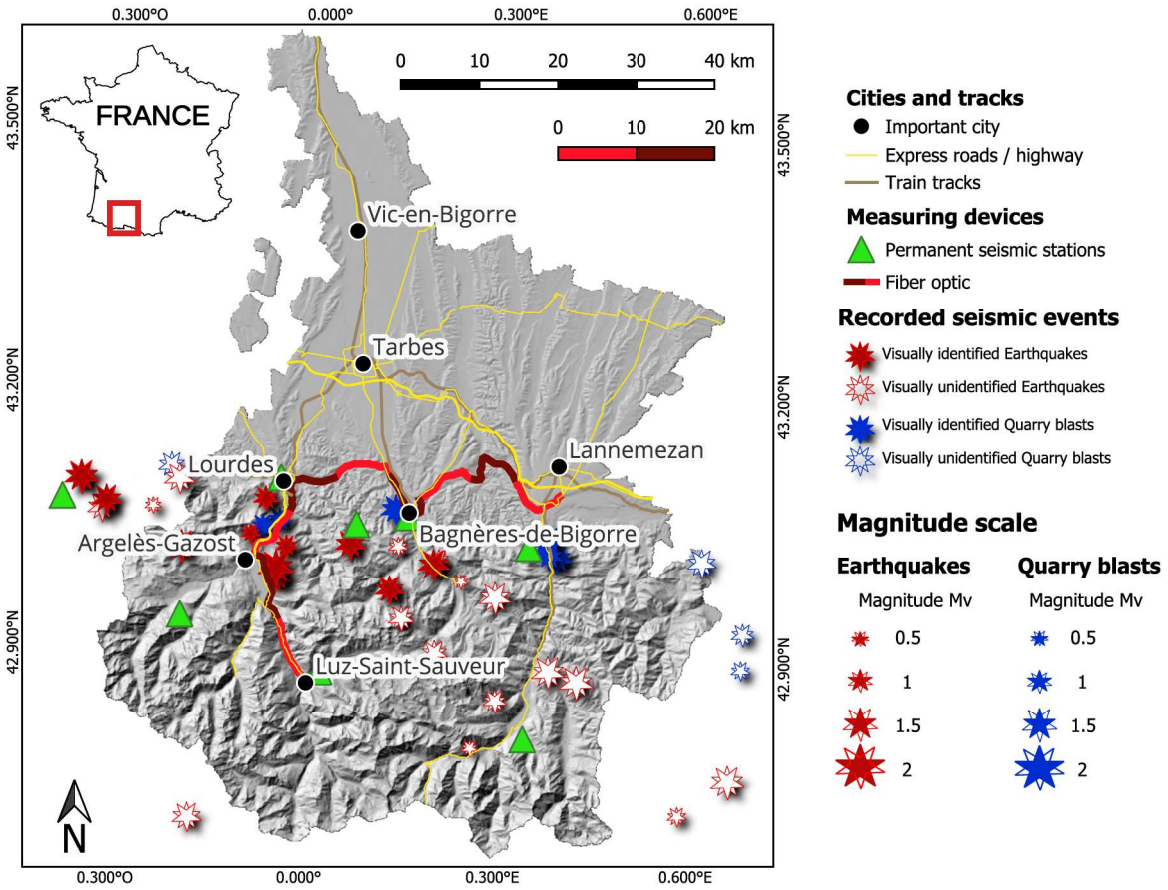


Figure 4.2: Fiber optic setup in the French area "Hautes-Pyrénées". The background corresponds to the topography of the area. The map includes location of main cities, roads and train tracks, as well as the positions of the cataloged earthquakes and quarry blasts from the BCSF-RENASS (using the permanent seismic stations represented by green triangles). The visually unidentified events on DAS data are represented by white-filled stars.

have been detected and listed in the BCSF-RENASS catalog for the study period. Among these events, we visually confirmed the presence of 13 earthquakes and 6 quarry blasts in the DAS recordings (Figure 4.2), with data available online (see "Data Availability" section). It appears that most of the unobserved events on the processed DAS data occur beyond Bagnères-De-Bigorre in South-East direction. The absence of detection can have multiple explanations, based on the instrumental spatial configuration and geometry of the fiber optic cables (fiber optic is insensitive to seismic waves that arrive perpendicular to the fiber optic) or on the local geological conditions (nature of the geological media can modify the propagation of the seismic signal in the ground). For our analysis, we focus on the classification of three classes of events: "earthquakes", "quarry blasts" and "noise". "Noise" groups all uninvestigated events (daily anthropogenic events as moving vehicles, wildlife events, aerial sections of fiber optic cable or instrumental noise). For purpose of detection, we will refer to earthquakes and quarry blasts as "events", and the other sources, such as uninvestigated anthropogenic noise, as "noise".

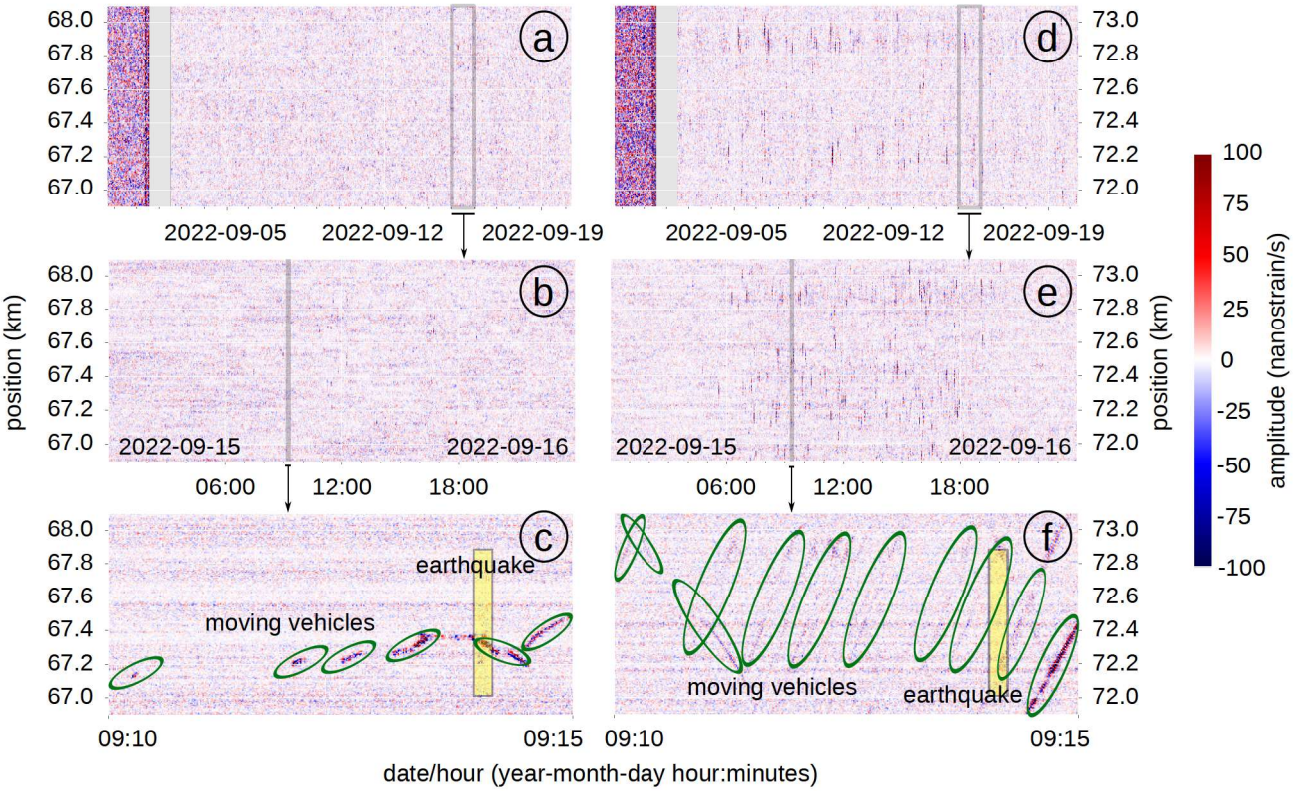


*FIGURE 4.3: Samples of SR records on the Hautes-Pyrénées DAS dataset for several time windows, in the countryside and close to a city (GPS coordinate: 43.01298 °N, -0.08660 °E). **(a)** represents the data recorded over the whole acquisition period, at position between 66.9 km and 68.1 km (countryside), **(b)** shows the data acquired over a period of 24 h around the same section, and **(c)** the data recorded over a period of 300 s around the same section. **(d,e,f)** represent similar records at position between 71.9 km and 73.1 km (close to Argelès-Gazost). Moving vehicles are delimited by green circles and earthquakes by yellow rectangles inside the 300 s view. The 0 km-position reference is located at the end of the fiber optic cable point close to Lannemezan.*

4.2.4 Methodology

4.2.4.1 Event Picking and Identification for Labeling

The massive seismological dataset recorded during this study requires high-performance analysis tools to detect events of interest. Identifying them by direct observation of the SR is time-consuming

especially when we are also interested in detecting low-magnitude events or in identifying certain types of event. We therefore work with a representation called Energy Band (EB), obtained for each channel along the fiber by summing the energy contained in the spectra (calculated using a Fast Fourier Transform - FFT) of a seismic trace window ($FFT_{trace}(f)$) between two frequency bounds f_0 and f_1:

$$E_{[f0,f1]} = \sum_{f \in [f0,f1]} FFT_{trace}(f) \tag{4.1}$$

The spectra are calculated using a sliding window of 2 s, with a shift of 0.5 s. f0 and f1 can be adapted using frequency response knowledge. In our analysis, values are chosen to cover the maximal range of the frequency content: f1 corresponds to the Nyquist frequency (100 Hz), whereas f0 corresponds to 1/window_width (0.5 Hz) . The resulting representation can be displayed into a new grid whose general representation is given in Figure 4.4: (d_{step} , t_{step}) refers to the spatial and temporal sampling of the EB representation, whereas (d_{win} , t_{win}) refers to the window on which is computed each point of the EB representation. In our case, as the energy is computed for each channel, spatial parameters d_{step} and d_{win} are taken as the distance between consecutive channels. t_{step} and t_{win} are respectively fixed to 0.5 s and 2 s. EB representation improves the visual detection of low-magnitude events because of the integration of the spectrum content over a duration of 2 s. For this reason, we keep both EB and SR representations in the following figures. The reader should keep in mind that the EB with these window parameters is not involved in the processing chain.

19 of 32 seismic events recorded by BCSF-RENASS are visually detected using the EB representation. They are listed in Table 4.1 and segmented from the rest of the signal. Segmentation grid resolution matches the EB resolution (4.8 m, 0.5 s). Segmentation process is achieved by drawing a rectangle box around the EB where an event is located: the points inside the box are labeled as events, whereas the points outside are considered as noise.

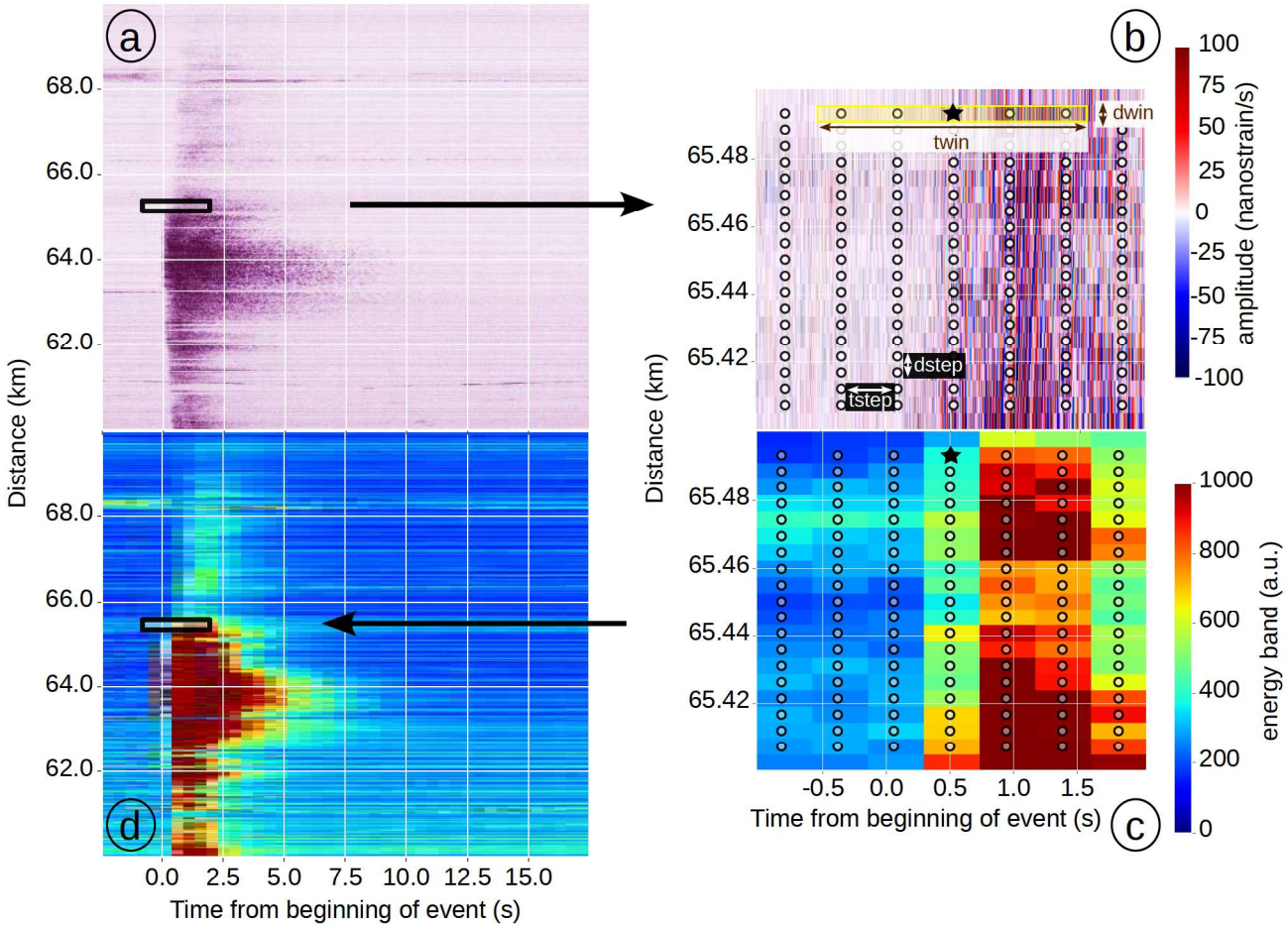


*FIGURE 4.4: EB representation from SR. The mapping is described inside a small part of the DAS data, delimited in **(a,d)** by a rectangle. We define a new grid above the SR representation represented in **(b)** by circles. Each circle is separated by t_{step} and d_{step}. Taking one element of the grid (for example the black star in **(b)**), energy is computed using the SR contained inside a window centered on the element (represented by yellow rectangle for the star element in **(b)**) and which size is given by (t_{win}, d_{win}). In the EB representation, each point corresponds to the computed energy at each grid position **(c)**.*

TABLE 4.1: Table listing all events identified on the DAS dataset. An identifier (id) is assigned, and information on event type (Class), magnitude (Mw) and measurement time (Time) are provided. The projected position of the epicenter on the fiber ($D_{\parallel fiber}$) and the distance from the epicenter to the fiber ($D_{\perp fiber}$) are used to check the consistency of what is observed. The reader can refer to Appendix B1 for the SR and EB visualization of all the listed events.

id	Class	Mw	$D_{\parallel fiber}(km)$	$D_{\perp fiber}(km)$	Time (UTC)
QB1	QB	0.7	59	1.8	2022-08-31 at 08:02:39
QB2	QB	0.8	3.4	5.8	2022-09-01 at 09:26:00
EQ1	EQ	0.6	59	1.7	2022-09-03 at 03:50:17
EQ2	EQ	1.0	30	10	2022-09-03 at 13:13:41
EQ3	EQ	2.4	62	85	2022-09-03 at 18:27:47
EQ4	EQ	1.4	61	25	2022-09-05 at 00:51:58
EQ5	EQ	1.2	30	6.5	2022-09-06 at 07:58:54
QB3	QB	1.0	3.5	4	2022-09-06 at 10:10:58
QB4	QB	1.1	32	1	2022-09-08 at 10:03:36
EQ6	EQ	0.8	53	3.2	2022-09-08 at 17:10:59
EQ7	EQ	2.0	61	50	2022-09-09 at 07:07:40
EQ8	EQ	0.4	61	1.5	2022-09-09 at 17:37:59
QB5	QB	0.6	57	0.5	2022-09-12 at 10:07:43
QB6	QB	1.1	3.5	6.3	2022-09-14 at 09:26:07
EQ9	EQ	1.1	57.5	8.5	2022-09-15 at 09:12:16
EQ10	EQ	1.6	66	1.4	2022-09-16 at 11:16:50
EQ11	EQ	1.1	61	20	2022-09-16 at 23:12:33
EQ12	EQ	0.8	68	1.0	2022-09-18 at 04:18:12
EQ13	EQ	1.3	63	9.5	2022-09-20 at 04:39:26

4.2.4.2 Processing Chain

Figure 4.5 details the processing chain. The input corresponds to the measured SR after baseline correction (average removal of each channel) (Figure 4.5a) and the output is the associated classification map encoded in three colors (Figure 4.5d). Initially, the continuous data stream is processed with a sliding window approach (Wenner et al., 2021). The sliding window uses two parameters that define the spatial and temporal sizes (d_{win} and t_{win}) and that defines the shift (d_{step} and t_{step}). The choice of the window size and the impact on the outcome of the processing chain is discussed in section 4.2.5.3.

Data contained in a window is labeled and pre-processed to extract a vector of 111 features described in section 4.2.4.3. Concatenating the vectors obtained at every spatio-temporal window centroid position outputs a feature matrix that is injected into the machine learning algorithm XGBoost for event classification. XGBoost produces a classification score for each window that can be combined to create a score map (Figure 4.5b). An initial classification map is generated using a threshold of 0.95 on the score map to reduce the amount of misclassification (Figure 4.5c). Then a final post-processing step (Markov Random Field) aggregates the classification, the score of each individual window, and the classification result of neighboring windows, to keep only the most significant classification results (Figure 4.5d).

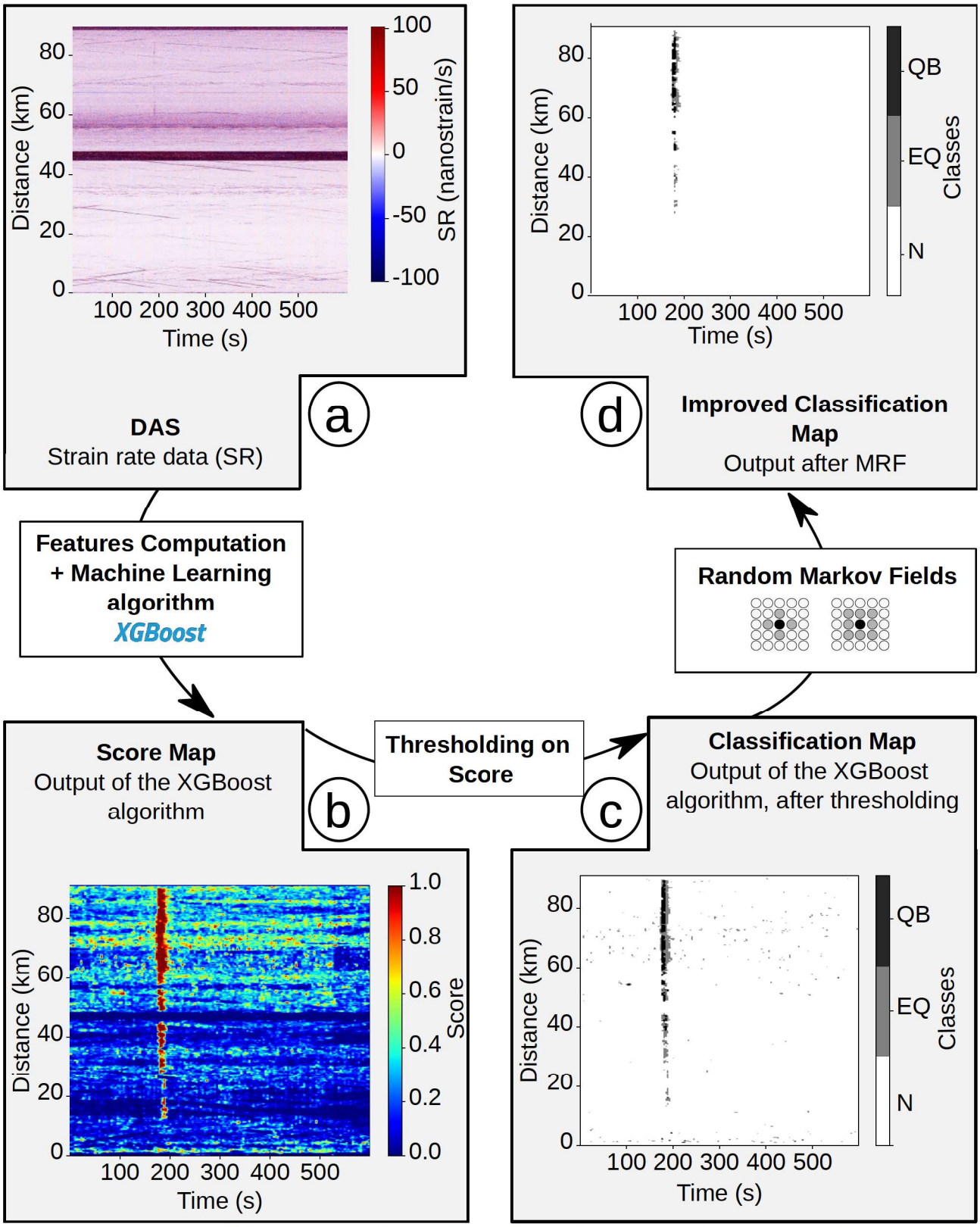


FIGURE 4.5: Overview of the processing chain. The SR data ***(a)*** *is introduced in a machine learning model and produces one score map per class* ***(b)****. Points with a score value higher than 0.95 are kept and produces a classification map* ***(c)****. A post-processing algorithm, Markov Random Field (MRF), is applied on the classification map to reduce the amount of noises* ***(d)****. N refers to noise class, EQ to earthquake and QB to quarry blast. To simplify the representation, the score map represented here corresponds to the sum of non-noise classes.*

4.2.4.3 Data Pre-Processing: Signal Description Using Features and Labelizations

Machine Learning algorithms rely on the use of features. Features are specific measurable properties or characteristics extracted from the data that help the algorithm to make predictions. Our work focuses on a set of 111 features divided into three major categories: temporal features (63), spatial features (24) and similarity features (24).

Temporal features are derived from previous works on seismic signal classification using traditional seismometers recordings (Provost et al., 2017; Maggi et al., 2017; Hibert, Provost, et al., 2017; Hibert et al., 2019; Malfante et al., 2018; Wenner et al., 2021; Chmiel et al., 2021; Falcin et al., 2021; Domel et al., 2023) and DAS recordings (Huynh et al., 2022). The temporal features are divided in three families, named waveform, spectral and spectrogram features. Waveform features describe the evolution of the signal and its envelope in the time domain, using tools such as energy, skewness and auto-correlation measurements. Spectral features provide information on the frequency content of each of the windowed channels and measure, for example, the value of the median FFT, energy between different frequency bands or the width of the spectral centroid. Spectrogram features measure changes in spectral content over time, such as the variation of the median FFT over time.

Spatial features are introduced in this work to account for the distributed nature of the DAS measurements. Unlike the use of a dense array of seismometers, the distance between two consecutive virtual sensors is constant. Figure 4.6b illustrates the waveform profile of a trace measured at one channel, and Figure 4.6c the measured signal at a fixed time along the fiber. Observation of the SR (Figure 4.6a) shows that, depending on the magnitude of the source, a larger or smaller portion of the fiber may record the event. Low energy sources, such as cars, low magnitude earthquakes and quarry blasts, are recorded only by a limited portion of the fiber but larger energetic sources as earthquakes can be recorded over several tens of kilometers along the fiber.

The proposed spatial features are based on the shape of the seismic signal measured at a given time along all channels contained in the window. It includes average, standard deviation and auto-correlation such as the one computed for the temporal waveform feature category. These features are designed to account for the spatial coherence of seismic signals, the apparent velocity of different phases, and dispersion effects. All these seismic signals characteristics help distinguishing visually between an earthquake and a moving vehicle, for example. We will thereafter refer to temporal trace for seismic signals recorded at one channel along the fiber (red in Figure 4.6b), and to spatial trace for a seismic signal recorded at a given time on all the channels of the fiber (blue in Figure 4.6c).

Similarity features are computed from the comparison of traces taken at different positions along the fiber. We choose two different methods to compute the similarity: cross-correlation and Dynamic Time Warping (DTW). Cross-correlation estimates the optimal time shift between the signal arrival time at the compared traces. DTW constructs an optimal mapping (warping path) between corresponding points of two sequences, allowing for non-linear stretching or compression of the time axes to achieve the best alignment, thereby accommodating temporal distortions (Sakoe & Chiba, 1978; Berndt & Clifford, 1994; Keogh & Ratanamahatana, 2005; Müller, 2007). Kumar et al. (2022) have shown the interest of using DTW for geophysical applications, including a higher sensitivity to minor variation compared to classical cross-correlation. Examples of applications are associated with full-waveform inversion in reflection seismology (Ma & Hale, 2013), GNSS displacement time series analysis (Kumar et al., 2023) and classification of volcano-seismic events (Ida et al., 2022). The reader may refer to Appendix B2 for an item-by-item description of 111 used features introduced in the Machine Learning model.

Labeling each feature vectors is not trivial due to the resolution difference between the segmen-

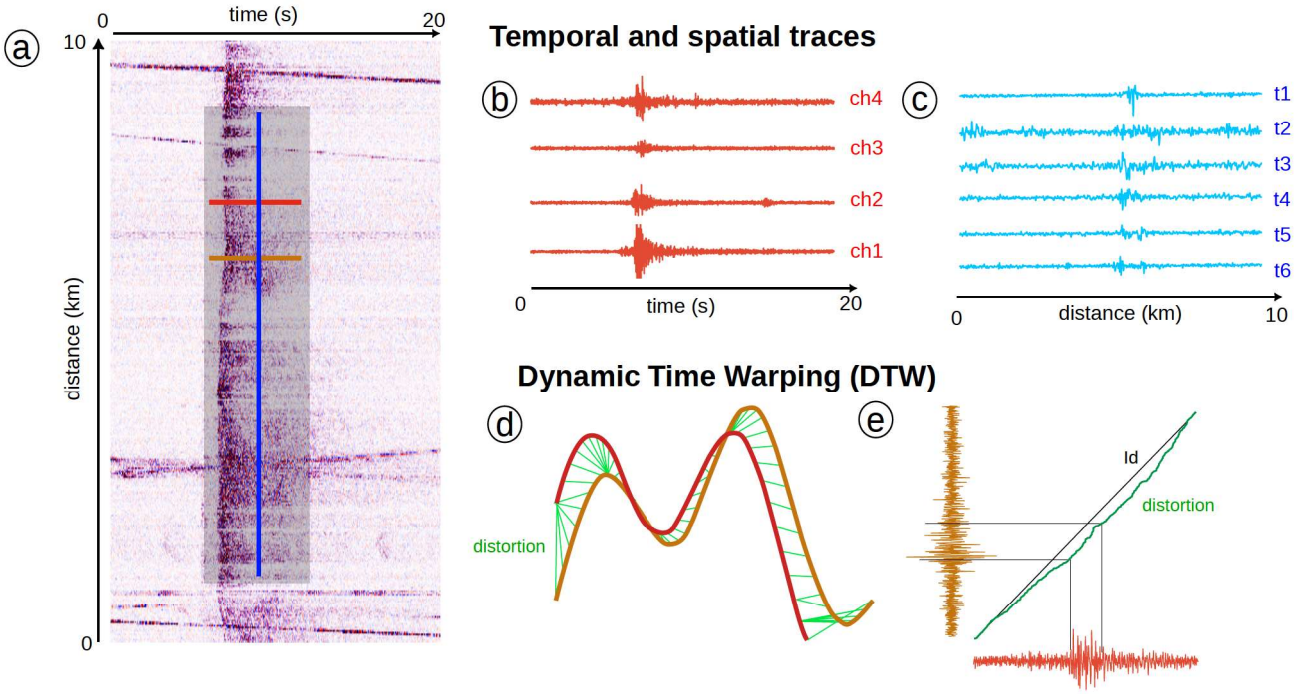


*FIGURE 4.6: From the SR **(a)**, features can be built by considering temporal trace observed at fixed position (**(b)**, represented in **(a)** with red and orange line) and spatial trace observed at fixed time (**(c)**, represented in **(b)** with blue line). Similarity features include DTW, built using an estimated distortion function that stretch two traces to fit together **(d,e)***

tation grid, as described in Section 4.2.4.1, and the grid used for feature vector computation. The labeling of each feature vector is determined by considering all the labels within the window used for feature computation. Given the segmentation grid resolution of 4.8 m by 0.5 s, and a window size of 998.4 m by 8 s, the feature vector labeling accounts for 3328 labels. If at least one of these labels corresponds to a "earthquake" or "quarry blast," the feature vector is labeled as either "earthquake" or "quarry blast," depending on which of these two classes has the majority of samples within the label grid.

4.2.4.4 Data Processing: Machine learning with XGBoost

We choose the machine learning algorithm called XGBoost as the identification tool in our workflow (Chen & Guestrin, 2016). XGBoost is a supervised ensemble learning method based on the use of a large number of weak learners (e.g. decision trees) to predict a class. Based on the boosting technique, decision trees are sequentially ordered. Misclassified samples made by one decision tree are kept for training the next one. Compared with other machine learning algorithms, features that poorly characterize the event do not reduce the performance of the XGBoost model. This algorithm is also able to deliver quantitative information about the influence of features in the classification. Several applications in geophysical contexts include rock facies classification (Zhang & Zhan, 2017), density log estimation for reservoir characterization (Zhong et al., 2020) and seismic source classification (Wang et al., 2023).

Once the training is completed, the XGBoost model uses a method based on vote analysis to provide a classification: each weak learner classifies the sample, and the class returned by the majority of them constitutes the output of the XGBoost model. A score can be deduced for each possible class by dividing the number of trees voting for one class by the total number of trees in the XGBoost model: a score close to 1 for a particular class indicates that most of the weak learners voted for

this class. Conversely, for a two-class problem, a score close to 0.5 indicates that the classification of an event has a very high uncertainty, as there are no clear majority in the vote casted by the weak learners. Scores are then a powerful indicator of the uncertainty of the classification and can be used to reduce the amount of false alarms.

4.2.4.5 Data Post-Processing: Markov Random Field (MRF)

Using a threshold on the score yielded by the XGBoost algorithm is a simple and efficient technique to improve the machine learning classification results. However, this technique does not consider the spatial and temporal relationships between feature vectors for data stream processing (Figure 4.4). As the classification is achieved for each SR windows which spacing is defined by the parameters t_{step} and d_{step} (Figure 4.4), we can assume that, in the case of a resolution at least two time smaller than the duration and the spatial footprint of an event, several neighboring windows include the same event. Neighboring classification is therefore a source of information that can be used in the Markov Random Field (MRF) method to mitigate uncertainties about event identification. MRF, originally used for image processing, is a spatial clustering algorithm relying on the classification score and the classification of neighbor points (Cross & Jain, 1983). The combination of these two parameters is achieved through a loss function C that gathers class neighborhood loss C_{neigh} and class likelihood loss C_{likeli}, defined as follows:

$$C_{neigh}(x) = \sum_{S \in V} \mathbb{1}_{\bar{x}}(Class(S)) \tag{4.2}$$

$$C_{likeli}(x) = \ln L(\theta|x) \tag{4.3}$$

$$C(x) = Potential \times C_{neigh}(x) + C_{likeli}(x) \tag{4.4}$$

where V is the neighborhood, x is one class among the output classes of XGBoost (here x takes a value among "noise", "quarry blast", or "earthquake"), $\mathbb{1}$ is the indicator function, θ is the parameter vector describing the corresponding observation (here associated with the features vector used in the XGBoost model), and L is the likelihood function. $Potential$ weights the contribution of C_{neigh} and C_{likeli} in the loss function C. In the case of a neighborhood defined by the 8 nearest neighbors, $C_{neigh} \in [0, 8]$ is equal to 0 if all neighbors are of class x, and to 8 if none of the neighbors is of class x.

The efficiency of the method has been demonstrated in previous studies for the real-time classification of anthropogenic events recorded by DAS interrogator on a controlled test bench (Huynh et al., 2022). In particular, it helps reducing the number of false alarms by avoiding isolated misclassified points on the classification map.

4.2.4.6 Performance of the Method

To define the performance of the method, we divided the DAS dataset in two parts: a training set, used to build the machine learning model and to define the parameters for feature computation ; and a test set used to measure the performance of the processing chain. During the three weeks of data acquisition, we recorded 19 events with several magnitudes and with hypocenters located at different geographical positions. Because of the small size of the dataset, we choose a leave-one-out cross-validation (LOOCV) evaluation technique, consisting in keeping a single event as a test set for

each cross-validation. Evaluation is performed for each event using the machine learning algorithm trained on the remaining events.

The method can be quantitatively validated by comparing the class predictions for each feature vector yielded by the machine learning algorithm to the manual classification of the corresponding portion of the DAS data. Quantitative criteria include multi-class precision, multi-class recall, and F1-score. These values are defined by the equations:

$$Precision = \frac{1}{N} \sum_{i \in Classes} \frac{TP_i}{TP_i + FP_i} \tag{4.5}$$

$$Recall = \frac{1}{N} \sum_{i \in Classes} \frac{TP_i}{TP_i + FN_i} \tag{4.6}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.7}$$

where N denotes the total number of classes, TP_i the amount of true positive considering class i, FP_i the amount of false positive, and FN_i the amount of false negative.

Qualitative validation consists of a visual inspection performed after reconstructing the classification map. Incorporating MRF as the final step in the processing chain helps to filter out areas with sparse detection and low classification scores, while preserving those with few detection but high relevance. As visual validation is prone to subjectivity, we present the results of each classification in Appendix B2.

4.2.5 Results

4.2.5.1 Impact of DAS Window Size Parameters

The dimensions of the window (t_{win}, d_{win}) used to compute the features have an impact on the results of the classification. We thus estimate the performance of our machine learning algorithm for several (t_{win}, d_{win}) pairs. For each pair, we use the LOOCV method for dataset splitting and the three criteria presented in section 4.2.4.6. The goal is to find the best pair that maximizes the F1-score (Figure 4.7c) derived from precision (Figure 4.7a) and recall (Figure 4.7b). Results given in Figure 4.7 show that the choice of a temporal window t_{win} smaller than 4 s, or higher than 20 s, degrades the recall. Similar conclusions can be drawn for the choice of a spatial window d_{win} smaller than 100 m. We note that the optimization of precision values is obtained for spatial windows d_{win} higher than 250 m. In the range $t_{win} \in [8, 12]$s and $d_{win} \in [500, 1250]$m, F1-score value is the highest with values between 0.65 and 0.69. Choosing large windows implies a longer computation time. With our dataset, we measure that doubling the duration or the spatial length of the window doubles the computation time. Looking for a compromise between model performance and computation time, we selected for the rest of our work the parameters: (t_{win}, d_{win}) = (8 s, 1000 m), and (t_{step}, d_{step}) = (4 s, 100 m). We note that, for these values, recall is very low compared to precision. This implies that our model does not detect a part of the points that belong to events but the alarm is reliable when a point corresponding to an event is detected. Figure B1.1 in Appendix B1 for several events indicates that some of the points associated with an event are undetected because of their low energy.

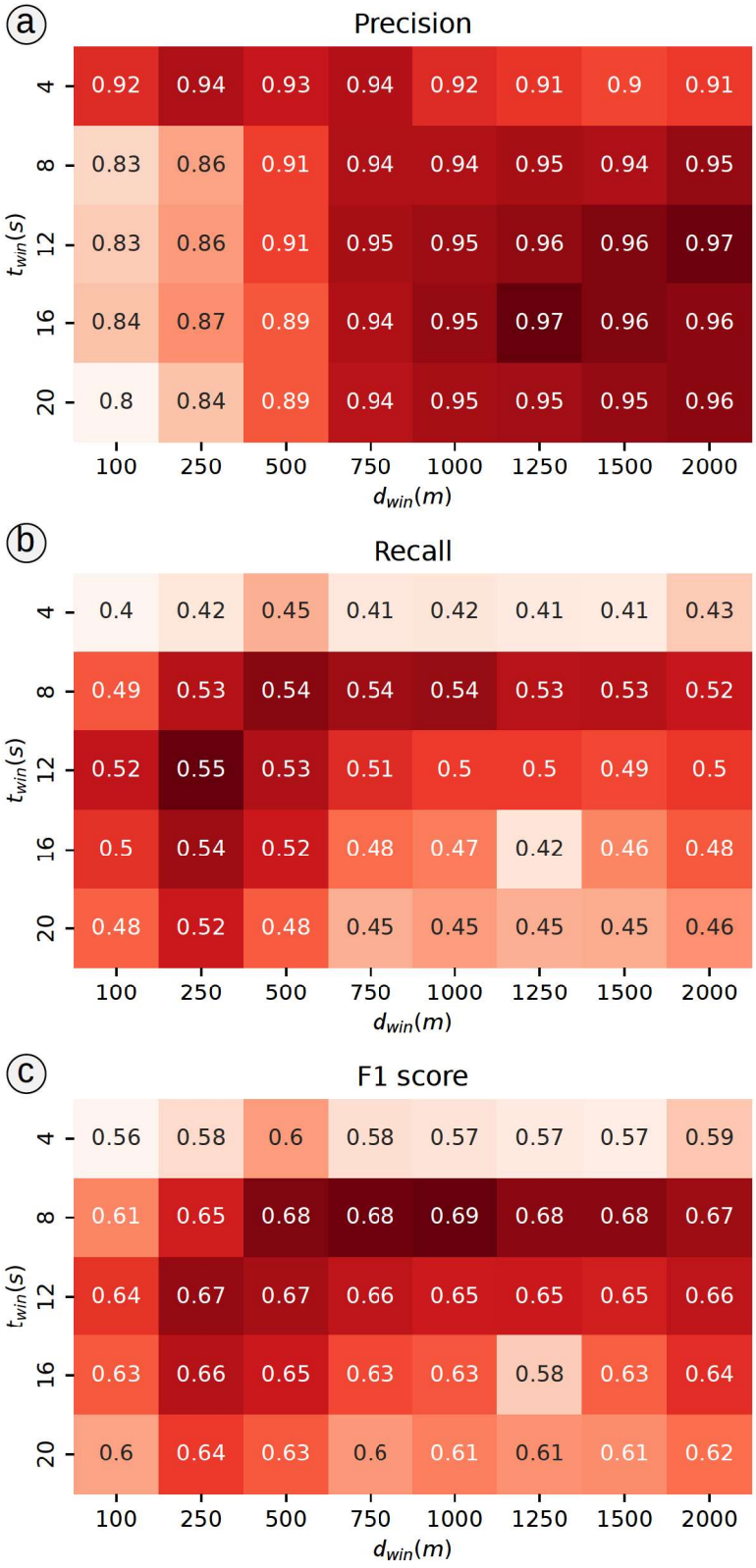


*FIGURE 4.7: Performance measurements for several time and spatial windows. Results are presented in terms of precision **(a)**, recall **(b)** and F1-score **(c)**.*

4.2.5.2 Impact of Post-Processing Parameters

The post-processing, based on the thresholding of the scores and on the application of the MRF algorithm, relies on the use of two parameters named $Score\_threshold$ and $Potential$. The $Score\_threshold$ is applied to the XGBoost score (section 4.2.4.4) and is chosen by visual inspection to reduce the number of false alarms. For our analysis, its value is set equal to 0.95. The $Potential$ value is chosen to maximize the F1-score on the training set after applying the MRF method on the training database. Figure 4.8 shows the calculated value of the normalized accuracy for several $Potential$ values; it indicates that 0.45 is the optimal value according to F1-score.

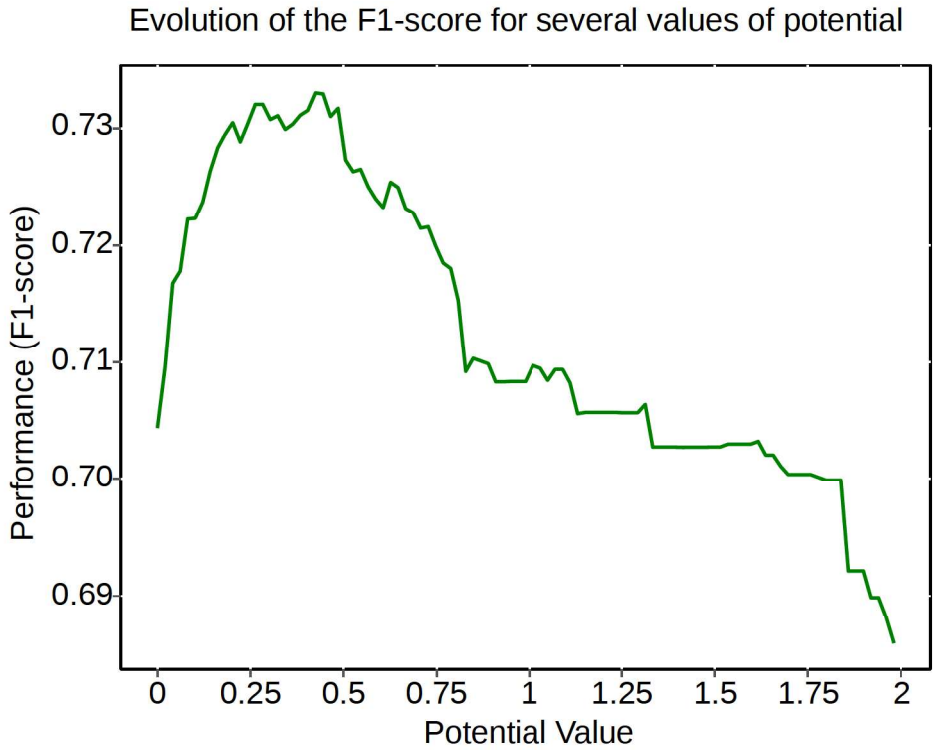


FIGURE 4.8: Performance estimation based on F1-score on the output of the Markov Random Field for several values of $Potential$. The optimal parameter which maximize this metric is $Potential = 0.45$.

4.2.5.3 Impacts of Spatial and Similarity Features

The influence of each feature on the model can be directly determined with the feature importance yielded by the XGBoost model (Figure 4.9). Because of evaluation with LOOCV, the final feature importance is averaged from each computed model. Features presented in section 4.2.4.3 are arbitrarily numbered from 1 to 111, and are named with their categories (temporal, spatial or similarity) followed by their number detailed in table B2.1 in Appendix B2. For example, (Spatial77) refers to the feature n°77, which belongs to the spatial features category. Figure 4.9 shows that the eight most important features are sorted in decreasing order of importance as follow: average of the raw spatial trace (Spatial77), energy filtered between $[0, \frac{1}{4}]$ of the Nyquist frequency band (Temporal40), number of peaks in auto-correlation function of spatial and temporal traces (Spatial84 and Temporal8), ratio between the energy in $[10, 30]$ Hz and $[30, 50]$ Hz (Temporal20), Kurtosis of the trace filtered in $[5, 10]$ Hz (Temporal23), indexes of maximum of cross-correlation function computed for two consecutive stacks of temporal traces (Similarity104) and maximum of cross-correlation function computed for two consecutive stacks of temporal traces (Similarity89).

Considering the eight most important features, two are spatial features (Spatial77, 84), two are similarity features (Similarity104, 89) and four are temporal features (Temporal40, 8, 20, 23). The contribution of spatial and similarity features can be qualitatively observed in the classification map (Figure 4.10) after using a score threshold of 0.95 on the output of XGBoost for each event. The choice is discussed in section 4.2.5.1. Taking the example of an earthquake that occurred on September 15, 2022 at 9:12:16 UTC (Figure 4.10a,b), we observe that relying solely on temporal features is insufficient for achieving accurate classification results, as anthropogenic events generate a lot of false alarms in Figure 4.10d. Spatial and similarity features have a lower sensitivity to anthropogenic events as depicted in Figure 4.10e,f. The combination of all the presented features results in a detection map with less false detections (Figure 4.10c). Each detection map is obtained using a grid defined as (t_{win}, d_{win}, t_{step}, d_{step}) = (8 s, 1000 m, 4 s, 100 m), as discussed in Section 4.2.5.1. Appendix B1 provides the detection maps for the 19 cataloged events recorded during the study period.

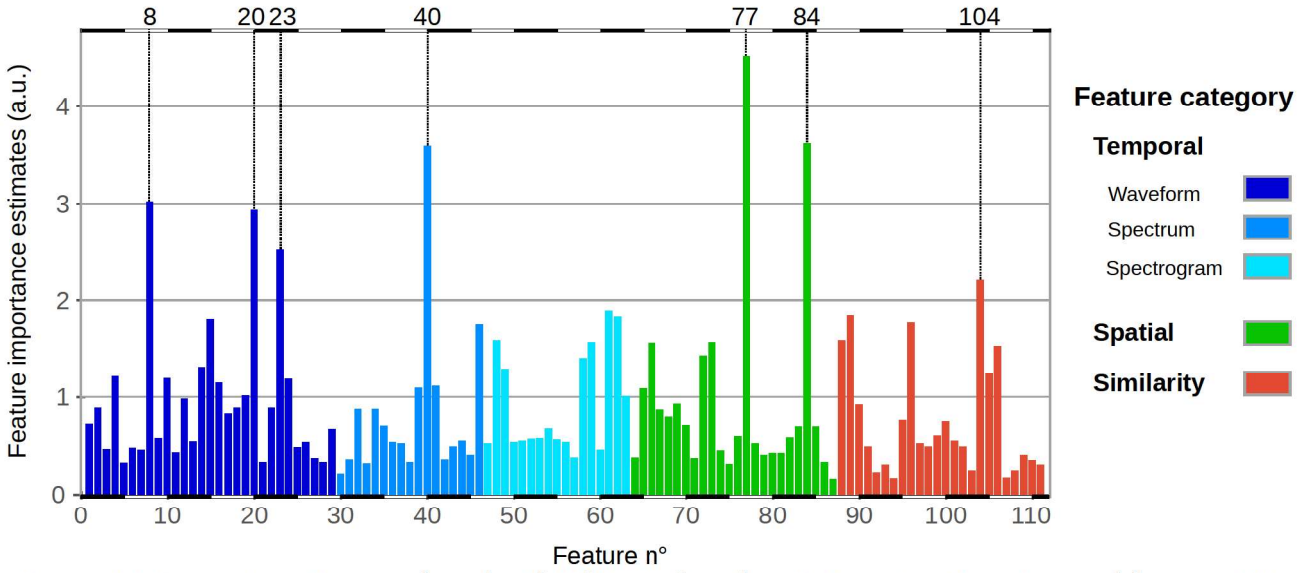


FIGURE 4.9: Feature importance evaluated with XGBoost algorithm. Features 1 to 63 are temporal features, 64 to 87 are spatial features, and 88 to 111 are similarity features.

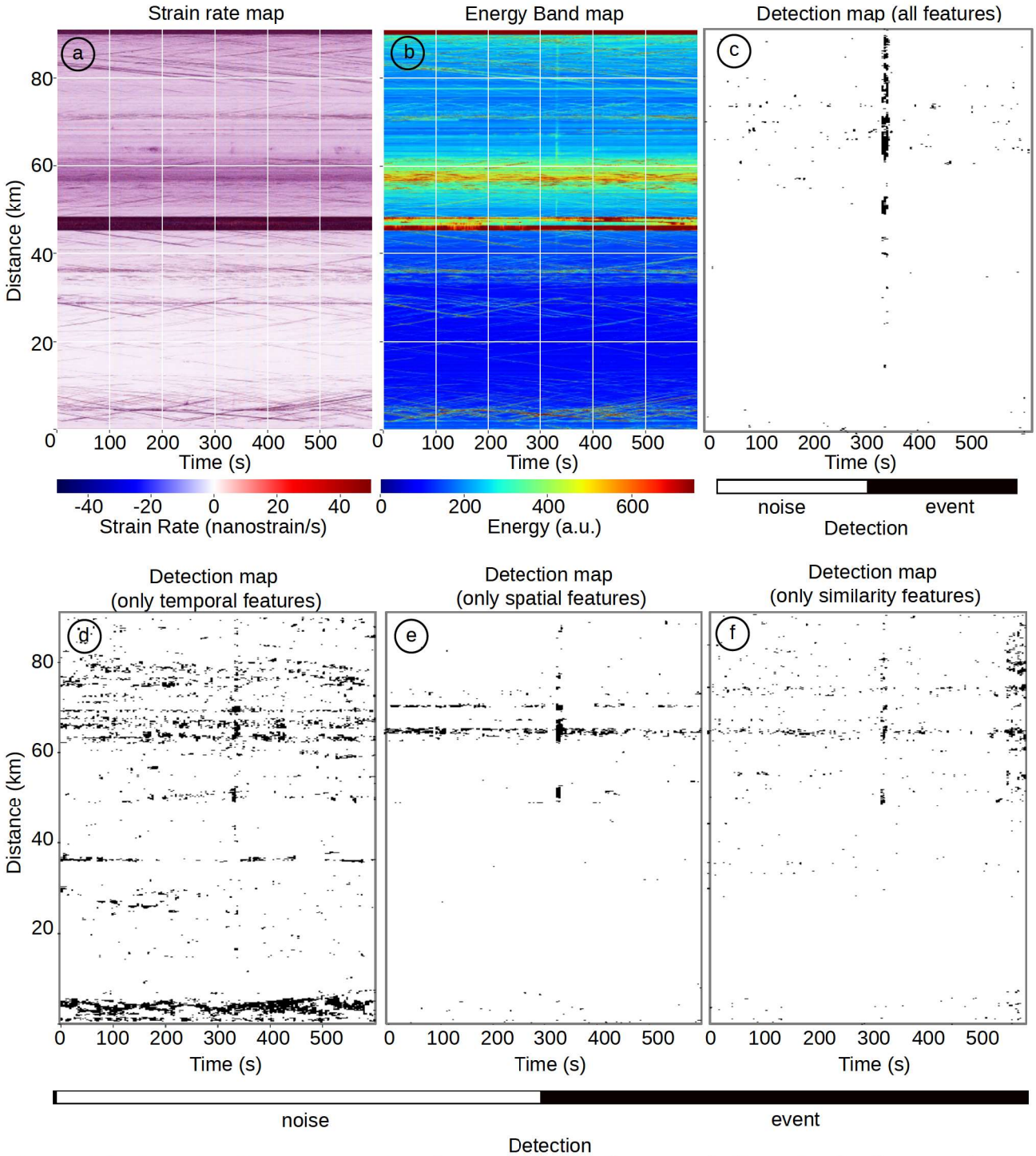


FIGURE 4.10: Detection Map representation of an earthquake of magnitude Mw = 1.1 that happens close to Argelès-Gazost, on Sept. 15, 2022 at 9:12:16 UTC. Detection map is obtained on the score map with a threshold of 0.95. ***(a)*** *represents the SR,* ***(b)*** *represents the EB,* ***(c)*** *the detection map obtained with a classifier that use all features,* ***(d)*** *the detection map obtained with a classifier that is only trained on temporal features,* ***(e)*** *the detection map obtained with a classifier that is only trained on spatial features,* ***(f)*** *the detection map obtained with a classifier that only uses similarity features.*

4.2.5.4 Performance of the Processing Chain

Using the different window parameters and features defined in the previous sections, we compute the features for all the identified events, train various models with several training sets using LOOCV technique and post-process the results with score thresholding and MRF method.

Figure 4.11 contains an example of the detection of an earthquake of magnitude Mw = 1.1 (Figure 4.11c) using the SR map (Figure 4.11a). We also plot the EB map (Figure 4.11b) to help the reader identify the event location. The detection and identification maps resulting from our workflow, along with the SR and the EB maps, are presented for each event recorded during the study period in Appendix B1.

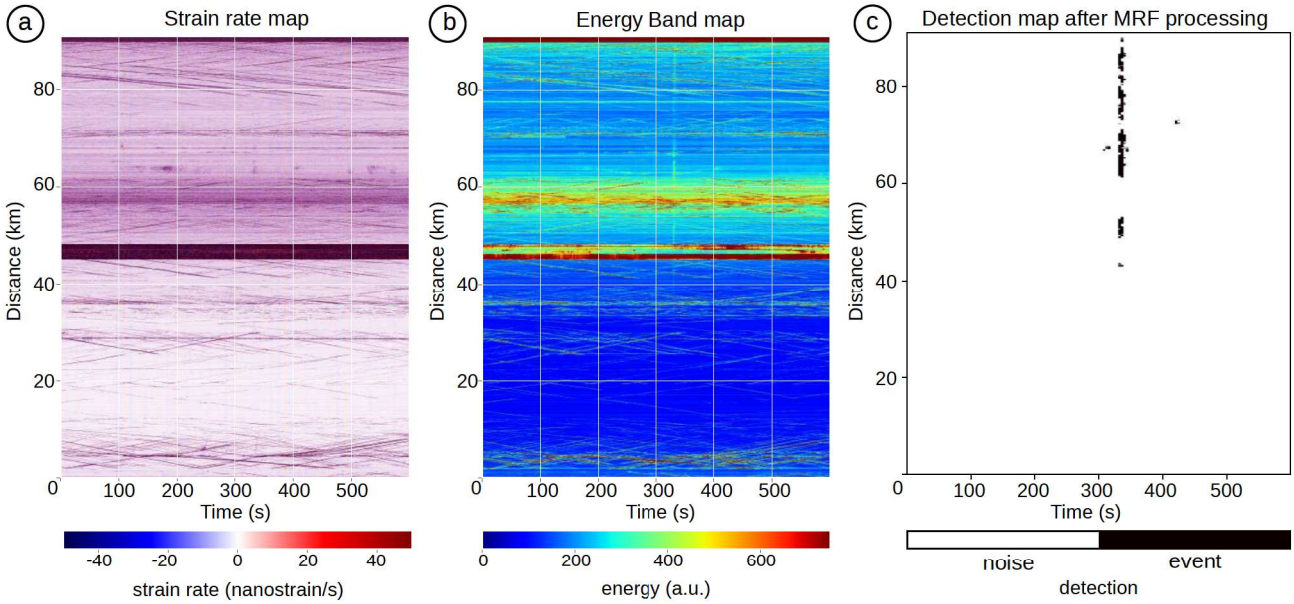


FIGURE 4.11: Detection Map representation of an earthquake of magnitude Mw = 1.1 that happens close to Argelès-Gazost, on Sept. 15, 2022 at 9:12:16 UTC. Detection map is obtained on the score map with application of MRF model, with a threshold of 0.95 and with use of MRF. ***(a)*** *represents the SR,* ***(b)*** *the EB,* ***(c)*** *and the detection map.*

The XGBoost algorithm combined with the proposed spatial and similarity features is able to correctly detect an earthquake of magnitude Mw = 0.4 at a maximum distance of 1.5 km (EQ8 in Table 4.1) and quarry blasts of magnitude Mw = 1.1 at a maximum distance of 1.0 km (QB4 in Table 4.1). Detection is achieved despite the surrounding anthropogenic noise, and quantified using the LOOCV splitting for which each event is processed by a model trained with the other events from our database. Amongst the 19 events visible on the DAS data, 13 of the 13 earthquakes are detected, as well as 3 of the 6 quarry blasts. One of the detected earthquakes is close to non-detection and corresponds to an event of magnitude Mw = 0.8 located at 1.0 km from the closest point along the fiber (EQ12 in Table 4.1). Undetected quarry blasts are located at distances higher than 4 km from the closest point sensor of the fiber (QB2, QB3 and QB6 in Table 4.1). Compared to the detected ones, they present lower energy and are difficult to detect above the noise.

4.2.6 Discussion and Conclusion

This paper introduces novel DAS-oriented features—spatial and similarity features—complementing the temporal features established in Huynh et al. (2022). These features enhance event detection and identification within a comprehensive processing chain. Our findings confirm that the integration

of these new features significantly improves classification outcomes, even under field conditions with a standard fiber optic telecommunication cable spanning nearly 100 km. The promising results yielded by our processing chain despite the presence of environmental and anthropogenic noises are showing the relevance of AI based processing chains for the detection of seismic events of interest with the DAS acquisition system for the monitoring of natural seismic sources at regional scales. We achieved effective earthquake detection, highlighting DAS capability to identify such events even in challenging conditions. Quarry blasts were accurately detected when located within 1.8 km of the fiber optic network, showcasing the method utility in near-field applications.

However, several limitations must be acknowledged. The relatively small dataset and the scarcity of natural seismic events constrained our analysis, impacting the interpretability of the results. The reliance on a supervised machine learning algorithm necessitated accurate data labeling, which was challenging given the potential mislabeling of low-magnitude events. Consequently, we opted to train our algorithm using only 10-min segments of DAS data associated with cataloged events, as opposed to utilizing the full three weeks of continuous data.

To address these limitations, future research could explore semi-supervised or self supervised machine learning approaches, which would leverage the extensive amount of data collected during the measurement campaign (Wang et al., 2022; Zhu et al., 2023; Rimpot, Hibert, Malet, et al., 2024). Additionally, extending the acquisition campaign or employing datasets from other locations could enhance the dataset size and diversity. Another promising avenue is to utilize the existing seismic knowledge from the six permanent seismometers in the region to inform DAS monitoring through transfer learning (Titos et al., 2020; Lapins et al., 2021; Donnadille et al., 2024).

The geometric configuration of the fiber optic network and its varying signal-to-noise ratio (SNR) based on position also influenced our results. The non-proportional increase in distance between event sources and fiber points affects the spatial and similarity features measured. Hence, incorporating the geographic distribution of events through localisation when possible is crucial for minimizing the influence of network geometry on model performance in different study areas.

The ability to detect very low magnitude earthquakes, starting from Mw=0.4 at a distance of 1.5 km from the fiber, and quarry blasts within 4 km, underscores the high sensitivity and efficacy of DAS for event monitoring. This high sensitivity, coupled with the distributed nature of the sensor network, establishes DAS as an excellent solution for long-term monitoring of natural processes over regional scales with unprecedented spatial resolution. Future research should address dataset limitations, explore semi- or self-supervised and transfer learning methods, and consider the implications of fiber geometry on detection performance.

4.3 Chapter Summary

In this chapter, we presented a functional approach for detecting earthquakes and quarry blasts using a real-scale fiber optic spanning 91 km in the Hautes-Pyrénées, the second most seismically active region in France after the Alps. This approach integrates human-engineered features that account for temporal information, spatial patterns, and trace-by-trace comparisons. Despite the continuous presence of anthropogenic noise along the fiber, the method demonstrated strong potential for seismic monitoring. Of the 13 earthquakes and 6 quarry blasts with magnitudes between Mw=0.4 and and Mw=2.4, 13 earthquakes and 3 quarry blasts were successfully detected. The results highlight the system ability to effectively detect seismogenic events, showcasing the capabilities of DAS even in challenging environments.

However, the study was limited to only 19 ten-minute files and did not process continuous data. These limitations are due to the short duration of the measurement campaign (three weeks) in a region with relatively low natural seismic activity. These limitations suggest that results could be more significant with longer monitoring periods or in more seismically active areas. In addition, the data labeling process relies only on an external pre-existing event catalog. For other applications, such as detecting nearby construction work, this approach would not be feasible. This underscores the need for a revised data labeling approach to avoid a time-consuming event-per-event labelling process. For example, we worked with FEBUS Optics engineers to test the detection of nearby construction work as part of a collaborative project (CITEPH). The goal was to monitor a buried pipeline, about 60 km long, using a fiber optic cable deployed near Pau in southwestern France. Construction activities, such as machinery driving, hitting, and digging near the pipeline, were conducted during the project. In addition to the events caused by natural and daily anthropogenic activity, we generated seismogenic events using various construction machines at known dates and locations. Due to the large variety of observed events, it was challenging to label the data not generated by the used machines. A machine learning model was trained only with data generated by the used machines, but many false alarms were observed.

The next chapter presents a methodology based on unsupervised learning to label data by groups of events rather than individually. This approach uses the Hautes-Pyrénées dataset from this chapter and a new continuous 44-day dataset collected with an 800 m fiber optic network. The first dataset will be used to test the clustering method on labeled data, while the second will assess the method in a more operational setting with continuous data over a longer period.

Chapter 5

Exploring DAS Data Using Human-Engineered and Self-Supervised Learning-Based Features

5.1 Introduction

In the previous chapters, we developed classification methods for DAS data using features derived from conventional seismology, followed by the creation of new, DAS-specific features. While these approaches proved effective, they relied on supervised learning, requiring precise data labeling. To address this challenge temporarily, we either used data generated from controlled events conducted in our experiments (Chapter 3) or cross-referenced DAS data with cataloged events from external web services like BCSF-RENASS (Chapter 4). However, these approaches are ultimately dependent on external annotations, which are not always available. Without them, manual data examination of each individual event becomes necessary, a process that is both time-consuming and prone to human bias.

This chapter explores an alternative approach using unsupervised methods to analyze DAS data, group similar data into clusters, and perform cluster-by-cluster labeling. We build on the latent space representations developed in Chapters 3 and 4, comparing them to a representation generated using a self-supervised learning approach called BYOL. The input data for BYOL consists of four common representations used in DAS data analysis, and introduced in Chapter 2: energy band, STA/LTA ratio, spectrogram, and power spectrum density. These representations are then transformed into image format for easier manipulation within BYOL. Clustering algorithms are applied to each latent space representation to organize the data into clusters, and each cluster is manually labeled. Finally, we evaluate the two representation methods by comparing the behavior of the classes extracted by each approach and assessing the feasibility of clustering for efficiently grouping similar data to facilitate rapid manual labeling.

This chapter consists in one submitted paper (section 5.2):

Huynh, C., Rimpot, J., Hibert, C., Turquet, A., Stangeland, T., Malet, J. P., and Lanticq, V. (2025). *Unsupervised Learning for the Comprehensive Exploration of Continuous-DAS Data.* Submitted to JGR.

5.2 Paper: Unsupervised Learning for the Comprehensive Exploration of Continuous-DAS Data (Journal of Geophysical Research, Submitted)

5.2.1 Abstract

Distributed Acoustic Sensing (DAS) offers new possibilities for seismological monitoring by using fiber optic cables as sensors, enabling cost-effective and a dense spatial detection of seismogenic activity. However, the vast amount of data produced by DAS systems presents a significant challenge in terms of data labeling and analysis. Traditional supervised machine learning approaches require extensive labeling, which is time-consuming and prone to user bias. To address this challenge, we propose a two-step processing chain. The first step aims to represent each data point in the dataset using several hundred features, which will form our data representation space, called the latent space. We compare two latent space constructions: one based on features constructed using signal processing metrics commonly used in seismology (human-engineered features), and the other based on self-supervised algorithms using common bidimensional representations of DAS data (image-BYOL). After obtaining the latent spaces, the second step focuses on reducing the entire dataset into a set of clusters using an unsupervised clustering method. We propose using two sequential clustering algorithms: the first reduces the dataset to 5000 clusters, and the second applies hierarchical clustering to group these 5000 clusters into 500 to 700 more interpretable clusters based on inconsistency criterion. This method is applied to continuous DAS data collected during two different recording experiments in the Hautes-Pyrénées mountains, and with two different configurations: a 6-week continuous measurement along a 800-m cable at Viella, and 19 different 10-min measurements along a 91-km cable. For Viella dataset, we achieved to detect 100% of the events with magnitude $M_w > 2.0$. The image-BYOL latent space generates a higher rate of false positives compared to human-engineered features. The results highlight the potential of clustering techniques for improving DAS data analysis, while also emphasizing the need for further refinement to reduce false positives, especially for smaller seismogenic events.

5.2.2 Introduction

Distributed Acoustic Sensing (DAS) is a promising tool for monitoring seismogenic activity. DAS systems use fiber optic cables as sensors to detect ground vibrations, offering a cost-effective and efficient alternative to traditional seismometers. Over recent years, DAS has seen widespread adoption across a variety of applications. It has been utilized for detecting natural seismogenic events, including earthquakes (Li et al., 2019; Lior et al., 2023), monitoring volcanic activity (Nishimura et al., 2021; Jousset et al., 2022), and monitoring water reservoirs (Zhu et al., 2021; Tribaldos & Ajo-Franklin, 2021). Additionally, it has proven valuable in marine geophysics (Sladen et al., 2019; Lindsey et al., 2019; Williams et al., 2019; Spica et al., 2022; Lin et al., 2024), and seismic imaging (Dou et al., 2017; Walter et al., 2020). Beyond natural seismogenic events, DAS is also highly effective for monitoring anthropogenic activities such as oil and gas pipelines (Tejedor et al., 2021b), train tracks, roads (Li & et al., 2020; Wiesmeyr et al., 2020; Yuan et al., 2021), traffic and avalanche monitoring (Kleine et al., 2024; Turquet et al., 2024a, 2024b), and even submarine power cables (Hicke et al., 2017a). The ability of DAS to capture seismic signals over long distances makes it especially advantageous for monitoring vast areas or regions with limited accessibility (Turquet et al., 2024a).

Because of its versatility in term of detectable seismogenic events, it is difficult to extract signals from specific sources without specific methods. These methods can either be signal processing-

based, such as slow-time average over long-time average (STA/LTA), temporal spectrum analysis, spatial spectrum analysis such as f-k filtering (Paitz et al., 2023), or kurtosis techniques (Trnkoczy, 2009; Vaezi & Van der Baan, 2015; Li et al., 2016; Kumar et al., 2018), or employ data-driven Machine Learning (ML) techniques. Among popular ML techniques, some integrate signal processing methods as temporal signal analysis (Hibert, Mangeney, et al., 2014; Hibert, Provost, et al., 2017; Hibert et al., 2019; Maggi et al., 2017; Provost et al., 2017; Chmiel et al., 2021; Domel et al., 2023), or derived parameters such as Fast Fourier Transform (Wiesmeyr et al., 2020; Tejedor et al., 2021b), wavelet decomposition (Wang et al., 2019), or Mel-frequency cepstral coefficients (Bublin, 2021). For the DAS application, Huynh et al. (2024) has implemented a set of 111 features that integrate temporal, spatial and trace similarity information. The goal of these algorithms is to identify relevant features and weight them according to the task at hand. Another approach involves deep neural networks, which can construct discriminative features for classification and detection of specific events, such as microseismogenic activity (Binder & Tura, 2020), earthquakes (Zhu & Beroza, 2019; Mousavi et al., 2020), or even footsteps (Jakkampudi et al., 2020). The weighting of the neurons in the network, like in the previous case, must be adjusted according to the task objective. This requires providing supervised models with labeled data from the target classes and non-target classes. However, this labeling process is time-consuming, resource-intensive as it demands a comprehensive review of the catalog used to train the model, and introduces selection bias in the catalogues (Yoon et al., 2015).

To address these chllenges, transfer learning techniques have been proposed to reduce the need for labeled data. Transfer learning is a machine learning technique that allows a model trained on one task to be adapted to another task with minimal additional training. For the DAS, it includes training dataset from a different location or a different time period, or even a different type of sensor like conventional seismometers. Donnadille et al. (2024) applies knowledge transfer from a model trained on conventional seismometers to a DAS dataset in three scenarios: classifying local vs. distant earthquakes, distinguishing quarry blasts from earthquakes, and jointly identifying all three event types. The model achieves an F1 score exceeding 76% across these tasks. Unsupervised learning techniques could also offer a promising alternative for full dataset labelling process. Initially used for data exploration to identify emerging classes of events, in particular in volcano-monitoring system (Esposito et al., 2008; Hammer et al., 2012; Unglert & Jellinek, 2017; Rimpot, Hibert, Retailleau, et al., 2024), induced seismicity (Beyreuther et al., 2012), earthquake monitoring (Yoon et al., 2015; Cesca, 2020), and landslides (Seydoux et al., 2020), unsupervised algorithms can reduce the time and effort required for manual labeling in a more general seismic context (Johnson et al., 2020). Rimpot, Hibert, Retailleau, et al. (2024) introduces a workflow that combines a latent space derived from Self-Supervised Learning (SSL) with clustering, enabling a detailed and comprehensive analysis of seismological data despite the prevalence of background noise. Originally designed for dense seismic node networks, this workflow can be adapted for DAS data. Additionally, the approach is flexible, allowing the integration of various latent feature spaces to enhance the detection of seismogenic events.

Building on the study by Rimpot, Hibert, Retailleau, et al. (2024), we propose a method for analyzing and labeling DAS data using an unsupervised machine learning algorithm. We compare two different latent feature spaces: one derived from human-engineered features and another from the SSL approach. Next, we apply a two-step clustering process. The first step reduces the dataset into 5,000 clusters, minimizing information loss, while the second step utilizes a tree-based hierarchical clustering algorithm to organize and merge these clusters into a more interpretable set of 500 to 700 clusters. These clusters are then manually labeled into five to six classes, one of which is earthquake. Its behavior compared to a reference database obtained from the “Bureau Central et Sismologique Français” and ”REseau NAtional de Surveillance Sismique” (BCSF-RENASS) and Observatoire Midi-Pyrenees (OMP) catalogs. We selected two datasets from the Hautes-Pyrénées mountains in France, each collected over distinct time periods and spatial scales, to test our processing chain. By using

one local dataset and one regional dataset, we aim to evaluate the scalability and efficiency of our approach across different spatial and temporal contexts. The scalability of this method is limited by noise variations along the fiber, especially in fibers over several tens of kilometers. The article is structured to first introduce the datasets, followed by a detailed explanation of the data processing methodology. The results are then presented and discussed for their implications, with a final section summarizing our findings and suggesting potential directions for future research.

5.2.3 Dataset

In this work, we propose to test our approach on two datasets: the "Pyrenees" dataset, which recorded during the period of August-September 2022 (Huynh et al., 2024), and the "Viella" dataset, which recorded during the period of December 2023-January 2024. Both datasets were collected in the Pyrenees mountains in France, but they differ in the length of the fiber optic cable and the scale of the study. The Pyrenees dataset focuses on data acquired at a regional scale and uses a 91 km long fiber optic cable (Huynh et al., 2024), while the Viella dataset focuses on data acquired at a local scale and uses a 800 m long cable. For both datasets, the strain rate (SR) representation, expressed in nanostrain per second (nstrain/s), is derived from the optical raw data measured with the DAS by setting the gauge length and the derivation time. The derivation time affects the maximum observable frequency, while the gauge length affects the spatial resolution and then the minimum detectable wavelengths. In this section, we separately describe the characteristics of each dataset.

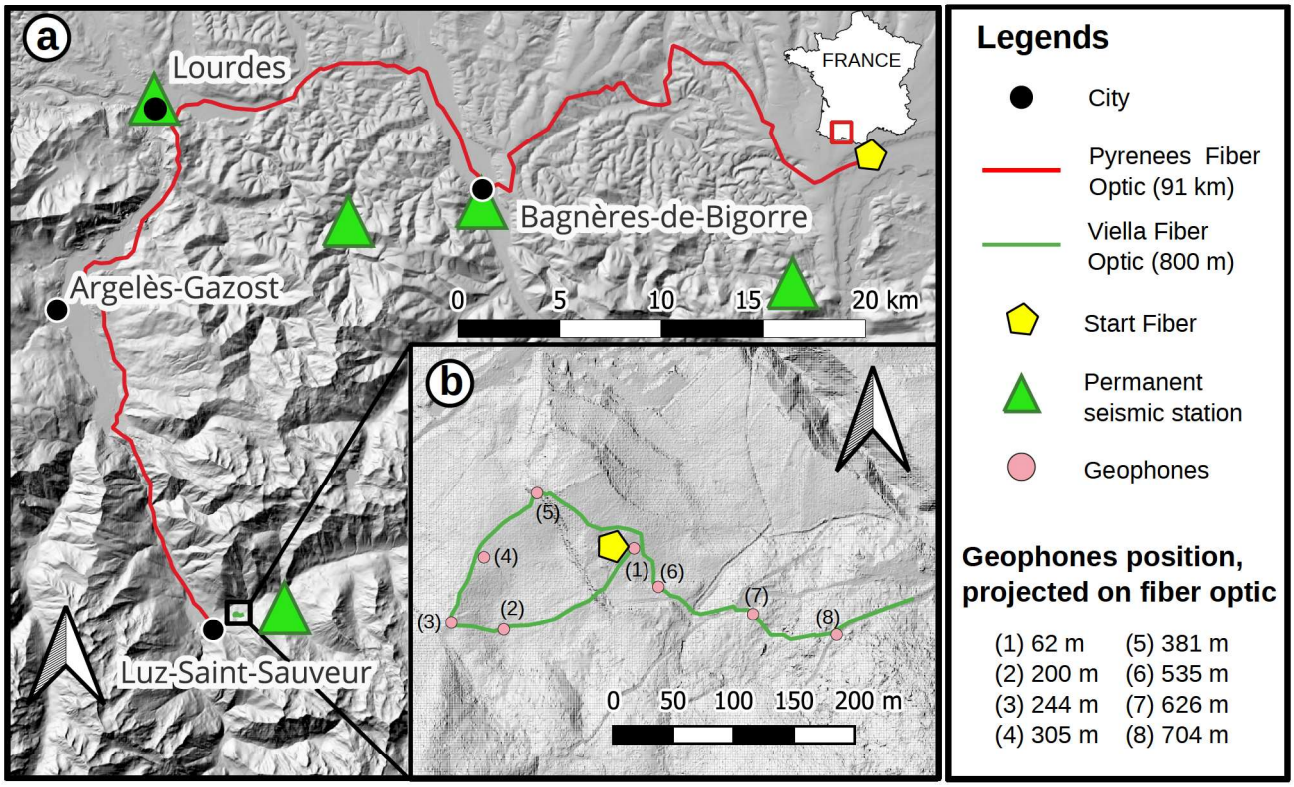


*FIGURE 5.1: Fiber path of Pyrenees **(a)** and Viella **(b)** field test. For Viella field test, 8 nodes are also installed along the fiber optic and indicated with light red circles.*

5.2.3.1 Pyrenees Dataset

The Pyrenees site is of particular interest due to the moderate seismicity of the Pyrenees mountains (Souriau & Pauchet, 1998; Rigo et al., 2005; Lacan & Ortuño, 2012; Sylvander et al., 2022) and various cataloged anthropogenic activities, such as quarry blasts. The presence of a 91-km pre-existing fiber optics telecom cable on a regional scale allows for the acquisition of data across a wide variety of seismic environments originated from departmental roads, urban areas, and natural surroundings. Data were collected using a FEBUS A1-R DAS system between August 30 and September 20, 2022, with seismic data extracted using a gauge length of 10 m and a sampling frequency of 200 Hz. The spatial sampling interval is 4.8 m, resulting in a total of 18 958 virtual sensors. To ensure a sufficiently high signal-to-noise ratio over long distances, a FEBUS range extension module is connected at the midpoint of the fiber optic. In order to build a catalog of referenced seismogenic events, we use the open-access catalog of the French facilities "Bureau Central et Sismologique Français" and "REseau NAtional de Surveillance Sismique" (BCSF-RENASS). The open-access catalog provides a list of natural earthquakes and quarry blasts that occurred during the study period. By cross-referencing the BCSF-RENASS catalog and our DAS data, we build a dataset consisting of 19 recordings, each 10 min long, containing a total of 13 natural earthquakes and 6 quarry blasts, with magnitudes ranging from $M_w = 0.4$ to $M_w = 2.4$. These 19 recordings represent a total of 206 GB of data, and are deeply presented in Huynh et al. (2024). Figure 5.1a provides the path of the deployed fiber optic for the Pyrenees field.

5.2.3.2 Viella Dataset

The small commune of Viella, near Luz-Saint-Sauveur in the Hautes-Pyrenees region of France, is situated close to the endpoint of the fiber optic cable used in the Pyrenees study for seismological monitoring. However, this experiment was conducted independently of the Pyrenees study. The region is vulnerable to a large variety of natural hazards. This includes meteorological events like storms, snowfalls, and droughts, as well as more destructive phenomena such as floods, landslides, avalanches, and earthquakes. Viella, classified as a region of moderate seismicity, faces added complexity in its seismological landscape due to its susceptibility to environmental seismogenic events, including avalanches and landslides. Viella is a village that has been affected by gravity hazards for centuries, with major landslides occurring in 1994, 1999, 2013, and 2018. These events have caused considerable damage, particularly to buildings, which have developed large cracks and fissures over time, while also leading to noticeable changes in the surrounding landscape. The 2018 landslide in particular highlighted the geological instability of the area and has continued to affect the region through frequent rockfalls and smaller landslides. These ground movements make Viella a key location for studying the interactions between natural seismicity and environmental triggers, providing valuable data for improving our understanding of regional seismic risks.

The Viella site involved the manual deployment of an 800 m fiber optic cable. To maximize sensitivity, the fiber was aligned perpendicular to the sliding slope. Continuous coupling with the ground was essential for accurate data measure. The cable was manually buried at a depth of 2–5 cm, with the soil compacted by hand on top. Garden staples is also used to ensure coupling in areas with harder media. Additionally, the fiber should be robust enough to withstand mechanical pressures, such as shifting earth, debris or even mole, without compromising its sensitivity to seismic waves. Despite these cautions, the fiber optic cable experienced several cuts at various points during our study period : on 2023-12-11 (fiber cut at 312.4 m), 2023-12-21 (fiber cut at 266.4 m), 2024-01-07 (fiber cut at 246 m) and 2024-01-16 (fiber cut at 214 m). The recording is performed using a FEBUS A1-R DAS system installed into a farm, to ensure enough power supply and protect the optoelectronic equipment from weather conditions. It has been carried out continuously over six weeks between

December 11, 2023 and January 24, 2024. Important measurement settings include a gauge length of 10 m and a sampling frequency of 400 Hz, as well as a spatial sampling of 2.4 m. Because of the fiber cuts, the amount of available virtual sensors is not constant over time: at the beginning of the study, the dataset contains 312 virtual sensors, while at the end, only 68 virtual sensors are available. The required data storage volume for Viella dataset is about 3 TB. Figure 5.1b provides the path of the deployed fiber optic at Viella.

In addition to the fiber optic cable, we also deployed a network of 8 portable nodes at the Viella site. These nodes are equipped with accelerometers that are able to capture the seismic variation of environmental events in the three-dimension space. The nodes are installed along the fiber optic to perform comparative study with the seismic data captured by the fiber optic. Data are measured with a sampling frequency of 500 Hz. The nodes are installed by hand at a depth of tens of centimeters in the ground, to ensure a good coupling with the soil. The nodes are powered by a battery and can record data for a duration of three to four weeks. Since the nodes were installed on the same day as the DAS setup, nodes measurement finished between January 1 and January 5, 2024. The nodes generate a total amount of 58 GB of data. Figure 5.1b provides the position of the deployed nodes at Viella.

The context of the installation site plays a crucial role in the nature of the collected data. As the fiber has been deployed close to a working farm, a portion of the data captured during the study is influenced by anthropogenic events. Routine farm activities, such as tractor movements and the movement of breeding animals, is frequently observed in the dataset. In addition to human activities, as the fiber optic cable is shallowly buried, the DAS system is also highly sensitive to environmental variations. For instance, precipitation events are detected in the data. The system also recorded several seismogenic events.

5.2.4 Methods

Our study employs a four-step methodology to analyze seismic data: 1) pre-processing; 2) latent space computation; 3) K-Means clustering; and 4) hierarchical clustering. During pre-processing, we remove the instrumental response from the seismic data (SR) and segment it into discrete windows for efficient streaming, referred to as SR data blocks. The latent space computation transforms these blocks into a feature-based representation, enabling clustering. We compare two approaches for constructing this latent space: one based on 111 hand-engineered features capturing seismogenic event characteristics across temporal, spatial, and trace-by-trace domains (Huynh et al., 2024), and another using self-supervised learning with image-based DAS representations (e.g., Energy Band, STA/LTA, spectrograms, Power Spectral Density). Using the BYOL algorithm (Grill et al., 2020), the latter learns a latent space tailored for clustering. K-Means clustering condenses data into 5000 diverse clusters, which hierarchical clustering further refines into 500–700 interpretable clusters via a dendrogram-based proximity metric like inconsistency.

5.2.4.1 Data Pre-Processing

As DAS use fiber optic to measure strain and deformation along the fiber, it is particularly sensitive to common-mode instrumental noise, which occurs when sound and vibrations in the vicinity of the DAS interrogator simultaneously affect all measuring position along the fiber, called data channels. The suppression of common-mode frequency components is computed for each channel by calculating the spatial mean over the entire length of the fiber optic at each time, and subsequently subtracted it from the input signal (Hartog, 2017).

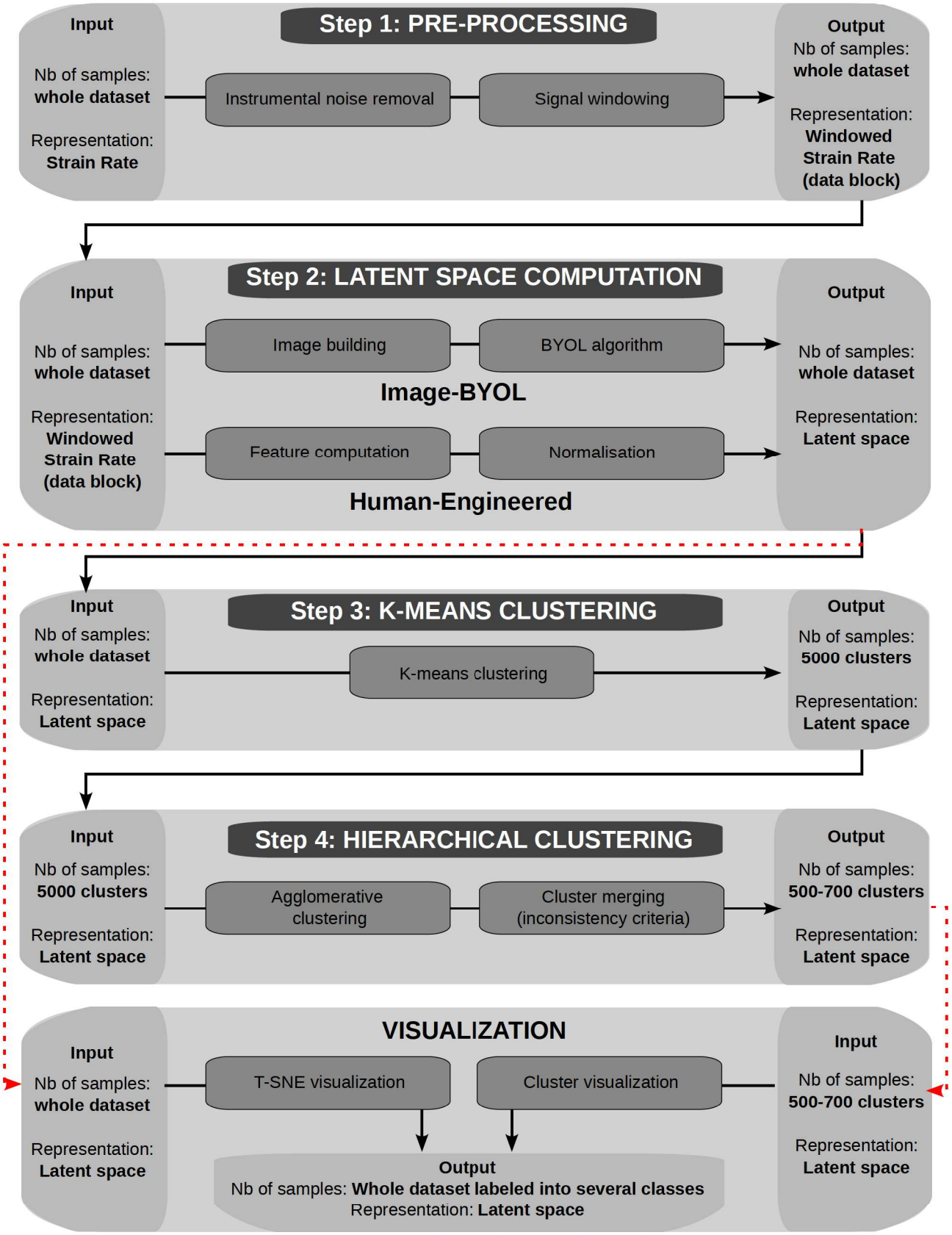


FIGURE 5.2: Overview of the processing chain.

	Pyrenees	Viella
Fiber Length (km)	91	0.8 (begin) - 0.214 (end)
Duration	19x 10-min	44 days
Sampling Frequency (Hz)	200	400
Spatial Sampling (m)	4.8	2.4
Gauge Length (m)	10	10
Data Volume (GB)	206	3000
Number of windows	17 480	142 727
Time window size (s)	60	60
Distance window size (m)	1000	100
Window containing EQ/QB	243	unknown

TABLE 5.1: Summary table of the two datasets. The upper part of the table provides acquisition configuration, while the lower part provides the number of windows and the size of the windows.

We then slice the SR data into discrete windows for the processing chain, named SR data blocks. This slicing is defined by two user-specified parameters: the time window size t_{win} and the distance window size d_{win}. Each window represents a SR data block that is defined in both the temporal and spatial dimensions, without any overlap between consecutive windows. For the Pyrenees dataset, we use a window size of 1000 m in distance and 1 min in time, whereas for the Viella dataset, a window size of 100 m and 1 min was applied. Given the different lengths of the fiber, this approach yields 90 windows per min for Pyrenees and between 1 and 7 windows per min for Viella. A total of 17 480 SR data blocks have been collected in the Pyrenees dataset, including 243 SR data blocks of earthquakes and 52 SR data blocks of quarry blasts. For Viella dataset, 142 727 SR data block have been collected. Table 5.1 gives an overview of the two data sets in terms of acquisition configuration and SR data block size and amount.

5.2.4.2 Image-BYOL Latent Space

Data Representations

Computing a latent space with the image-BYOL approach requires the preparation of images from the DAS data. Ideally, these images should contain elements that help differentiate between various types of potentially seismogenic sources. Such elements can include contours, shapes, sizes, textures, pixel intensity, or symmetries within the image. These images should contain diverse representations of the seismogenic source data able to capture their temporal and spatial dynamics. Leveraging our observation and knowledge in DAS data visualization, we selected four distinct representations: Energy Band (EB), Short-Term Average over Long-Term Average (STA/LTA), spectrogram, and Power Spectral Density (PSD). Except for the STA/LTA ratio directly computed using the SR representation, EB and PSD are obtained by summing values calculated from the spectrogram representation, which ensures a low computational cost.

The first two representations, EB and STA/LTA, are primarily designed to detect events within the data. EB is a widely used representation for visually inspecting DAS data and is an effective tool for triggering alarms in industrial applications, such as basic intrusion detection (Johannessen et al., 2012; Parker et al., 2014). It captures the movement of seismogenic sources, including their direction and speed, through its representation in both space and time. Our EB representation is calculated by integrating the spectrum of a windowed DAS channel over the 0.2-100 Hz frequency

range. The window duration is 0.5 s with an overlap of 0.8, resulting in a representation with a temporal resolution of 0.1 s. STA/LTA is a well-established technique in seismology for detecting seismic wave arrivals from natural seismogenic events, such as earthquakes. STA/LTA is computed by taking the ratio of the short-term average to the long-term average of the signal. For both datasets, we use a window size of 1 s for the short-term average and 10 s for the long-term average. The STA/LTA is calculated every 4 ms.

The other two representations, spectrogram and PSD, focus more on the spectral analysis of the data. A spectrogram is a visual representation of the frequency content of a signal over time, typically displayed with time on the x-axis, frequency on the y-axis, and color intensity representing the spectrum amplitude, expressed in dB ($20\log_{10}$ function), at each frequency and time point. We compute the spectrogram between 0.2-30 Hz instead of 0.2-100 Hz, as this frequency range contains the majority of seismic information, using a sliding window of 5 s with a 0.98 overlap. The resulting spectrogram has a frequency resolution of 0.2 Hz and a temporal resolution of 0.1 s. The PSD complements the spectrogram by providing a frequency content across different channels of the fiber optic cable. For the PSD representations, we compute the image representation between 0.2-30 Hz. The PSD image representation resolution is 0.2 Hz for frequency domain, 4.8 m for the Pyrenees dataset and 2.4 m for the Viella dataset for spatial domain, with the amplitude expressed in dB using the $20\log_{10}$ function.

Image Parametrization

During the image preparation process, it is essential to use a colormap that ensures weaker events remain visible, as stronger events could otherwise dominate the visualization. The objective is to preserve crucial seismic details, especially when dealing with events of different magnitudes. For this purpose, we use the jet colormap with white saturation at the upper limit. This means that any value exceeding a specified threshold is represented in white. This colormap is designated for representations of EB, spectrograms, and PSD, where intensity is directly linked to the seismic power of the source and where peak values may exist. By selecting an appropriate upper limit, high-value anomalies can be accounted for without distorting the overall representation. In the case of STA/LTA, however, the interpretation of the signal requires a different approach. A strong incoming signal produces a high STA/LTA value (greater than 1), while the value decreases between 0 and 1 as the signal fades, with stronger signals having values closer to 0. To better visualize the subtle transitions in STA/LTA values, and to have high value for strong signal incoming and fades, we apply a transformation using the absolute logarithm function.

In conjunction with the choice of colormap, selecting appropriate upper and lower bounds for intensity thresholds ensures that both weak and strong events are accurately represented without overshadowing important details. For the data, the upper bound is determined based on the 99th percentile of intensity values, allowing most signal variations to be captured while filtering out extreme outliers that could distort the visualization. Additionally, the upper bound is constrained by a minimum threshold, ensuring that background noise stays at the lower end of the color scale, even when no events are present. This minimum threshold is visually set for each representation.

Figure 5.3 presents four examples of the final image used as an input into the SSL algorithm after applying these bounds and colormaps, with annotations indicating the correspondence of the x and y axes and the content of each thumbnail. Figure 5.3a,c shows what the images look like if the measured signal is weak, 5.3b,d if the signal is strong. Generated images are saved in PNG format with RGB color encoding and a resolution of 256x256 pixels. This resolution is sufficient to

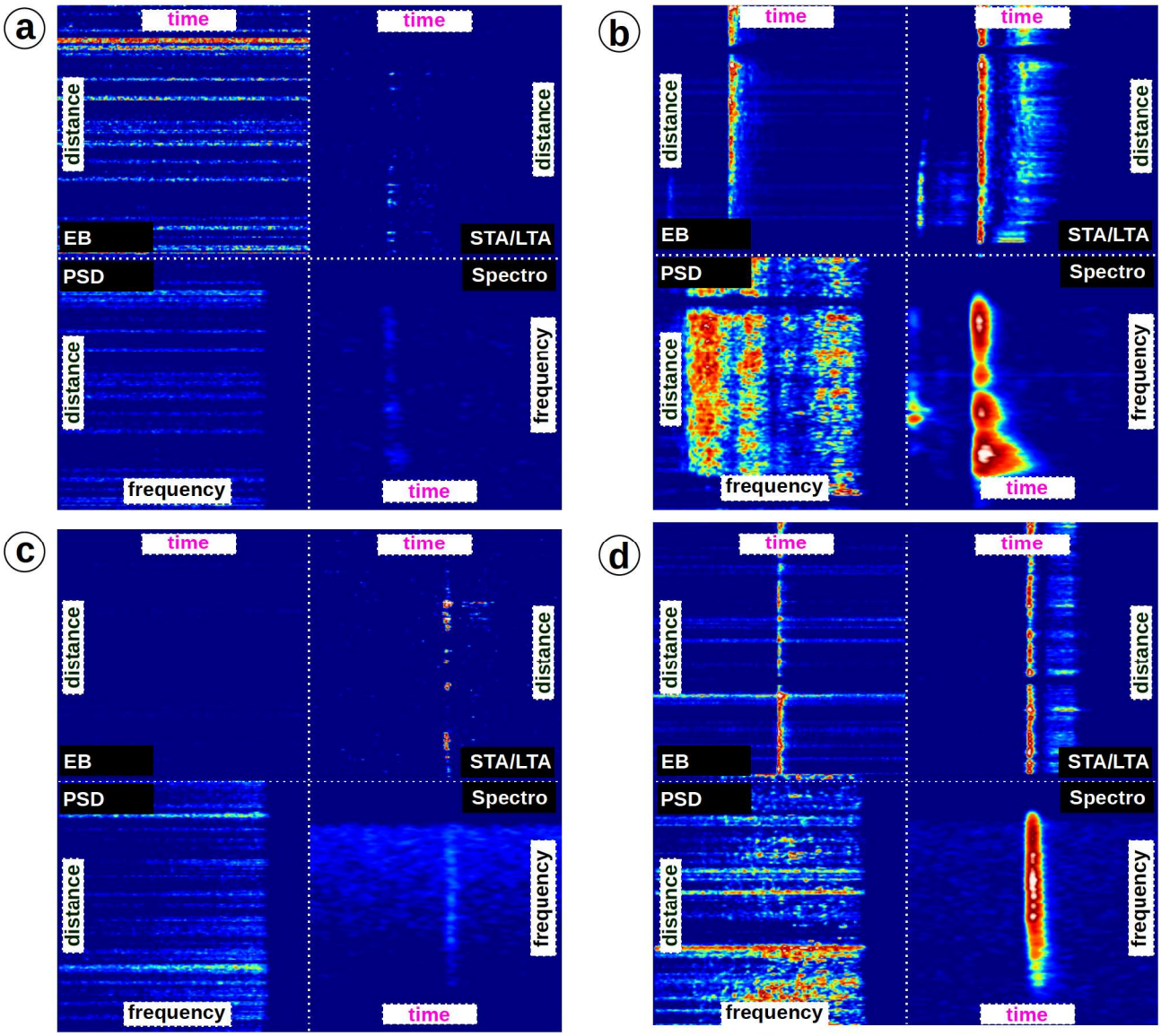


FIGURE 5.3: Image representation of the same event, measured at two different points on the fiber optic. ***(a)*** *and* ***(b)*** *represents a quarry blast measured at two different points distant of 6 km.* ***(c)*** *and* ***(d)*** *represent an earthquake measured at two different points distant of 31 km. Taken from Pyrenees dataset.*

capture the essential details of the seismogenic events while keeping the file size manageable for further processing.

BYOL Algorithm

BYOL (Bootstrap Your Own Latent) is a self-supervised learning algorithm designed to learn useful representations from unlabeled data (Grill et al., 2020). The algorithm is based on a contrastive learning approach, where the model is trained to produce similar features for augmented versions of the same input and different features for augmented versions of different inputs. This methodology encourages the model to learn a meaningful latent space that capture the underlying structure of the data without requiring explicit labels. Compared to other self-supervised learning algorithms, BYOL has several advantages. It is simple to implement, as it does not require negative samples or complex training objectives. The architecture of the algorithm is also highly flexible, as it can be applied to a wide range of data types like images (Grill et al., 2020), videos (Recasens et al., 2021;

Wang et al., 2022), and even audio (Niizumi et al., 2021) data. In our work, we use the BYOL model to learn a latent space based on 512 features extracted from the images.

BYOL is however highly dependent on the quality of the data augmentation process. The algorithm requires a diverse set of augmented versions of the input data to effectively learn a robust and generalizable latent space. In our study, we consider several classical image transformations, including flipping, rotation, color jitter, and random cropping. Techniques like flipping, rotation, and random cropping directly interact with the content of the images, making the model less sensitive to variations in the position and orientation of shapes that caracterize specific types of event. Meanwhile, color jitter introduces noise into the input data by modifying elements such as hue, saturation, contrast, and brightness. This technique helps make the features less sensitive to effects caused by the quality of the instrumentation or the influence of the event location relative to the fiber optic or the interrogator.

5.2.4.3 Human-Engineered Latent Space

Feature Engineering

The human-engineered latent space is designed to capture key characteristics of the DAS seismic signal. The latent space is built using 111 features presented in Huynh et al. (2024) and directly computed using the pre-processed SR data. The features that constitute the latent space are subdivided into three families: 63 temporal features, 24 spatial features, and 24 similarity features.

The temporal features, based on prior research on seismic signal classification using both traditional seismometer recordings (Provost et al., 2017; Maggi et al., 2017; Hibert, Provost, et al., 2017; Hibert et al., 2019; Malfante et al., 2018; Wenner et al., 2021; Chmiel et al., 2021; Falcin et al., 2021; Domel et al., 2023) and DAS recordings (Huynh et al., 2022), are divided into three categories: waveform, spectral, and spectrogram features. For example, waveform features involve envelope, autocorrelation or statistical measurement like kurtosis; spectral features include median FFT or energy; and spectrogram features include measuring the variation of the FFT median over time. The spatial features are derived from the shape of the seismic signal measured across all channels at a given time, including metrics like average, standard deviation, and auto-correlation. Similarity features are calculated by comparing traces from different fiber positions using cross-correlation and dynamic time warping.

Feature Normalization

Normalization ensures all features in the latent space are on a comparable scale, preventing any single feature from disproportionately affecting the clustering model. This is particularly important for human-engineered latent space, where features with different value ranges, such as energy values (100–1000) versus energy ratios (0–5), can skew the clustering process if not properly normalized. We choose to apply Z-score normalization. It involves transforming each single feature to have a mean of zero and a standard deviation of one. The Z-score normalization formula is given in Equation 5.1.

$$Standardized\_Value = \frac{Value - Mean}{Standard\ Deviation} \tag{5.1}$$

5.2.4.4 K-Means Clustering and Centroid Computation

K-Means clustering is an algorithm used for partitioning a dataset into a specified number of clusters. The algorithm works iteratively, beginning with random initialization of cluster centroids. Each data point is then assigned to the nearest centroid, based on a distance metric as Euclidean distance. Once the points are assigned, the centroids are recalculated as the mean position of all points within the cluster. This process repeats until convergence, either when the centroids stabilize or the assignments stop changing. K-Means is well-suited for large datasets and is highly scalable, making it a popular choice for clustering applications across various fields.

K-Means clustering is employed to partition the Pyrenees and Viella datasets, each represented in two distinct latent spaces. Variations in the latent spaces result in corresponding differences in the clustering outcomes. The objective is to aggregate data blocks with similar feature vectors, thereby reducing the dataset size while preserving essential characteristics of the original data. The Pyrenees dataset consists of 17,480 data blocks, whereas the Viella dataset comprises 142,727 data blocks, both clustered into 5,000 groups. This high number of clusters is intentionally selected to capture subtle patterns and relationships that might be overlooked with coarser clustering resolutions. Each cluster is represented by its centroid, expressed in latent space coordinates. In Section 5.2.4.5, these centroids are referred to as data points, with the term "cluster" designating the groups formed by the K-Means algorithm.

5.2.4.5 Hierarchical Clustering: An Agglomerative Approach

Hierarchical clustering provides a detailed view of how data points group together at different levels of granularity. It can be performed in two ways: divisive, starting with all points in one cluster and splitting them, or agglomerative, starting with each point as its own cluster and merging them. The agglomerative approach is faster and was chosen for this study. Initially, each cluster is a singleton containing a single data point. The algorithm then identifies the pair of clusters that are closest to each other and merges them into a new cluster. This closeness is quantified using a measure known as distance. This process continues iteratively, with clusters being merged based on their proximity, until all data points are grouped into a single cluster. This agglomerative approach results in a tree-like structure with nodes and leafs known as a dendrogram. The dendrogram, a representation of which is provided in Figures 5.5 and 5.6, visually illustrates the relationships and distances between clusters at different levels of the hierarchy. By analyzing the dendrogram, the number of clusters can be chosen based on various criteria, such as distance thresholds metric or cluster inconsistency.

In our study, we use agglomerative clustering to reduce the number of clusters generated in the K-Means step, aiming to achieve approximately 500 to 700 clusters. We tune the inconsistency parameter, as defined in Equation 5.2, until we achieve a number of clusters that is both reasonable and not too small, allowing rare events to appear in separate clusters. The inconsistency parameter compares the distance of the cluster parent node d_{parent} with the distance among its children $d_{children}$. A cluster is considered highly inconsistent if the height of its parent node is significantly larger than the average height of its children. This helps identify rare events and enables the formation of clusters with a small number of data points when those points are significantly different from the rest of the dataset. Since the inconsistency threshold and the resulting number of clusters are related, we perform several tests with different inconsistency values to obtain between 500 and 700 clusters. For Pyrenees and viella dataset, the inconsistency threshold value is set to 1.152. For the Pyrenees dataset, this approach produced 582 clusters with human-engineered latent space and 570 clusters with image-BYOL. Similarly, for the Viella dataset, we obtained 612 clusters with human-engineered latent space and 640 clusters with image-BYOL. The content and characteristics of these

clusters are further analyzed in Section 5.2.5.

$$Inconsistency = \frac{d_{parent} \quad mean(d_{children})}{std(d_{children})} \tag{5.2}$$

5.2.5 Results

In this section, we analyze the outcomes of the agglomerative clustering algorithm applied to the Viella and the Pyrenees dataset, and using either the image-BYOL and the human-engineered features sets. First, we assess the influence of temporal and spatial factors on the data distribution within the 4 computed latent spaces, providing insights into the underlying patterns of the dataset. We then look at the content of the clusters produced by agglomerative clustering from polar dendrograms to visualize inter-cluster relationships and propose a seismogenic source labeling scheme. To validate this labeling, we cross-reference the identified groups with an existing seismic catalog. For the Viella dataset, we further examine the temporal occurrence and daily periodicity of the classes derived from continuous data acquisition, comparing them against expected patterns from known seismogenic sources.

5.2.5.1 Impact of Time and Distance on the Data Repartition in Latent Spaces

Visualizing the data distribution in latent space requires to reduce the dimension of the latent space due to the high dimensionality of these spaces—111 dimensions for hand-engineered features and 512 dimensions for image-BYOL features. Several dimensionality reduction techniques can be used, such as Principal Component Analysis PCA (Wold et al., 1987), t-Distributed Stochastic Neighbor Embedding t-SNE (Van der Maaten & Hinton, 2008), and Uniform Manifold Approximation and Projection UMAP (McInnes et al., 2018).

In our study, we use t-SNE to reduce the multidimensional latent spaces (human-engineered and image-BYOL) to two dimensions for the Pyrenees and Viella datasets. t-SNE is primarily influenced by the perplexity and early exaggeration parameters, which impact the preservation of local and global structures, as well as the distances between clusters. To identify the most suitable t-SNE representation, several visual tests were conducted to determine the optimal parameter values. In our case, we set the perplexity to 50 and the early exaggeration to 12. Each point of the t-SNE representation encodes a data block in Figure 5.4. Figure 5.4a,d,g,j show respectively the t-SNE representation of the Pyrenees dataset computed with human-engineered and image-BYOL latent spaces, and the Viella dataset computed with human-engineered and image-BYOL latent spaces. The colors encode the cluster indices after the application of the whole processing chain, and there is no correspondence between cluster indices in the different t-SNE representations.

We represented the t-SNE points in different colors based on time and position along the fiber optic. The results are illustrated in Figure 5.4b,e,h,k for time influence and Figure 5.4c,f,i,l for position influence. We also provide a t-SNE representation with the EB thumbnails for both datasets in Appendix C2 to visually clarify the relationships between points in the t-SNE space.

For the Pyrenees dataset, the time does not influence the localization of points within the feature space, while the spatial influence is strong. The spatial influence is particularly evident when using the human-engineered latent space, which appear to form two distinct clusters (Figure 5.4c). The cluster at the top of the representation consists of points located within a distance of less than 30 km, while the other encompasses the remaining data points. In contrast, the image-BYOL latent space

show a more uniform distribution of points, with no clear separation based on time and less clear separation based on position along the fiber (Figure 5.4f). Distance-based separation is expected in the t-SNE representation with the human-engineered latent space. The natural attenuation of the optical signal along the fiber results in a signal-to-noise ratio degradation that is not entirely mitigated by the common-mode frequency removal technique. This is evident in the t-SNE representation, where the impact of distance is pronounced (Figure 5.4c). In contrast, the quality of image construction and effective data augmentation strategies in generating the image-BYOL latent space considerably reduce distance influence, as shown by a more uniform distribution of points across positions along the fiber in the t-SNE representation (Figure 5.4f). Temporal influence is negligible for the Pyrenees dataset.

For the Viella dataset, the influence of time and position is more prominently reflected in the t-SNE representations derived from both the human-engineered latent space (Figure 5.4h,i) and the image-BYOL latent space (Figure 5.4k,l). In particular, certain areas of the t-SNE representation contain only points measured during specific periods: for example, points measured within the first 11 days or beyond 37 days of the experiment are well-defined in the t-SNE representation space. Regarding distance, as the fiber lost more than half of its length on the first day of the experiment, most points are measured within 300 m. A distribution into two distinct clusters is observed, one corresponding to the first 100 m and the other to the remaining length of the fiber. Temporal influence is present due to the occurences of fiber cuts, with points measured at specific periods more distinctly clustered. In particular, data block points measured after day 37 form an isolated cluster in the image-BYOL representation, and a set of smaller clusters in the human-engineered feature representation. Distance-based separation also reflects differing environmental conditions, particularly for the first 100 m of the fiber, due to the noise from the farm.

5.2.5.2 Clusters Exploration and Labelisation

Dendrogram-based Visualization

Using t-SNE, we previously provided a visual representation of the proximity of data blocks within each of the latent spaces produced. The processing chain then clusters the data blocks within the latent space using K-Means, followed by agglomerative clustering. To display the connection between clusters, we have simplified the dendrogram representation produced by agglomerative clustering. We only show the final clusters, obtained after merging the initial 5000 K-Means clusters down to approximately 500 to 700 clusters. We have generated dendrograms for both datasets, Pyrenees (Figure 5.5) and Viella (Figure 5.6), using the human-engineered (Figure 5.5a and Figure 5.6a) and image-BYOL latent space (Figure 5.5b and Figure 5.6b). Each leaf in the dendrogram represents one of the merged clusters, while each node represents a cluster formed by merging two smaller clusters. Each cluster is depicted by an image that corresponds to the EB representation of the cluster centroid.

For the Pyrenees dataset, the dendrogram generated with the human-engineered latent space highlights certain groups of clusters that appear noisier than others, notably clusters 0 to 30, 75 to 150, and 240 to 338. This noise cluster gathering is less pronounced in the image-BYOL dendrogram. In both dendrograms, a high number of vehicle movement events are identifiable, grouped according to the number of vehicles, their proximity to the fiber, and their direction of travel. A few rarer events are also present, although they are more challenging to distinguish within the clustering structure.

For the Viella dataset, the human-engineered feature dendrogram emphasizes a particularly noisy group of clusters, from cluster 400 to 435 (Figure 5.6a). Examination of Figure 5.4g,h shows

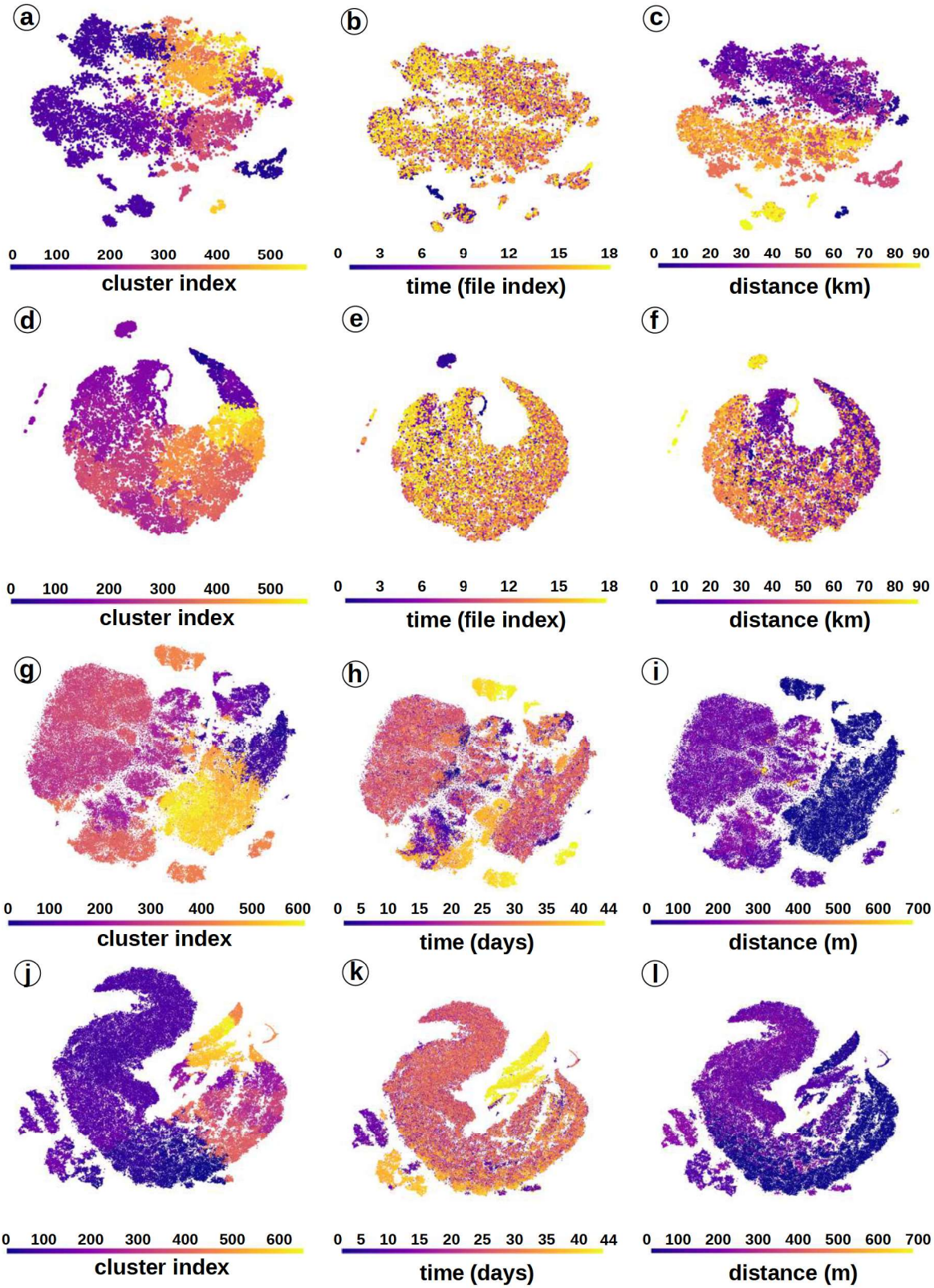


FIGURE 5.4: *t-SNE representation for Pyrenees* ***(a-f)*** *and Viella dataset* ***(g-l)****. The t-SNE processed using the human-engineered latent space is displayed as follows:* ***(a,g)*** *with cluster index encoded by color,* ***(b,h)*** *with time encoded by color, and* ***(c,i)*** *with the position on the fiber encoded by color. The dataset processed using image-BYOL is shown as:* ***(d,j)*** *with cluster index encoded by color,* ***(e,k)*** *with time encoded by color, and* ***(f,l)*** *with the position on the fiber encoded by color.*

that cluster 400 to 435 likely corresponds to measurements taken after 37 days. In the image-BYOL dendrogram, this noisy group expands in size, covering clusters 540 to 640 and 430 to 475. Distinct classes of events are visible: continuous anthropogenic activities (filled circular arc in Figure 5.6); anthropogenic events with intermittent pauses (long dotted circular arc in in Figure 5.6); and isolated anthropogenic events (short dotted circular arc in in Figure 5.6). As with the Pyrenees dataset, rarer events are present but remain challenging to identify in the clustering structure.

Cluster Labeling

The dendrograms offer an initial understanding of the events contained within the clusters; however, a more detailed analysis is required to identify rare seismogenic events. To identify seismogenic event class candidates, we visually examine EB, traces stacked along channels and spectrogram stacked along channels of 25 random data blocks from each of the 500-700 clusters. For the Pyrenees dataset, we identified 5 classes: vehicles, continuous events (those with a constant signal lasting at least 10 s), impact events (signals lasting less than 10 s), earthquakes or quarry blasts (EQ/QB), and noise. For the Viella dataset, we identified 6 classes: continuous short events, continuous long events, multiple impact events, unique impact events, earthquakes, and noise. Appendix C3 provides several examples of data block cluster centers for each cataloged class. As the purity of the clusters varies, labeling is based on the population of each cluster. For events with equivalent occurence frequency, particularly among anthropogenic events, the predominant class within the cluster is assigned as the label. For rarer events, such as earthquakes, a cluster is labeled as such if at least 25% of the data blocks belong to the rarer class. We manually rank the events as follows:

1. Very frequent events: In the Pyrenees dataset, the most common events identified are vehicles and noise, while in the Viella dataset, noise is the predominant very frequent event.

2. Frequent events: The frequent events identified in the Pyrenees dataset include continuous events and impact events. In contrast, the Viella dataset shows a broader range of frequent events, encompassing continuous short events, continuous long events, multiple impact events, and unique impact events.

3. Rare events: The rare events detected in both datasets are earthquakes, and quarry blasts for Pyrenees dataset.

To obtain an overview of the distribution of the data block forming each of the identified classes, we used the t-SNE map (Figure 5.7). In Figure 5.7, each point represents a data block expressed in latent space, and each color indicates the manually assigned class label. Figure 5.7a,b,c,d respectively shows the Pyrenees dataset with human-engineered and image-BYOL latent space, and the Viella dataset with human-engineered and image-BYOL latent space. We observe that the more common classes are well-separated on the t-SNE map. Rarer classes tend to cluster in small isolated groups, often mixed with more frequent classes, particularly for earthquake or quarry blast (EQ/QB) events. This tendency does not suggest that the features are ineffective but rather illustrates the limitations of the t-SNE representation. The objective is to display the entire dataset in a highly reduced space, which means the major trends among the classes are prioritized. This tendency does not suggest that the features are ineffective but rather illustrates the limitations of the t-SNE representation. The objective of the representation is to display the entire dataset in a highly reduced space, which means the major trends among the classes are prioritized.

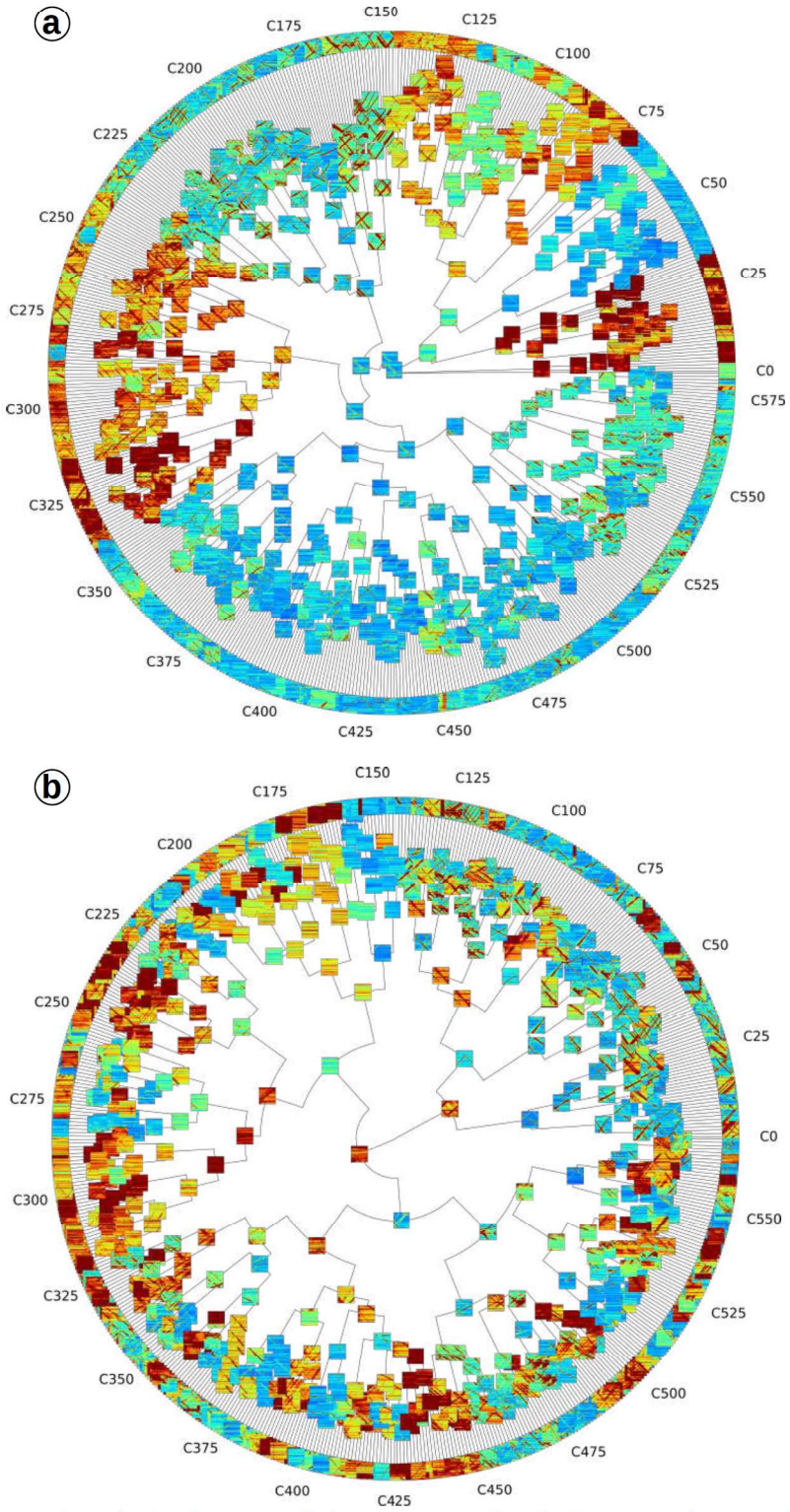


FIGURE 5.5: Dendrogram-based visualization of cluster content for the Pyrenees dataset. ***(a)*** *shows the dendrogram generated using the human-engineered latent space, while* ***(b)*** *displays the dendrogram generated with the image-BYOL latent space. C xxx marks the cluster index for every 25 clusters. The images correspond to the EB representation of the centroid for each original cluster (leaf nodes of the dendrogram) and for each cluster formed through agglomerative clustering (internal nodes of the dendrogram).*

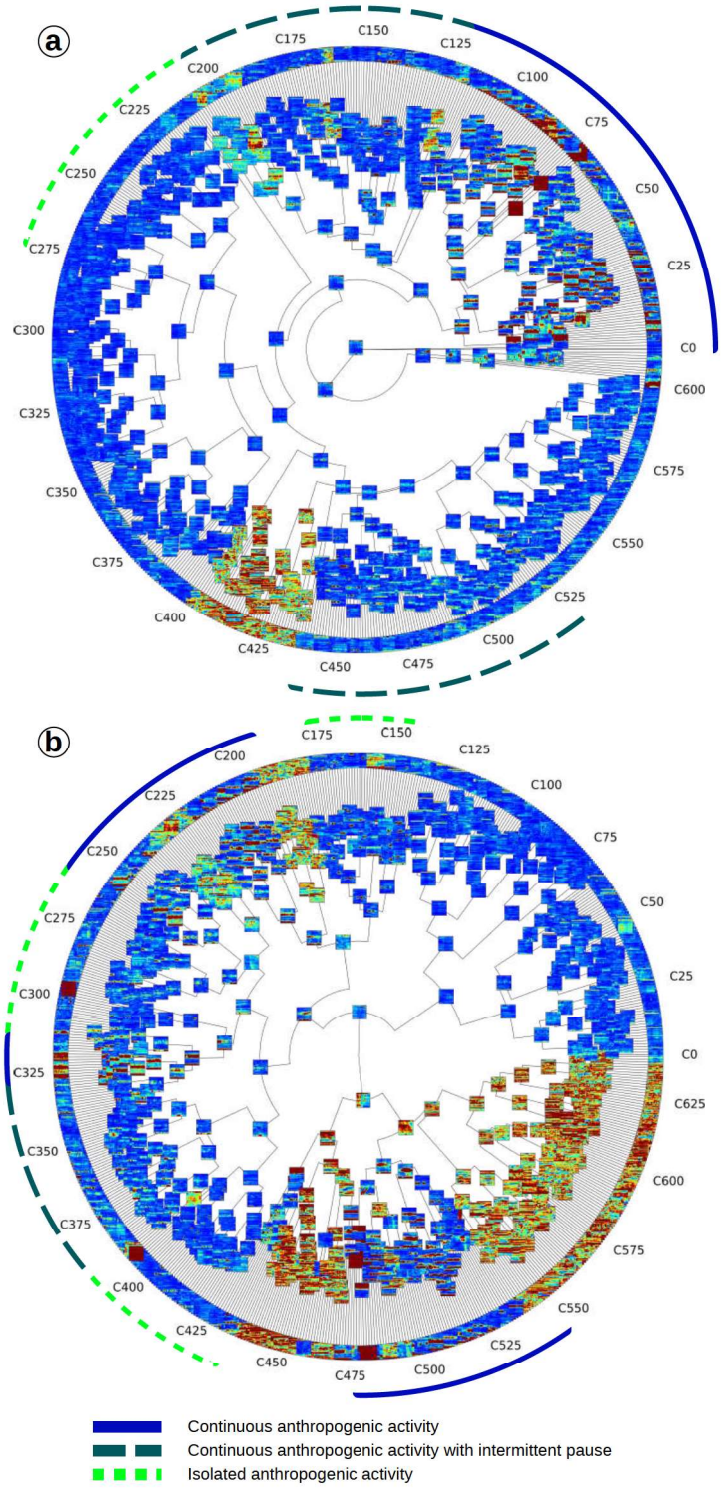


FIGURE 5.6: Dendrogram-based visualization of cluster content for the Viella dataset. ***(a)*** *shows the dendrogram generated using the human-engineered latent space, while* ***(b)*** *displays the dendrogram generated with the image-BYOL latent space. C xxx marks the cluster index for every 25 clusters. The images correspond to the EB representation of the centroid for each original cluster (leaf nodes of the dendrogram) and for each cluster formed through agglomerative clustering (internal nodes of the dendrogram).*

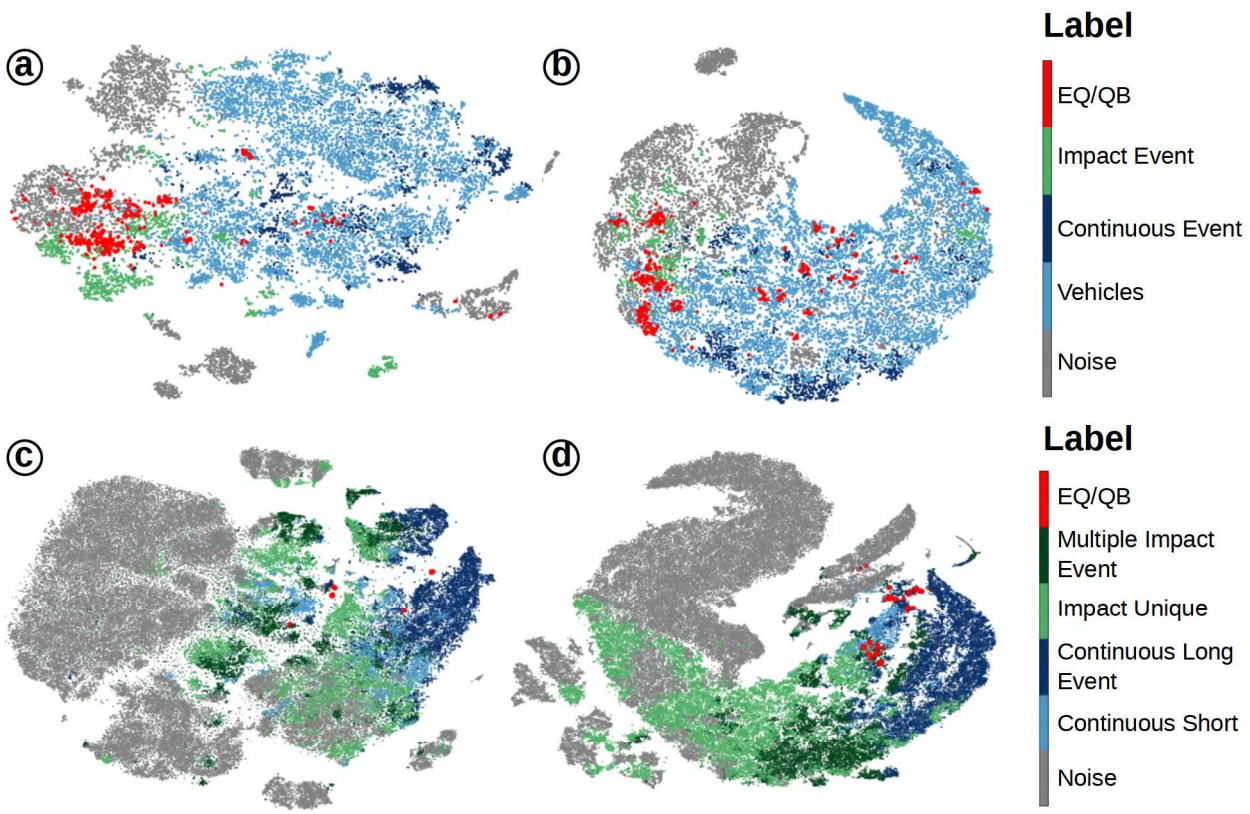


*FIGURE 5.7: t-SNE representation for Pyrenees **(a,b)** and Viella dataset **(c,d)**. The colors are encoding the hand-attributed class. **(a,c)** are obtained with the human-engineered latent space, **(b,d)** are obtained with the image-BYOL latent space. For vizualization purpose, the earthquake class is represented with points scaled to size 5 compared to the others for the Pyrenees dataset, and with points scaled to size 150 for the Viella dataset.*

5.2.5.3 Validation

The validation process involves comparing the observed behavior of the constructed classes with expected behaviors, using a reference catalog such as BCSF-RENASS for point and known seismogenic source events, or predefined patterns for periodic events. For the Pyrenees dataset, "earthquake or quarry blast" (EQ/QB) class is validated by comparing the class occurence with a reference catalog for EQ/QB detection. For the Viella dataset, we perform multiple validation: first, we analyze the daily and weekly periodicity of the classes to identify patterns related to anthropogenic events, then we compare the cluster labels with a reference catalog for earthquake detection. Additionally, we validate the earthquake detection by applying the same clustering and a earthquake versus noise labeling process to seismic node data to assess the robustness and accuracy of the results.

Pyrenees Dataset: Comparison with a Reference Catalog for Earthquake Detection

The validation of the handmade classes of seismogenic events in the Pyrenees dataset is carried out by comparing the proportion of data blocks classified as earthquake or quarry blast (EQ/QB) for each time step (1-min interval) and compare this proportion with the median of the data over 10 min. Detection was considered successful if the proportion exceeded the median value plus two median absolue deviation (MAD) of the data over 10 min. The median absolute deviation (MAD) is a more robust measure of variability compared to the standard deviation, as it is less sensitive to

Id	Class	Mag(M_w)	Dist(km)	Time(UTC)	Detected by
QB1	QB	0.7	1.8	2022-08-31 at 08:02:39	human-engineered, image-BYOL
QB2	QB	0.8	5.8	2022-09-01 at 09:26:00	-
EQ1	EQ	0.6	1.7	2022-09-03 at 03:50:17	human-engineered, image-BYOL
EQ2	EQ	1.0	10	2022-09-03 at 13:13:41	-
EQ3	EQ	2.4	85	2022-09-03 at 18:27:47	image-BYOL
EQ4	EQ	1.4	25	2022-09-05 at 00:51:58	image-BYOL
EQ5	EQ	1.2	6.5	2022-09-06 at 07:58:54	image-BYOL
QB3	QB	1.0	4	2022-09-06 at 10:10:58	human-engineered, image-BYOL
QB4	QB	1.1	1	2022-09-08 at 10:03:36	human-engineered
EQ6	EQ	0.8	3.2	2022-09-08 at 17:10:59	human-engineered, image-BYOL
EQ7	EQ	2.0	50	2022-09-09 at 07:07:40	-
EQ8	EQ	0.4	1.5	2022-09-09 at 17:37:59	-
QB5	QB	0.6	0.5	2022-09-12 at 10:07:43	human-engineered, image-BYOL
QB6	QB	1.1	6.3	2022-09-14 at 09:26:07	-
EQ9	EQ	1.1	8.5	2022-09-15 at 09:12:16	human-engineered, image-BYOL
EQ10	EQ	1.6	1.4	2022-09-16 at 11:16:50	human-engineered
EQ11	EQ	1.1	20	2022-09-16 at 23:12:33	-
EQ12	EQ	0.8	1.0	2022-09-18 at 04:18:12	human-engineered, image-BYOL
EQ13	EQ	1.3	9.5	2022-09-20 at 04:39:26	human-engineered, image-BYOL

TABLE 5.2: Pyrenees dataset: Earthquakes (EQ) and Quarry Blast (QB) detected by BCSF-RENASS and visually confirmed on DAS data.

outliers and extreme values in the data (Leys et al., 2013). The validation process was carried out using the BCSF-RENASS reference catalog for a direct comparison with known seismogenic events. Using this validation process, we successfully detect 9 out of 19 true earthquake or quarry blast events with the human-engineered latent space and 11 of 19 with the image-BYOL latent space. A detailed list of the detected EQ/QB events is provided in Table 5.2.

Among the undetected events using both methods (QB2, EQ2, EQ7, EQ8, QB6, EQ11), it is noteworthy that 4 of these undetected events have a magnitude greater than 1.0. Three of these events, which were earthquakes (EQ), are located at distances greater than 10 km from the fiber. Some events are detected by BYOL but not by the hand-engineered method (BYOL: EQ3, EQ4, EQ5), and vice versa (hand-engineered: QB4, EQ10). Specifically, for events detected using image-BYOL approach only, two events are located at more than 25 km from the fiber optic (EQ3, EQ4) and one event is located at 6.5 km (EQ5). For events detected using the hand-engineered method only, two event are located nearby the fiber optic (QB4,EQ10).

The mixed results can be explained by the presence of a cluster containing several classes of events labeled as EQ, which tend to produce detections even in areas where no real events occur. These clusters, called impure clusters, may increase the proportion of data blocks required to be considered as EQ as the criteria involves median and MAD over a 10-min record, thereby questioning the reliability of the decision criterion. As a result, events in regions with low seismogenic activity or high noise levels may be incorrectly identified, leading to false positives.

Viella Dataset: Periodicity Analysis for Anthropogenic Events

The Viella dataset provides continuous data, enabling a detailed temporal analysis of the detection results. We can examine the temporal distribution of the detected classes over the entire acquisition period in Figure 5.8a-f. Additionally, to explore any periodic patterns, we present the distribution on a daily and weekly basis in Figure 5.8g,h.

Observations from these figures reveal notable patterns in the periodicity of different classes. The noise class exhibits a clear daily periodicity, characterized by peaks of silence at night and troughs during the day, as shown in Figure 5.8a,g,h. The anthropogenic classes display varying degrees of daily and weekly periodicity: the continuous long events class shows a strong periodicity, with event proportions reaching up to 0.2 during peak activity times, compared to an average of 0.1 in the evenings (Figure 5.8g,h). The continuous short and impact multiple classes exhibit less pronounced periodicity, with slight increases during the day. In contrast, the impact unique class does not display any particular periodicity. The results can be interpreted in light of these periodic patterns. The periodicity observed in the continuous long events class suggests that high-energy localized signals are likely to originate from anthropogenic activities, particularly agricultural work in the context of Viella. On the other hand, the lack of periodicity in the impact unique class indicates that these events are sporadic and may not solely be of anthropogenic origin, suggesting a need for further investigation into their sources.

Viella Dataset: Comparison with a Reference Catalog for Earthquake Detection

Several potential earthquakes are detected during the measurement period, making cross-referencing with established seismic picking relevant. For this purpose, we combine the seismic databases provided by the Observatoire Midi-Pyrenees (OMP) and BCSF-RENASS. Using the OMP and BCSF-RENASS databases, a total of 20 earthquakes are visually confirmed on the DAS data. Table 5.3 provides details of these events, including their date, magnitude and epicenter distance from the fiber network. Detection is effective for moderate-magnitude events for both methods, with all of events above $M_w = 2.0$ identified (up to a distance of 63 km), and the farthest event detected at 143 km for a magnitude of $M_w = 2.8$. However, detection is more inconsistent for lower-magnitude events across both approaches.

Regarding detection volume, clustering performed using the cataloged latent space identified a total of 16 events, including 8 cataloged earthquakes (with the smallest at $M_w = 1.5$ and a distance of 13 km) and 3 visually verified but uncataloged potential earthquakes. In comparison, the image-BYOL method detects 55 events, including 9 cataloged earthquakes (smallest at $M_w = 1.4$ and 28 km away) and 5 uncataloged potential earthquakes. Figure 5.9 shows the temporal distribution of detected earthquakes using the cataloged latent space (Figure 5.9b) and the image-BYOL latent space (Figure 5.9c). The cataloged latent space approach is promising, as it raises fewer false alarms than the image-BYOL method and allows filtering by setting an occurrence threshold to retain only 11 potential earthquakes. In contrast, the image-BYOL method generates more false positives and does not offer a direct selection criterion based on occurrence proportion.

Viella Dataset: Comparison with Nodes Data for Earthquake Detection

To further assess the results obtained with the DAS, we also analyzed data from the seismic nodes that were installed alongside the fiber optic cable. Due to a fiber breakage on the first day of measurement, with a loss of 800 m to 312.4 m of fiber optic cable, we chose to use only data from seismic

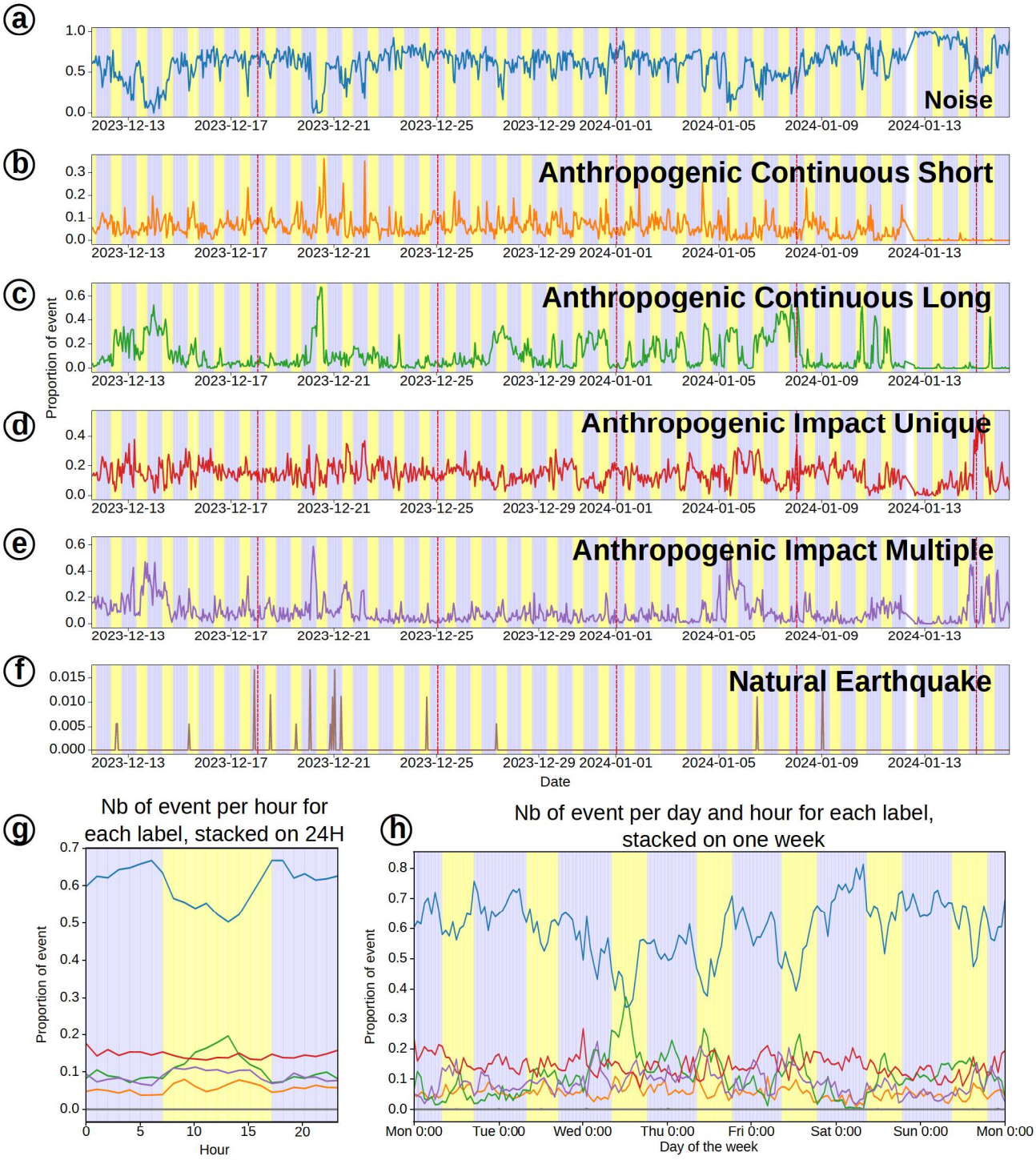


FIGURE 5.8: Temporal distribution of the manual labels for Viella dataset (fulltime measurement period). The x-axis represents the time, and the y-axis represents the proportion of events per class.

Id	Date	Magnitude(M_w)	Distance(km)	Detected by
1	2023-12-12 07:20:00	0.4	17	image-BYOL
-	2023-12-12 22:59:17	1.5	14	-
2	2023-12-14 10:43:51	2	63	human-engineered, image-BYOL
3	2023-12-17 21:52:13	2.5	48	human-engineered, image-BYOL
-	2023-12-19 01:26:08	1.6	48	-
4	2023-12-20 02:21:04	2.7	41	human-engineered, image-BYOL
5	2023-12-20 23:24:43	3.4	143	human-engineered, image-BYOL
6	2023-12-21 01:20:08	2.8	143	human-engineered, image-BYOL
7	2023-12-21 06:58:05	3.1	143	human-engineered, image-BYOL
-	2023-12-21 19:56:07	1.0	19	-
-	2023-12-23 04:12:09	1.0	22	-
-	2023-12-23 04:59:37	0.4	8	-
8	2023-12-24 15:41:09	2.2	41	human-engineered, image-BYOL
-	2023-12-24 19:37:54	1.1	8	-
-	2023-12-28 02:46:17	0.8	18	-
9	2023-12-29 01:03:48	1.4	28	image-BYOL
-	2023-12-31 11:26:18	0.5	17	-
-	2024-01-06 04:57:28	1.4	20	-
10	2024-01-09 01:47:26	1.5	13	human-engineered
-	2024-01-09 03:52:04	1.0	13	-

TABLE 5.3: Viella dataset: Earthquakes detected by BCSF-RENASS or OMP and visually confirmed on DAS data.

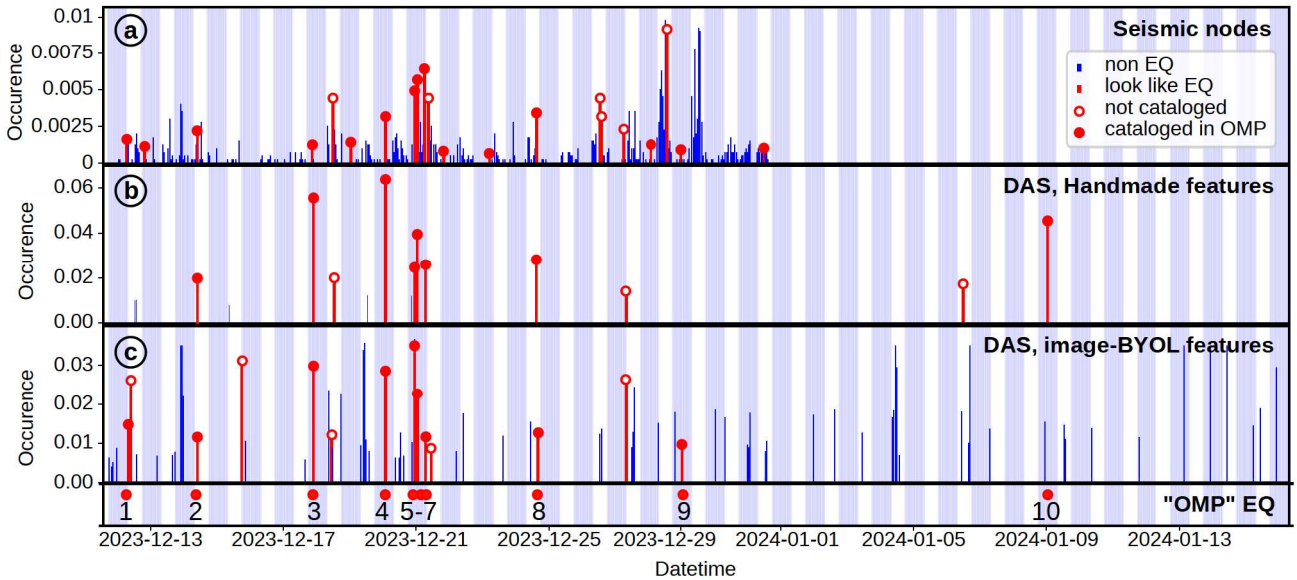


FIGURE 5.9: Comparison of detection raised by the seismic nodes using image-BYOL latent space (a), the cataloged latent space (b), and the image-BYOL latent space (c) with the BCSF-RENASS or OMP database. Events represented by blue bar are raised by the system but did not correspond actually to an earthquake, events marked by red bars visually align with earthquake characteristics. Events represented by red bars with filled circles are confirmed as earthquakes and cataloged in the BCSF-RENASS or OMP database. The x-axis represents the time, and the y-axis represents the proportion of events per class.

nodes 1 to 6 in Figure 5.1. Additionally, since the nodes had limited operational autonomy due to power supply constraints, we restricted the analysis to data recorded up until January 1st. For this analysis, we first remove the geophone instrumental response from the seismic trace through deconvolution. We then segment the data into blocks of the same temporal size as the DAS data (1 min) and apply the image-BYOL method to images that combine the seismic traces with their corresponding spectrograms. We also use the same data augmentation as for the DAS data. The analysis successfully detected all 21 events identified by the BCSF-RENASS or OMP (Table 5.3), but it also resulted in a higher number of false positives compared to the image-BYOL approach applied to DAS data (Figure 5.9a). This suggests that while the image-BYOL model provides comprehensive event detection, further refinement is necessary to reduce the occurrence of false positives, particularly when detecting weaker seismogenic events.

5.2.6 Discussion and Conclusion

The objective of this paper is to propose a solution for labeling continuously measured DAS data in preparation for supervised classification learning. To achieve this, we propose a clustering-based approach to perform per-cluster labeling. We rely on two different latent space that we would like to compare: an human-engineered latent feature space and a latent space learned using a self-supervised learning model approach called image-BYOL. For both latent space, we applied a clustering approach to group similar data, followed by hierarchical clustering to reduce the number of clusters. Finally, we labeled these clusters into a few distinct classes.

Our method enables the automatic identification of key factors differentiating data, whether due to the nature of the measured event (anthropogenic, natural) or instrumental factors (influence of distance on the data, effect of fiber breakage), and offers direct applications in anomaly detection. The method can be adapted for data labeling per clusters, where human intervention is necessary. For frequent events such as anthropogenic activities linked to agriculture in the Viella dataset, the method detects temporal periodicities and distinguishes them from other classes. In the case of rare events, such as earthquakes, the approach requires the definition of new rules to address the issue that clusters are not always pure, such as "consider the cluster as an earthquake if 25% of the data blocks actually contain earthquakes". This rule detects all earthquakes with a magnitude greater than 2.0 in the Viella dataset using both latent spaces, but generates a higher false positive rate for events with smaller magnitudes. For the Pyrenees dataset, the results obtained using both human-engineered features and those learned through image-BYOL do not show a clear advantage for either approach. In some cases, detections appeared inconsistent, with events occasionally missed despite their presence in the data. These observations point to challenges inherent in the clustering method applied, particularly the cluster labeling rules, which may require refinement to better handle the complexity of such environments. Complex settings, characterized by overlapping signals or strong background noise, pose difficulties for the clustering method to accurately differentiate between events. Furthermore, the long distances covered by the fiber optic cable can lead to signal attenuation and delays, introducing additional challenges for the reliable detection and classification of seismic events. A potential solution to improve the detection of low-magnitude events is to apply a re-clustering method on inpure clusters. The clustering process will be applied iteratively to the data blocks that were not correctly classified in the previous iteration, allowing the creation of new clusters that are potentially purer in next iteration.

Both human-engineered and image-BYOL feature extraction show similar effectiveness overall. On the Viella dataset, human-engineered features result in fewer false detections, while the difference is less pronounced on the Pyrenees dataset. Therefore, both methods should be tested to identify the most suitable approach for a specific application. From computational considera-

tion, human-engineered features are faster to compute, as the features are predefined. In contrast, BYOL SSL approach requires a training phase, which is more time-intensive but eliminates the need for human feature engineering. For readers looking to replicate this work and who have access to computational resources such as GPU, we recommend starting with the BYOL approach. This method is entirely independent of prior feature engineering efforts and allows for a full evaluation of the processing chain performance. For those without access to such resources, the use of human-engineered features provides a less computationally intensive alternative. This second approach requires downloading the feature calculation code provided in (Huynh et al., 2024).

The application of unsupervised learning techniques, particularly clustering of latent spaces from SSL algorithms or human-engineered features, offers a promising approach for rapidly generating seismic datasets that can be used to train models for real-time monitoring or large-scale analysis of pre-existing data. By leveraging clustering methods, such as K-Means and agglomerative clustering, we can efficiently group seismic data and identify patterns without the need for manual labeling. This capability significantly enhances the potential of DAS as a monitoring tool, which has been a key challenge in the field. Furthermore, this approach allows for the development of intelligent event-triggering systems, enabling monitoring beyond traditional event-based detection and providing a more continuous and adaptive framework for seismogenic event monitoring.

5.3 Chapter Summary

In this chapter, we presented a method for identifying clusters of seismogenic events from DAS measurements. The approach focuses on building or selecting a robust latent space representation, clustering the DAS data within this space, and using the clusters to create a labeled dataset. This method was tested on two datasets: the Pyrenees dataset, which contains 19 ten-minute recordings of events and noise over a 91-km fiber optic cable, and the Viella dataset, comprising 44 days of continuous recordings over an 800-m fiber optic cable. Validation was carried out by comparing results with reference catalogs, analyzing periodic patterns in the data, and cross-referencing with seismic node measurements.

The method was evaluated on both frequent events (such as farm activity, occurring almost daily) and rare events (like earthquakes). The results show that the method is effective for frequent events, which are tedious and difficult to label manually. For rare events, such as earthquakes, the method works well for labeling seismogenic events, especially those with magnitudes greater than Mw=2.0. Looking ahead, the method can be used to explore data and identify classes of events, which can then be applied in supervised models. However, with larger datasets, optimizing processing time and storage will be necessary. To improve accuracy and reduce false alarms, re-clustering impure clusters could help enhance sensitivity to lower-magnitude events. Additionally, the method has potential for other tasks, such as anomaly detection.

Chapter 6

Conclusions and Outlook

6.1 Conclusions

6.1.1 Result Summary

Through this thesis, we aimed to design and implement complete processing chains for the automatic classification of Distributed Acoustic Sensing (DAS) data. Each chapter addressed a key aspect, presenting encouraging results while also highlighting several limitations.

In **Chapter 2**, we provided a comprehensive overview of DAS applications in seismology, focusing on the types of seismic signals DAS can capture and the conventional and advanced data processing techniques used in this field. This overview guided the design of complete processing chains for the automatic classification of DAS data streams, which are detailed in Chapter 3. The processing chain presented in **Chapter 3** consists of three essential steps: pre-processing, classification, and post-processing. Pre-processing ensures that the instrumental response is removed and that the data is segmented into fixed spatial and temporal windows suitable for streaming analysis. The classification step generates a latent space representation using 57 features derived from conventional seismology and applies supervised machine learning algorithms like Random Forest or XGBoost to classify the data. Finally, post-processing incorporates the spatial coherence of DAS data through a Markov random field model to refine the classification. The method, tested in a controlled environment at the FEBUS Optics test center, achieved a classification accuracy of 87% for six event classes, with spatial redundancy improving accuracy by 4%. However, the tests were conducted in controlled conditions with minimal external noise and a trench length of 22 m, which limits their applicability to in-situ scenarios. Building on this work, **Chapter 4** introduced new features that account for the spatial characteristics of DAS data, expanding the feature set to 111. Using a 91-km telecommunication fiber cable deployed in the Hautes-Pyrénées, we demonstrated the system ability to detect seismogenic events, including earthquakes as small as Mw=0.4 within 1.5 km of the fiber and quarry blasts within 4 km. This approach proved effectiveness even under challenging conditions with variable SNR and multiple anthropogenic sources along the fiber. However, supervised learning remained constrained by the need for precise data labeling. The model can be evaluated using the 19 seismogenic events identified from the BCSF-RENASS catalog and leave-one-out cross-validation; however, the results cannot be generalized to other situations. Manual labeling was labor-intensive and limited the model genericity. To address this challenge, **Chapter 5** explored an unsupervised clustering approach to group similar data, enabling cluster-based labeling. For the Viella dataset, which consisted of 44 days of continuous recording, this method reduced the label-

ing effort from 142 727 individual data windows to 640 clusters using BYOL-based features and 612 clusters with human-engineered features. The clustering approach identified frequent events, such as vehicle movements and agricultural activities, along with their daily or weekly periodic patterns. It also facilitated the detection of rare and natural events, such as local earthquakes of tectonic or landslide origin, using pre-defined rules. However, cluster impurity and the reliance on arbitrary pre-defined rules to control false positives and negatives remained significant challenges, sometimes leading to an overclassification of non-earthquake events as earthquakes.

Across all approaches, efforts were made to minimize the influence of subjective human labeling. During the measurements conducted at the FEBUS Optics test bench, events were generated under supervision (Chapter 3), and for the Hautes-Pyrénées dataset, cataloged events from BCSF-RENASS provided an objective reference (Chapter 4). In Chapter 5, we have chosen not to rely on an external catalog. The shift to cluster-based labeling reduced labeling subjectivity but introduced challenges related to cluster purity and the selection of the appropriate number of clusters. Additionally, the need for pre-defined rules to manage false positives and negatives highlighted ongoing challenges in detecting rare events.

6.1.2 Thesis Contributions

This PhD thesis bridges the gap between conventional seismology and fiber optic-based seismology. It presents approaches that **combine traditional signal processing tools**, focused on temporal waveform, spectrum, and spectrogram analysis, with new tools that capture additional information, such as **spatial waveform analysis** and **channel similarity**. By extracting features for event classification, this approach enables the effective use of DAS data for seismic analyses. Another contribution of this PhD thesis is to provide a solution to the challenges posed by the large volume of DAS data when applying supervised learning algorithms. A scalable solution is particularly needed for continuous DAS data collected over long periods and across extensive fiber networks. In this PhD thesis, we explore **grouping detected seismological events by similarity**. This approach proves effective for both frequent and rare events, such as monitoring natural seismicity in active areas or tracking anthropogenic sources like traffic and construction, regardless of dataset size.

The methods can have a significant impact, particularly in areas where conventional seismology is less effective or where new challenges arise. DAS offers promising solutions for monitoring construction activities, detecting intrusions, identifying pipeline leaks, and enhancing earthquake detection in noisy environments. However, DAS is sensitive to a variety of sources, making event classification challenging and requiring the use of AI algorithms. For dataset creation, **clustering** techniques provide a key solution by enabling the **rapid annotation of datasets** based on event similarity measures. For **classification** process, the proposed approach is **interpretable**, with features based on physical quantities and their importance quantified using methods like Random Forest and XGBoost. These advancements could **ultimately contribute to the development of early warning systems**, identifying clusters of events that precede major incidents such as earthquakes or landslides.

Despite the promising results, several limitations were encountered during this research. The studies were carried out over relatively short periods, which restricted the amount of data available for training supervised classification models, particularly for rare events. The scalability of the approach to continuous, long-term monitoring was not fully addressed, and the optimal amount of data required to maintain a functional system over extended periods remains unclear. Additionally, the models developed during this thesis were site-specific, and the results could not be easily generalized to other environments. Future work could focus on training models with data from multiple

sites to create more generalized systems that are less dependent on specific measurement configurations. Additionally, transfer learning approaches could be explored for applications across different sites or with different seismological instruments. From an applicative and industrial perspectives, a classification model that is generalizable to different sites and environments would be a significant asset for deploying large-scale seismic monitoring systems or temporary monitoring solutions. The need for new data to monitor a specific site often means that the system cannot be operational quickly and requires extensive fieldwork. For long-term monitoring, combining data from different sites could enable pre-training of the model, which could then be fine-tuned with data from new sites.

6.2 Perspectives and Future Research

6.2.1 Methodological Improvements for DAS Data Processing and Classification

Improving Clustering with Iterative Re-Clustering on Impure Clusters

Chapter 5 highlighted the subjectivity in unsupervised clustering approaches, especially for identifying rare events in continuous measurements (e.g., earthquakes), which require predefined rules to control false positives and negatives. To improve cluster purity and reduce the reliance on these rules, one possible solution is to introduce an **iterative re-clustering method**. This method would refine impure clusters by reapplying the clustering process to misclassified data, allowing more pure clusters to form in subsequent iterations. This approach, which has been successfully applied in fields such as image segmentation (Bensaid et al., 1996) and autonomous driving systems (Kruber et al., 2018), could be adapted for DAS data classification, enabling more accurate identification of seismogenic events, including those of low and very low magnitudes.

Transfer Learning Across Different Sites and for Several Measurement Instruments

Transfer learning is a powerful strategy that reduces the need for extensive training data while preserving high system performance. By leveraging knowledge from one domain, it enables models to adapt to new tasks or datasets with minimal additional training. This approach is particularly valuable in fields like seismology, where acquiring large amounts of labeled data can be challenging. To this end, data sharing within the scientific community can be of interest, as developed in the following Section 6.2.2. As illustrated in the Equation 1.1, the measured seismic signal $u(t)$ is influenced by several factors, including the source $s(t)$, the signal propagation through the ground $g(t)$, the sensor placement, and the instrument response $i(t)$.

$$u(t) = s(t) * g(t) * i(t) \quad \text{where } * \text{ denotes the convolution operator} \tag{1.1}$$

When transferring knowledge between several FO-DAS datasets, **several challenges** arise due to variations in environmental conditions, sensor placements, and site-specific factors. These differences can lead to changes in the recorded data, causing feature distributions to vary among sites. As a result, the model ability to generalize across datasets may be compromised, making it difficult to apply a model trained on one dataset to a new, unseen environment. Transferring knowledge

from conventional seismometers to FO-DAS instruments presents additional hurdles. Conventional seismometers have a fixed frequency response, typically measuring frequencies from a few Hz to several kHz. In contrast, FO-DAS instruments feature a variable frequency response, influenced by factors such as system configuration, fiber type, and measurement conditions. DAS data also suffers from a lower SNR ratio, which can distort the data and lead to biased results in AI models.

6.2.2 Data Sharing and Collaboration in FO-DAS Seismology

Sharing DAS data is challenging due to the large size of datasets and the use of proprietary **file formats** by different manufacturers. For instance, Silixa uses the TDMS format, while Schlumberger, Omnisens, OptaSense, and Fotech Solutions employ other proprietary formats. While these manufacturers provide tools to convert data into more accessible formats like HDF5 or SEG-Y, companies such as FEBUS Optics and ASN directly use non-proprietary formats, such as HDF5, for storing DAS data. However, a key issue remains the inconsistency and poor documentation of important **metadata**, such as optical configuration, fiber GPS location, measurement timestamps and fiber properties (telecommunication fiber cable, loose-tube cable, tight cable or cable specially engineered for measurement purposes) across datasets. To address this, Lai et al. (2024) proposes a comprehensive metadata standard that includes both instrument configurations (e.g., interrogator type, acquisition parameters) and sensor locations (e.g., cable installation, fiber properties).

These challenges hinder the sharing of DAS data and make comparisons across studies difficult. Nevertheless, there is an increasing effort within the scientific community to promote DAS **data sharing**, particularly for the development and comparison of signal processing algorithms and AI models. PubDAS, an open-source repository introduced by Spica et al. (2022), is a prime example, hosting 90 TB of DAS data from experiments across various geological environments. Our own contribution to this effort includes sharing the Hautes-Pyrénées dataset, available through our publication under the DOI: 10.57932/5b1302d6-57cd-44e4-81ac-5d585a7f8951.

In parallel with data sharing, providing **standardized processing tools libraries** is crucial for advancing DAS research. Hu and Li (2024) offers a Python library that includes tools for preprocessing, filtering, frequency analysis, signal decomposition, and denoising techniques like curvelet denoising. Similarly, Trabattoni et al. (2024) presents a Python library designed for efficient data loading and file combining, which mimics the API of popular Python libraries like NumPy/Scipy/Xarray to encourage compact and effective code. As AI models such as PhaseNet (Zhu & Beroza, 2019) and its DAS version (Zhu et al., 2023) gain in attractiveness, the **formatting of trained models** and the **metadata associated to the classification process** required becomes increasingly important. This highlights the need for standardized data formats across DAS manufacturers. Metadata should encompass model parameters (accessible through Python libraries like Scikit-learn or PyTorch), preprocessing/post-processing algorithms, evaluation methods, and the training data used to build the models. Our own contribution to this effort includes sharing the processing chain for classification process, available through our publication in the GitLab repository "EOST/seis-learning-spatial".

6.2.3 Computing Resources and Scalability

When processing DAS data, several solutions are available, each offering different trade-offs between computational power, latency, and use case requirements. These solutions can be ranked based on computational capacity, though higher capacity often results in increased latency due to data transfer to remote servers. Solutions can be ranked as follows:

1. **Edge computing**
   The simplest approach is to process data directly on the DAS instrument, minimizing latency since no data is sent elsewhere. However, this option is limited by the computational resources of DAS instrument, making it suitable for simpler tasks with manageable data volumes that require real-time results. During the thesis, I worked with several FEBUS software engineers to implement the processing chain presented in Chapter 3 into FEBUS A1-R DAS for demonstrations to several company customers. For this implementation, given the time constraints of real-time processing, we choose the features to implement based on their importance in classification process.

2. **Local computing using a portable workstation**
   More powerful options are the use of portable workstations, offering greater computing power for real-time analysis. However, this introduces some latency due to the need to transfer data from the DAS instrument to the workstation. This latency can be minimized by connecting the devices via LAN 100 GHz cable.

3. **Distant computing using high-performance computing (HPC) systems**
   HPC systems offer substantial computational power, enabling parallel processing of large datasets. However, it introduces significant latency due to data transmission to the HPC cluster and requires a reliable internet connection. This approach is ideal for tasks like prototyping new algorithms, training AI models, or processing large datasets that require extensive computational resources. However, it is less suited for real-time applications.

Depending on the objective of the work, certain solutions are more suitable than others. For field applications involving classification tasks in a real-time application, we believe the best approach is to use Solution 2, as it offers the best trade-off between computational power and latency time for data processing. If the data can be easily uploaded or retrieved before actual application, it is possible to perform pre-recordings, collect the data, and train a model on an HPC system before uploading the trained model to the workstation. For data exploration tasks, where we aim to explore one or more datasets spanning several weeks or even years, the goal is to construct catalogs solely based on the content of the data. In this case, there are no time latency constraints, making HPC processing a more advantageous choice.

When considering Solution 3, data is exchanged between the DAS instrument and the processing system, with latency time proportional to the amount of data being exchanged. As a result, exploring data compression algorithms can be advantageous. These algorithms reduce data volume for real-time processing, allowing for lossy compression when processing is not impacted, and for long-term storage, where lossless compression is crucial. For instance, Dong et al. (2022) proposes a lossless compression method for DAS data, achieving up to 40% compression without any data loss.

References

Abancó, C., Hürlimann, M., Fritschi, B., Graf, C., & Moya, J. (2012). Transformation of ground vibration signal for debris-flow monitoring and detection in alarm systems. *Sensors*, *12*(4), 4870–4891.

Ackerley, N., Beer, M., Kougioumtzoglou, I., Patelli, E., & Au, S. (2014). Principles of broadband seismometry. *Encyclopedia of Earthquake Engineering,(Springer, Berlin)*.

Ajo-Franklin, J. B., Dou, S., Lindsey, N. J., Monga, I., Tracy, C., Robertson, M., … Li, X. (2019, December). Distributed Acoustic Sensing Using Dark Fiber for Near-Surface Characterization and Broadband Seismic Event Detection. *Sci Rep*, *9*(1), 1328. Retrieved 2021-12-01, from http://www.nature.com/articles/s41598-018-36675-8 doi: 10.1038/s41598-018-36675-8

Akansu, A. N., & Haddad, R. A. (2001). *Multiresolution signal decomposition: transforms, subbands, and wavelets*. Academic press.

Aki, K., & Chouet, B. (1975). Origin of coda waves: source, attenuation, and scattering effects. *Journal of geophysical research*, *80*(23), 3322–3342.

Aki, K., Fehler, M., & Das, S. (1977). Source mechanism of volcanic tremor: Fluid-driven crack models and their application to the 1963 kilauea eruption. *Journal of volcanology and geothermal research*, *2*(3), 259–287.

Aki, K., & Richards, P. G. (2002). Quantitative seismology. sausalito. *Calif: University Science Books*.

Al Hasani, M., & Drijkoningen, G. (2023). Experiences with distributed acoustic sensing using both straight and helically wound fibers in surface-deployed cables—a case history in groningen, the netherlands. *Geophysics*, *88*(6), B369–B380.

Allen, R. (1982). Automatic phase pickers: Their present use and future prospects. *Bulletin of the Seismological Society of America*, *72*(6B), S225–S242.

Allen, R. M., & Stogaitis, M. (2022). Global growth of earthquake early warning. *Science*, *375*(6582), 717–718.

Allmann, B. P., Shearer, P. M., & Hauksson, E. (2008). Spectral discrimination between quarry blasts and earthquakes in southern california. *Bulletin of the Seismological Society of America*, *98*(4), 2073–2079.

Amundson, J. M., Truffer, M., Lüthi, M. P., Fahnestock, M., West, M., & Motyka, R. J. (2008). Glacier, fjord, and seismic response to recent large calving events, jakobshavn isbræ, greenland. *Geophysical Research Letters*, *35*(22).

Arattano, M., Abancó, C., Coviello, V., & Hürlimann, M. (2014). Processing the ground vibration signal produced by debris flows: the methods of amplitude and impulses compared. *Computers & Geosciences*, *73*, 17–27.

Aster, R. C., & Winberry, J. P. (2017). Glacial seismology. *Reports on Progress in Physics*, *80*(12), 126801.

Atterholt, J., Zhan, Z., Shen, Z., & Li, Z. (2022). A unified wavefield-partitioning approach for distributed acoustic sensing. *Geophysical Journal International*, *228*(2), 1410–1418.

Bai, Y., Xing, J., Xie, F., Liu, S., & Li, J. (2019, December). Detection and identification of external intrusion signals from 33 km optical fiber sensing system based on deep learning. *Optical Fiber Technology*, *53*, 102060. Retrieved 2022-02-21, from https://linkinghub.elsevier.com/retrieve/pii/S106852001930077X doi: 10.1016/j.yofte.2019.102060

Bartholomaus, T., Larsen, C. F., O'Neel, S., & West, M. (2012). Calving seismicity from iceberg–sea surface interactions. *Journal of Geophysical Research: Earth Surface*, *117*(F4).

Batista, G. E., Bazzan, A. L., Monard, M. C., & et al. (2003). Balancing training data for automated annotation of keywords: a case study. *Wob*, *3*, 10–18.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29.

Battista, B. M., Knapp, C., McGee, T., & Goebel, V. (2007). Application of the empirical mode decomposition and hilbert-huang transform to seismic reflection data. *Geophysics*, *72*(2), H29–H37.

Becker, M. W., Ciervo, C., Cole, M., Coleman, T., & Mondanos, M. (2017). Fracture hydromechanical response measured by fiber optic distributed acoustic sensing at millihertz frequencies. *Geophysical Research Letters*, *44*(14), 7295–7302.

Bedard, A. (2005). Low-frequency atmospheric acoustic energy associated with vortices produced by thunderstorms. *Monthly Weather Review*, *133*(1), 241–263.

Benoit, J. P., & McNutt, S. R. (1997). New constraints on source processes of volcanic tremor at arenal volcano, costa rica, using broadband seismic data. *Geophysical Research Letters*, *24*(4), 449–452.

Bensaid, A. M., Hall, L. O., Bezdek, J. C., Clarke, L. P., Silbiger, M. L., Arrington, J. A., & Murtagh, R. F. (1996). Validity-guided (re) clustering with applications to image segmentation. *IEEE Transactions on fuzzy systems*, *4*(2), 112–123.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining* (pp. 359–370).

Bessason, B., Eiríksson, G., Thorarinsson, O., Thórarinsson, A., & Einarsson, S. (2007). Automatic detection of avalanches and debris flows by seismic methods. *J. Glaciol.*, *53*(182), 461–472. Retrieved 2022-01-10, from https://www.cambridge.org/core/product/identifier/S0022143000201172/type/journal_article doi: 10.3189/002214307783258468

Beyreuther, M., Hammer, C., Wassermann, J., Ohrnberger, M., & Megies, T. (2012). Constructing a hidden markov model based earthquake detector: application to induced seismicity. *Geophysical Journal International*, *189*(1), 602–610.

Binder, G., & Tura, A. (2020). Convolutional neural networks for automated microseismic detection in downhole distributed acoustic sensing data and comparison to a surface geophone array. *Geophysical Prospecting*, *68*(9), 2770–2782. Retrieved 2021-12-01, from https://onlinelibrary.wiley.com/doi/10.1111/1365-2478.13027 doi: 10.1111/1365-2478.13027

Bolshakova, A., Inoue, S., Kolesov, S., Matsumoto, H., Nosov, M., & Ohmachi, T. (2011). Hydroacoustic effects in the 2003 tokachi-oki tsunami source. *Russian Journal of Earth Sciences*, *12*(2), 1–14.

Booth, A. D., Christoffersen, P., Schoonman, C., Clarke, A., Hubbard, B., Law, R., … Chalari, A.

(2020). Distributed acoustic sensing of seismic properties in a borehole drilled on a fast-flowing greenlandic outlet glacier. *Geophysical Research Letters*, *47*(13), e2020GL088148.

Bordoni, P., Haines, J., Di Giulio, G., Milana, G., Augliera, P., Cercato, M., ... Team, C. E. (2007). Cavola experiment site: Geophysical investigations and deployment of a dense seismic array on a landslide. *University of Leicester*.

Brehmer, P., Gerlotto, F., Guillard, J., Sanguinède, F., Guénnegan, Y., & Buestel, D. (2003). New applications of hydroacoustic methods for monitoring shallow water aquatic ecosystems: the case of mussel culture grounds. *Aquatic Living Resources*, *16*(3), 333–338.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. Retrieved 2021-12-07, from http://link.springer.com/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Bublin, M. (2021, January). Event Detection for Distributed Acoustic Sensing: Combining Knowledge-Based, Classical Machine Learning, and Deep Learning Approaches. *Sensors*, *21*(22), 7527. Retrieved 2021-12-08, from https://www.mdpi.com/1424-8220/21/22/7527 (Number: 22 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/s21227527

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121–167.

Burtin, A., Bollinger, L., Vergne, J., Cattin, R., & Nábělek, J. (2008). Spectral analysis of seismic noise induced by rivers: A new tool to monitor spatiotemporal changes in stream hydrodynamics. *Journal of Geophysical Research: Solid Earth*, *113*(B5).

Burtin, A., Hovius, N., & Turowski, J. M. (2016). Seismic monitoring of torrential and fluvial processes. *Earth Surface Dynamics*, *4*(2), 285–307.

Cai, C., Pu, H., Wang, P., Chen, Z., & Luo, J. (2021). We hear your pace: Passive acoustic localization of multiple walking persons. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(2), 1–24.

Canassy, P. D., Faillettaz, J., Walter, F., & Huss, M. (2012). Seismic activity and surface motion of a steep temperate glacier: a study on triftgletscher, switzerland. *Journal of Glaciology*, *58*(209), 513–528.

Cao, C., Fan, X., Liu, Q., & He, Z. (2015). Practical pattern recognition system for distributed optical fiber intrusion monitoring system based on phase-sensitive coherent otdr. In *Asia communications and photonics conference* (pp. ASu2A–145).

Cao, J., Wang, W., Wang, J., & Wang, R. (2016). Excavation equipment recognition based on novel acoustic statistical features. *IEEE Transactions on Cybernetics*, *47*(12), 4392–4404.

Caudron, C., Miao, Y., Spica, Z. J., Wollin, C., Haberland, C., Jousset, P., ... et al. (2024). Monitoring underwater volcano degassing using fiber-optic sensing. *Scientific reports*, *14*(1), 3128.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.

Cesca, S. (2020). Seiscloud, a tool for density-based seismicity clustering and visualization. *Journal of Seismology*, *24*(3), 443–457.

Chakraborty, A., & Okaya, D. (1995). Frequency-time decomposition of seismic data using wavelet-based methods. *Geophysics*, *60*(6), 1906–1916.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). SMOTE: Synthetic Minority Over-sampling Technique. *jair*, *16*, 321–357. Retrieved 2022-02-11, from https://www.jair.org/index.php/jair/article/view/10302 doi: 10.1613/jair.953

Chen, J., Li, H., Ai, K., Shi, Z., Xiao, X., Yan, Z., … Sun, Q. (2024). Low-altitude uav surveillance system via highly sensitive distributed acoustic sensing. *IEEE Sensors Journal*.

Chen, S., & He, H. (2011). Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach. *Evolving Systems*, *2*(1), 35–50.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). San Francisco California USA: ACM. Retrieved 2023-07-13, from https://dl.acm.org/doi/10.1145/2939672.2939785 doi: 10.1145/2939672.2939785

Chen, Y. (2020, 06). Automatic microseismic event picking via unsupervised machine learning. *Geophysical Journal International*, *222*(3), 1750-1764. Retrieved from https://doi.org/10.1093/gji/ggaa186 doi: 10.1093/gji/ggaa186

Chen, Y., Savvaidis, A., Fomel, S., Chen, Y., Saad, O. M., Wang, H., … Chen, W. (2023). Denoising of distributed acoustic sensing seismic data using an integrated framework. *Seismological Society of America*, *94*(1), 457–472.

Chilamkuri, K., & Kone, V. (2020). Monitoring of varadhi road bridge using accelerometer sensor. *Materials Today: Proceedings*, *33*, 367–371.

Chmiel, M., Walter, F., Wenner, M., Zhang, Z., McArdell, B. W., & Hibert, C. (2021, February). Machine Learning Improves Debris Flow Warning. *Geophys Res Lett*, *48*(3). Retrieved 2022-02-02, from https://onlinelibrary.wiley.com/doi/10.1029/2020GL090874 doi: 10.1029/2020GL090874

Choi, B. H., Pelinovsky, E., Kim, K., & Lee, J. (2003). Simulation of the trans-oceanic tsunami propagation due to the 1883 krakatau volcanic eruption. *Natural Hazards and Earth System Sciences*, *3*(5), 321–332.

Collins, J., Vernon, F., Orcutt, J., Stephen, R., Peal, K., Wooding, F., … Hildebrand, J. (2001). Broadband seismology in the oceans: Lessons from the ocean seismic network pilot experiment. *Geophysical Research Letters*, *28*(1), 49–52.

Conrad, C. P., Bilek, S., & Lithgow-Bertelloni, C. (2004). Great earthquakes and slab pull: interaction between seismic coupling and plate–slab coupling. *Earth and Planetary Science Letters*, *218*(1-2), 109–122.

Cross, G. R., & Jain, A. K. (1983, January). Markov Random Field Texture Models. *IEEE Trans. Pattern Anal. Mach. Intell., PAMI-5*(1), 25–39. Retrieved 2022-01-10, from http://ieeexplore.ieee.org/document/4767341/ doi: 10.1109/TPAMI.1983.4767341

D'Alessandro, A., Luzio, D., & D'Anna, G. (2014). Urban mems based seismic network for post-earthquakes rapid disaster assessment. *Advances in Geosciences*, *40*, 1–9.

Daley, T. M., Freifeld, B. M., Ajo-Franklin, J., Dou, S., Pevzner, R., Shulakova, V., … et al. (2013). Field testing of fiber-optic distributed acoustic sensing (das) for subsurface seismic monitoring. *The Leading Edge*, *32*(6), 699–706.

Daubechies, I. (1992). Ten lectures on wavelets. *Society for industrial and applied mathematics*.

Dautermann, T., Calais, E., Lognonné, P., & Mattioli, G. S. (2009). Lithosphere—atmosphere—ionosphere coupling after the 2003 explosive eruption of the soufriere hills volcano, montserrat. *Geophysical Journal International*, *179*(3), 1537–1546.

Davy, C., Barruol, G., Fontaine, F. R., Sigloch, K., & Stutzmann, E. (2014). Tracking major storms from microseismic and hydroacoustic observations on the seafloor. *Geophysical Research Letters*, *41*(24), 8825–8831.

Dean, T., Cuny, T., & Hartog, A. H. (2017). The effect of gauge length on axially incident p-waves measured using fibre optic distributed vibration sensing. *Geophysical Prospecting*, *65*(1), 184–

193.

De Angelis, S., Fee, D., Haney, M., & Schneider, D. (2012). Detecting hidden volcanic explosions from mt. cleveland volcano, alaska with infrasound and ground-coupled airwaves. *Geophysical Research Letters*, *39*(21).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Díaz, J., DeFelipe, I., Ruiz, M., Andrés, J., Ayarza, P., & Carbonell, R. (2022). Identification of natural and anthropogenic signals in controlled source seismic experiments. *Scientific reports*, *12*(1), 3171.

Diaz, J., Ruiz, M., Udina, M., Polls, F., Martí, D., & Bech, J. (2023). Monitoring storm evolution using a high-density seismic network. *Scientific Reports*, *13*(1), 1853. Retrieved 2024-10-27, from https://www.nature.com/articles/s41598-023-28902-8 doi: 10.1038/s41598-023-28902-8

Domel, P., Hibert, C., Schlindwein, V., & Plaza-Faverola, A. (2023, May). Event recognition in marine seismological data using Random Forest machine learning classifier. *Geophysical Journal International*, *235*(1), 589–609. Retrieved 2023-09-23, from https://academic.oup.com/gji/article/235/1/589/7199654 doi: 10.1093/gji/ggad244

Dong, B., Popescu, A., Tribaldos, V. R., Byna, S., Ajo-Franklin, J., Wu, K., & et al. (2022). Real-time and post-hoc compression for data from distributed acoustic sensing. *Computers & Geosciences*, *166*, 105181.

Donnadille, M., Turquet, A., Hibert, C., & Richard, C. (2024). Distributed acoustic sensing automated classificrs dcsign via transfcr lcarning for scismology. *European Geosciences Union*.

Dou, S., Lindsey, N., Wagner, A. M., Daley, T. M., Freifeld, B., Robertson, M., … Ajo-Franklin, J. B. (2017, December). Distributed Acoustic Sensing for Seismic Monitoring of The Near Surface: A Traffic-Noise Interferometry Case Study. *Sci Rep*, *7*(1), 11620. Retrieved 2021-12-01, from http://www.nature.com/articles/s41598-017-11986-4 doi: 10.1038/s41598-017-11986-4

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in neural information processing systems*, *9*.

Dumont, V., Tribaldos, V. R., Ajo-Franklin, J., & Wu, K. (2020, December). Deep Learning for Surface Wave Identification in Distributed Acoustic Sensing Data. *2020 IEEE International Conference on Big Data (Big Data)*, 1293–1300. Retrieved 2021-11-23, from http://arxiv.org/abs/2010.10352 (arXiv: 2010.10352) doi: 10.1109/BigData50022.2020.9378084

Dziewonski, A., Bloch, S., & Landisman, M. (1969). A technique for the analysis of transient seismic signals. *Bulletin of the seismological Society of America*, *59*(1), 427–444.

Ende, M. V. D., Lior, I., Ampuero, J.-P., Sladen, A., Ferrari, A., & Richard, C. (2021). *A Self-Supervised Deep Learning Approach for Blind Denoising and Waveform Coherence Enhancement in Distributed Acoustic Sensing data.* figshare. Retrieved 2023-09-14, from https://figshare.com/articles/software/A_Self-Supervised_Deep_Learning_Approach_for_Blind_Denoising_and_Waveform_Coherence_Enhancement_in_Distributed_Acoustic_Sensing_data/14152277 (Artwork Size: 543610900 Bytes Pages: 543610900 Bytes) doi: 10.6084/M9.FIGSHARE.14152277

Engdahl, E. R., van der Hilst, R., & Buland, R. (1998). Global teleseismic earthquake relocation with

improved travel times and procedures for depth determination. *Bulletin of the Seismological Society of America*, *88*(3), 722–743.

England, P. (2018). On shear stresses, temperatures, and the maximum magnitudes of earthquakes at convergent plate boundaries. *Journal of Geophysical Research: Solid Earth*, *123*(8), 7165–7202.

Esposito, A. M., Giudicepietro, F., D'Auria, L., Scarpetta, S., Martini, M. G., Coltelli, M., & Marinaro, M. (2008). Unsupervised neural analysis of very-long-period events at stromboli volcano using the self-organizing maps. *Bulletin of the Seismological Society of America*, *98*(5), 2449–2459.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, pp. 226–231).

Falcin, A., Métaxian, J.-P., Mars, J., Stutzmann, E., Komorowski, J.-C., Moretti, R., … Lemarchand, A. (2021, March). A machine-learning approach for automatic classification of volcanic seismicity at La Soufrière Volcano, Guadeloupe. *Journal of Volcanology and Geothermal Research*, *411*, 107151. Retrieved 2023-09-23, from https://linkinghub.elsevier.com/retrieve/pii/S0377027320305874 doi: 10.1016/j.jvolgeores.2020.107151

Fang, G., Li, Y. E., Zhao, Y., & Martin, E. R. (2020). Urban near-surface seismic monitoring using distributed acoustic sensing. *Geophysical Research Letters*, *47*(6), e2019GL086115.

Fang, J., Li, Y., Ji, P. N., & Wang, T. (2022). Drone detection and localization using enhanced fiber-optic acoustic sensor and distributed acoustic sensing technology. *Journal of Lightwave Technology*, *41*(3), 822–831.

Fernandes, R. M. S., Miranda, J. M., Meijninger, B. M. L., Bos, M. S., Noomen, R., Bastos, L., … Riva, R. E. M. (2007, April). Surface velocity field of the Ibero-Maghrebian segment of the Eurasia-Nubia plate boundary. *Geophysical Journal International*, *169*(1), 315–324. Retrieved 2024-08-28, from https://academic.oup.com/gji/article-lookup/doi/10.1111/j.1365-246X.2006.03252.x doi: 10.1111/j.1365-246X.2006.03252.x

Ferrick, M., Qamar, A., & St. Lawrence, W. (1982). Source mechanism of volcanic tremor. *Journal of Geophysical Research: Solid Earth*, *87*(B10), 8675–8683.

Flanagan, M. P., & Shearer, P. M. (1998). Global mapping of topography on transition zone velocity discontinuities by stacking ss precursors. *Journal of Geophysical Research: Solid Earth*, *103*(B2), 2673–2692.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, *21*, 768–769.

Fox, C. G., & Dziak, R. P. (1998). Hydroacoustic detection of volcanic activity on the gorda ridge, february–march 1996. *Deep Sea Research Part II: Topical Studies in Oceanography*, *45*(12), 2513–2530.

Fukushima, S., Shinohara, M., Nishida, K., Takeo, A., Yamada, T., & Yomogida, K. (2022, December). Detailed S-wave velocity structure of sediment and crust off Sanriku, Japan by a new analysis method for distributed acoustic sensing data using a seafloor cable and seismic interferometry. *Earth Planets Space*, *74*(1), 92. Retrieved 2024-08-29, from https://earth-planets-space.springeropen.com/articles/10.1186/s40623-022-01652-z doi: 10.1186/s40623-022-01652-z

Gaci, S. (2013). The use of wavelet-based denoising techniques to enhance the first-arrival picking on seismic traces. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(8), 4558–4563.

Gaci, S. (2016). A new ensemble empirical mode decomposition (eemd) denoising method for seismic signals. *Energy Procedia*, *97*, 84–91.

Gariano, S. L., & Guzzetti, F. (2016). Landslides in a changing climate. *Earth-science reviews*, *162*, 227–252.

Gath, I., & Geva, A. B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, *11*(7), 773–780.

Geng, J., Bock, Y., Melgar, D., Crowell, B. W., & Haase, J. S. (2013). A new seismogeodetic approach applied to gps and accelerometer observations of the 2012 brawley seismic swarm: Implications for earthquake early warning. *Geochemistry, Geophysics, Geosystems*, *14*(7), 2124–2142.

Govi, M., Maraga, F., & Moia, F. (1993). Seismic detectors for continuous bed load monitoring in a gravel stream. *Hydrological sciences journal*, *38*(2), 123–132.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., … et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, *33*, 21271–21284.

Grollmisch, S., Abeßer, J., Liebetrau, J., & Lukashevich, H. (2019). Sounding industry: Challenges and datasets for industrial sound analysis. In *2019 27th european signal processing conference (eusipco)* (pp. 1–5).

Hammer, C., Beyreuther, M., & Ohrnberger, M. (2012). A seismic-event spotting system for volcano fast-response systems. *Bulletin of the Seismological Society of America*, *102*(3), 948–960.

Han, H., Wang, J., Meng, X., & Liu, H. (2016). Analysis of the dynamic response of a long span bridge using gps/accelerometer/anemometer under typhoon loading. *Engineering Structures*, *122*, 238–250.

Han, J., & van der Baan, M. (2013). Empirical mode decomposition for seismic time-frequency analysis. *Geophysics*, *78*(2), O9–O19.

Hansen, S. M., & Schmandt, B. (2015). Automated detection and location of microseismicity at mount st. hclcns with a largc-n gcophonc array. *Geophysical Research Letters*, *42*(18), 7390–7397.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Hartog, A. H. (2017). *An introduction to distributed optical fibre sensors*. Boca Raton: CRC Press, Taylor & Francis Group.

Havskov, J., & Alguacil, G. (2004). *Instrumentation in earthquake seismology* (Vol. 358). Springer.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1322–1328).

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18–28.

Heck, M., Hobiger, M., Van Herwijnen, A., Schweizer, J., & Fäh, D. (2019). Localization of seismic events produced by avalanches using multiple signal classification. *Geophysical Journal International*, *216*(1), 201–217.

Helmstetter, A., & Garambois, S. (2010). Seismic monitoring of séchilienne rockslide (french alps): Analysis of seismic signals and their correlation with rainfalls. *Journal of Geophysical Research: Earth Surface*, *115*(F3).

Hernández, P. D., Ramírez, J. A., & Soto, M. A. (2021). Deep-learning-based earthquake detection for fiber-optic distributed acoustic sensing. *Journal of Lightwave Technology*, *40*(8), 2639–2650.

Hibert, C., Ekström, G., & Stark, C. P. (2014). Dynamics of the bingham canyon mine landslides from seismic signal analysis. *Geophysical research letters*, *41*(13), 4535–4541.

Hibert, C., Ekström, G., & Stark, C. P. (2017). The relationship between bulk-mass momentum and short-period seismic radiation in catastrophic landslides. *Journal of Geophysical Research:*

Earth Surface, *122*(5), 1201–1215.

Hibert, C., Grandjean, G., Bitri, A., Travelletti, J., & Malet, J.-P. (2012). Characterizing landslides through geophysical data fusion: Example of the la valette landslide (france). *Engineering Geology*, *128*, 23–29.

Hibert, C., Mangeney, A., Grandjean, G., Baillard, C., Rivet, D., Shapiro, N., … Crawford, W. (2014, May). Automated identification, location, and volume estimation of rockfalls at Piton de la Fournaise volcano. *J. Geophys. Res. Earth Surf.*, *119*(5), 1082–1105. Retrieved 2021-12-01, from http://doi.wiley.com/10.1002/2013JF002970 doi: 10.1002/2013JF002970

Hibert, C., Michéa, D., Provost, F., Malet, J.-P., & Geertsema, M. (2019, November). Exploration of continuous seismic recordings with a machine learning approach to document 20 yr of landslide activity in Alaska. *Geophysical Journal International*, *219*(2), 1138–1147. Retrieved 2021-12-01, from https://academic.oup.com/gji/article/219/2/1138/5542200 doi: 10.1093/gji/ggz354

Hibert, C., Provost, F., Malet, J.-P., Maggi, A., Stumpf, A., & Ferrazzini, V. (2017, June). Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest algorithm. *Journal of Volcanology and Geothermal Research*, *340*, 130–142. Retrieved 2021-12-01, from https://linkinghub.elsevier.com/retrieve/pii/S0377027316303948 doi: 10.1016/j.jvolgeores.2017.04.015

Hicke, K., Hussels, M.-T., Eisermann, R., Chruscicki, S., & Krebber, K. (2017a). Condition monitoring of industrial infrastructures using distributed fibre optic acoustic sensors. In Y. Chung, W. Jin, B. Lee, J. Canning, K. Nakamura, & L. Yuan (Eds.), (p. 103230J). Retrieved 2021-12-08, from http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2272463 doi: 10.1117/12.2272463

Hicke, K., Hussels, M.-T., Eisermann, R., Chruscicki, S., & Krebber, K. (2017b, April). Condition monitoring of industrial infrastructures using distributed fibre optic acoustic sensors. In Y. Chung, W. Jin, B. Lee, J. Canning, K. Nakamura, & L. Yuan (Eds.), (p. 103230J). Jeju, Korea, Republic of. Retrieved 2021-12-08, from http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2272463 doi: 10.1117/12.2272463

Hildebrand, J. A. (2009). Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series*, *395*, 5–20.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832–844.

Hoover, G., & O'Brien, J. (1980). The influence of the planted geophone on seismic land data. *Geophysics*, *45*(8), 1239–1253.

Horasan, G., Güney, A. B., Küsmezer, A., Bekler, F., Öğütçü, Z., & Musaoğlu, N. (2009). Contamination of seismicity catalogs by quarry blasts: an example from istanbul and its vicinity, northwestern turkey. *Journal of Asian Earth Sciences*, *34*(1), 90–99.

Hou, Y., Jiao, R., & Yu, H. (2021). Mems based geophones and seismometers. *Sensors and Actuators A: Physical*, *318*, 112498.

Houston, K. M., & McGaffigan, D. P. (2003). Spectrum analysis techniques for personnel detection using seismic sensors. In *Unattended ground sensor technologies and applications v* (Vol. 5090, pp. 162–173).

Hu, M., & Li, Z. (2024). Daspy: A python toolbox for das seismology. *Seismological Research Letters*, *95*(5), 3055–3066.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., … Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, *454*(1971), 903–995.

Hudson, T. S., Baird, A. F., Kendall, J. M., Kufner, S. K., Brisbourne, A. M., Smith, A. M., … Clarke, A. (2021, July). Distributed Acoustic Sensing (DAS) for Natural Microseismicity Studies: A Case Study From Antarctica. *JGR Solid Earth*, *126*(7), e2020JB021493. Retrieved 2024-08-29, from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020JB021493 doi: 10.1029/2020JB021493

Hungr, O., Corominas, J., & Eberhardt, E. (2005). Estimating landslide motion mechanism, travel distance and velocity. In *Landslide risk management* (pp. 109–138). CRC Press.

Huot, F., Martin, E. R., & Biondi, B. (2018, October). Automated ambient noise processing applied to fiber optic seismic acquisition (DAS). OnePetro. Retrieved 2021-12-07, from https://onepetro.org/SEGAM/proceedings/SEG18/All-SEG18/SEG-2018-2997880/103704 doi: 10.1190/segam2018-2997880.1

Hürlimann, M., Abancó, C., Moya, J., & Vilajosana, I. (2014). Results and experiences gathered at the rebaixader debris-flow monitoring site, central pyrenees, spain. *Landslides*, *11*, 939–953.

Huynh, C., Hibert, C., Jestin, C., Malet, J., & Lanticq, V. (2024). A real scale application of a novel set of spatial and similarity features for detection and classification of natural seismic sources from distributed acoustic sensing data. *Geophysical Journal International*, ggae382.

Huynh, C., Hibert, C., Jestin, C., Malet, J.-P., Clément, P., & Lanticq, V. (2022, September). Real-Time Classification of Anthropogenic Seismic Sources from Distributed Acoustic Sensing Data: Application for Pipeline Monitoring. *Seismological Research Letters*, *93*(5), 2570–2583. Retrieved 2023-07-03, from https://pubs.geoscienceworld.org/srl/article/93/5/2570/615834/Real-Time-Classification-of-Anthropogenic-Seismic doi: 10.1785/0220220078

Ida, Y., Fujita, E., & Hirose, T. (2022). Classification of volcano-seismic events using waveforms in the method of k-means clustering and dynamic time warping. *Journal of Volcanology and Geothermal Research*, *429*, 107616.

Iverson, R. M., Reid, M. E., & LaHusen, R. G. (1997). Debris-flow mobilization from landslides. *Annual Review of Earth and Planetary Sciences*, *25*(1), 85–138.

Jakkampudi, S., Shen, J., Li, W., Dev, A., Zhu, T., & Martin, E. R. (2020). Footstep detection in urban seismic data with a convolutional neural network. *The Leading Edge*, 7. doi: 10.1190/tle39090654.1

Johannessen, K., Drakeley, B., & Farhadiroushan, M. (2012). Distributed acoustic sensing-a new way of listening to your well/reservoir. In *Spe intelligent energy international conference and exhibition* (pp. SPE–149602).

Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes of seismic noise signals using unsupervised learning. *Geophysical Research Letters*, *47*(15), e2020GL088353.

Johnson, J., Aster, R., Ruiz, M., Malone, S., McChesney, P., Lees, J., & Kyle, P. (2003). Interpretation and utility of infrasonic records from erupting volcanoes. *Journal of volcanology and geothermal research*, *121*(1-2), 15–63.

Johnson, J. B., & Aster, R. (2005). Relative partitioning of acoustic and seismic energy during strom-

bolian eruptions. *Journal of Volcanology and Geothermal Research*, *148*(3-4), 334–354.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Jousset, P., Currenti, G., Schwarz, B., Chalari, A., Tilmann, F., Reinsch, T., … Krawczyk, C. M. (2022). Fibre optic distributed acoustic sensing of volcanic events. *Nature Communications*, *13*(1), 1753. Retrieved 2023-08-02, from https://www.nature.com/articles/s41467-022-29184-w doi: 10.1038/s41467-022-29184-w

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Kagan, Y. (1992). Correlations of earthquake focal mechanisms. *Geophysical Journal International*, *110*(2), 305–320.

Kaneko, L., Ide, S., & Nakano, M. (2018). Slow earthquakes in the microseism frequency band (0.1–1.0 hz) off kii peninsula, japan. *Geophysical Research Letters*, *45*(6), 2618–2624.

Kappus, M. E., & Vernon, F. L. (1991). Acoustic signature of thunder from seismic records. *Journal of Geophysical Research: Atmospheres*, *96*(D6), 10989–11006.

Karrenbach, M., Cole, S., Ridge, A., Boone, K., Kahn, D., Rich, J., … Langton, D. (2019). Fiber-optic distributed acoustic sensing of microseismicity, strain and temperature during hydraulic fracturing. *Geophysics*, *84*(1), D11–D23.

Kavousi, P., Carr, T., Wilson, T., Amini, S., Wilson, C., Thomas, M., … et al. (2017). Correlating distributed acoustic sensing (das) to natural fracture intensity for the marcellus shale. In *Seg technical program expanded abstracts 2017* (pp. 5386–5390). Society of Exploration Geophysicists.

Kawakatsu, H., Kaneshima, S., Matsubayashi, H., Ohminato, T., Sudo, Y., Tsutsui, T., … Legrand, D. (2000). Aso94: Aso seismic observation with broadband instruments. *Journal of Volcanology and Geothermal Research*, *101*(1-2), 129–154.

Keefer, D. K. (1984). Landslides caused by earthquakes. *Geological Society of America Bulletin*, *95*(4), 406–421.

Keefer, D. K. (2002). Investigating landslides caused by earthquakes–a historical review. *Surveys in geophysics*, *23*, 473–510.

Keogh, E., & Ratanamahatana, C. A. (2005, March). Exact indexing of dynamic time warping. *Knowl Inf Syst*, *7*(3), 358–386. Retrieved 2024-08-30, from http://link.springer.com/10.1007/s10115-004-0154-9 doi: 10.1007/s10115-004-0154-9

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*(6), 4909–4926.

Klaasen, S., Paitz, P., Lindner, N., Dettmer, J., & Fichtner, A. (2021). Distributed acoustic sensing in volcano-glacial environments—mount meager, british columbia. *Journal of Geophysical Research: Solid Earth*, *126*(11), e2021JB022358.

Kleine, F., Bruland, C., Wüstefeld, A., Oye, V., & Landrø, M. (2024). Seismic signal classification of snow avalanches using distributed acoustic sensing in grasdalen, western norway. *Natural Hazards and Earth System Sciences Discussions*, *2024*, 1–22.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238–1274.

Köhler, A., Maupin, V., Nuth, C., & Van Pelt, W. (2019). Characterization of seasonal glacial seismicity from a single-station on-ice record at holtedahlfonna, svalbard. *Annals of Glaciology*, *60*(79), 23–36.

Kowarik, S., Hussels, M.-T., Chruscicki, S., Münzenberger, S., Lämmerhirt, A., Pohl, P., & Schubert, M. (2020, January). Fiber Optic Train Monitoring with Distributed Acoustic Sensing: Conventional and Neural Network Data Analysis. *Sensors*, *20*(2), 450. Retrieved 2021-12-08, from https://www.mdpi.com/1424-8220/20/2/450 (Number: 2 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/s20020450

Koyanagi, R. Y., Chouet, B., & Aki, K. (1987). Origin of volcanic tremor in hawaii. *US Geological Survey Professional Paper*, *1350*(2), 1221–1257.

Kriegel, H.-P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(3), 231–240.

Krieger, K. J., & Wing, B. L. (1986). Hydroacoustic monitoring of prey to determine humpback whale movements. *National Oceanic and Atmospheric Administration*.

Krohn, C. E. (1984). Geophone ground coupling. *Geophysics*, *49*(6), 722–731.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, *6*, 1–15.

Kruber, F., Wurst, J., & Botsch, M. (2018). An unsupervised random forest clustering technique for automatic traffic scenario categorization. In *2018 21st international conference on intelligent transportation systems (itsc)* (pp. 2811–2818).

Kumar, S., Vig, R., & Kapur, P. (2018). Development of earthquake event detection technique based on sta/lta algorithm for seismic alert system. *Journal of the Geological Society of India*, *92*, 679–686.

Kumar, U., Legendre, C. P., Lee, J.-C., Zhao, L., & Chao, B. F. (2023). On analyzing gnss displacement field variability of taiwan: Hierarchical agglomerative clustering based on dynamic time warping technique. *Computers & Geosciences*, *169*, 105243.

Kumar, U., Legendre, C. P., Zhao, L., & Chao, B. F. (2022). Dynamic time warping as an alternative to windowed cross correlation in seismological applications. *Seismological Society of America*, *93*(3), 1909–1921.

Kuvshinov, B. (2016). Interaction of helically wound fibre-optic cables with plane seismic waves. *Geophysical Prospecting*, *64*(3), 671–688.

Lacan, P., & Ortuño, M. (2012, September). Active Tectonics of the Pyrenees: A review. *Journal of Iberian Geology*, *38*(1), 9–30. Retrieved 2023-10-02, from http://revistas.ucm.es/index.php/JIGE/article/view/39203 doi: 10.5209/rev_JIGE.2012.v38.n1.39203

Lahr, J. C., Chouet, B. A., Stephens, C. D., Power, J. A., & Page, R. A. (1994). Earthquake classification, location, and error analysis in a volcanic environment: Implications for the magmatic system of the 1989–1990 eruptions at redoubt volcano, alaska. *Journal of Volcanology and Geothermal Research*, *62*(1-4), 137–151.

Lai, V. H., Hodgkinson, K. M., Porritt, R. W., & Mellors, R. (2024). Toward a metadata standard for distributed acoustic sensing (das) data collection. *Seismological Research Letters*, *95*(3), 1986–1999.

Landrø, M., Bouffaut, L., Kriesell, H. J., Potter, J. R., Rørstadbotnen, R. A., Taweesintananon, K., … et al. (2022). Sensing whales, storms, ships and earthquakes using an arctic fibre optic cable. *Scientific Reports*, *12*(1), 19226.

Lapins, S., Butcher, A., Kendall, J.-M., Hudson, T. S., Stork, A. L., Werner, M. J., … Brisbourne, A. M. (2024). Das-n2n: machine learning distributed acoustic sensing (das) signal denoising without clean data. *Geophysical Journal International*, *236*(2), 1026–1041.

Lapins, S., Goitom, B., Kendall, J., Werner, M. J., Cashman, K. V., & Hammond, J. O. S. (2021, July). A Little Data Goes a Long Way: Automating Seismic Phase Arrival Picking at Nabro Volcano With Transfer Learning. *JGR Solid Earth*, *126*(7), e2021JB021910. Retrieved 2024-03-29, from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021JB021910 doi: 10.1029/2021JB021910

Lapins, S., Roman, D. C., Rougier, J., De Angelis, S., Cashman, K. V., & Kendall, J.-M. (2020). An examination of the continuous wavelet transform for volcano-seismic spectral analysis. *Journal of Volcanology and Geothermal Research*, *389*, 106728.

Leary, P. (1997). Rock as a critical-point system and the inherent implausibility of reliable earthquake prediction. *Geophysical Journal International*, *131*(3), 451–466.

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

Lee, L., Lawrence, D., & Price, M. (2006). Analysis of water-level response to rainfall and implications for recharge pathways in the chalk aquifer, se england. *Journal of hydrology*, *330*(3-4), 604–620.

Lellouch, A., & et al. (2019). Velocity-based earthquake detection using downhole distributed acoustic sensing—examples from the san andreas fault observatory at depth. *Bulletin of the Seismological Society of America*, *109*(6), 2491–2500.

Lellouch, A., Lindsey, N. J., Ellsworth, W. L., & Biondi, B. L. (2020, November). Comparison between Distributed Acoustic Sensing and Geophones: Downhole Microseismic Monitoring of the FORGE Geothermal Experiment. *Seismological Research Letters*, *91*(6), 3256–3268. Retrieved 2022-05-31, from https://pubs.geoscienceworld.org/ssa/srl/article/91/6/3256/590004/Comparison-between-Distributed-Acoustic-Sensing doi: 10.1785/0220200149

Lellouch, A., Yuan, S., Spica, Z., Biondi, B., & Ellsworth, W. L. (2019). Seismic velocity estimation using passive downhole distributed acoustic sensing records: Examples from the san andreas fault observatory at depth. *Journal of Geophysical Research: Solid Earth*, *124*(7), 6931–6948.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, *49*(4), 764–766.

Li, X., Shang, X., Wang, Z., Dong, L., & Weng, L. (2016). Identifying p-phase arrivals with noise: An improved kurtosis method based on dwt and sta/lta. *Journal of Applied Geophysics*, *133*, 50–61.

Li, Y., Cui, Y., Hu, X., Lu, Z., Wang, Y., Wang, H., … Zhou, X. (2024, 04). Glacier retreat in eastern himalaya drives catastrophic glacier hazard chain. *Geophysical Research Letters*, *51*. doi: 10.1029/2024GL108202

Li, Z., & et al. (2020, February). Fiber distributed acoustic sensing using convolutional long short-term memory network: a field test on high-speed railway intrusion detection. *Opt. Express*, *28*(3), 2925. Retrieved 2021-11-10, from https://www.osapublishing.org/abstract.cfm?URI=oe-28-3-2925 doi: 10.1364/OE.28.002925

Li, Z., Shen, Z., Yang, Y., Williams, E., Wang, X., & Zhan, Z. (2021). Rapid response to the 2019 ridgecrest earthquake with distributed acoustic sensing. *AGU Advances*, *2*(2), e2021AV000395.

Li, Z., Zhang, H., Singh, V. P., Yu, R., & Zhang, S. (2019). A simple early warning system for flash floods in an ungauged catchment and application in the loess plateau, china. *Water*, *11*(3),

426.

Li, Z., Zhang, J., Wang, M., Chai, J., Wu, Y., & Peng, F. (2020, June). An anti-noise phi-OTDR based distributed acoustic sensing system for high-speed railway intrusion detection. *Laser Physics*, *30*(8), 085103. Retrieved from https://doi.org/10.1088/1555-6611/ab9119 (Publisher: IOP Publishing) doi: 10.1088/1555-6611/ab9119

Lienkaemper, J. J., Baker, B., & McFarland, F. S. (2006). Surface slip associated with the 2004 parkfield, california, earthquake measured on alinement arrays. *Bulletin of the Seismological Society of America*, *96*(4B), S239–S249.

Lin, J., Fang, S., He, R., Tang, Q., Qu, F., Wang, B., & Xu, W. (2024). Monitoring ocean currents during the passage of typhoon muifa using optical-fiber distributed acoustic sensing. *Nature Communications*, *15*(1), 1111. Retrieved 2024-10-31, from https://www.nature.com/articles/s41467-024-45412-x doi: 10.1038/s41467-024-45412-x

Lindsey, N. J., Dawe, T. C., & Ajo-Franklin, J. B. (2019, November). Illuminating seafloor faults and ocean dynamics with dark fiber distributed acoustic sensing. *Science*, *366*(6469), 1103–1107. Retrieved 2021-12-01, from https://www.science.org/doi/10.1126/science.aay5881 doi: 10.1126/science.aay5881

Lindsey, N. J., Martin, E. R., Dreger, D. S., Freifeld, B., Cole, S., James, S. R., … Ajo-Franklin, J. B. (2017, December). Fiber-Optic Network Observations of Earthquake Wavefields. *Geophys. Res. Lett.*, *44*(23). Retrieved 2021-12-01, from https://onlinelibrary.wiley.com/doi/10.1002/2017GL075722 doi: 10.1002/2017GL075722

Lior, I., Rivet, D., Ampuero, J.-P., Sladen, A., Barrientos, S., Sánchez-Olavarría, R., … Bustamante Prado, J. A. (2023). Magnitude estimation and ground motion prediction to harness fiber optic distributed acoustic sensing for earthquake early warning. *Scientific Reports*, *13*(1), 424. Retrieved 2024-11-10, from https://www.nature.com/articles/s41598-023-27444-3 doi: 10.1038/s41598-023-27444-3

Lior, I., Sladen, A., Rivet, D., Ampuero, J.-P., Hello, Y., Becerril, C., … et al. (2021). On the detection capabilities of underwater distributed acoustic sensing. *Journal of Geophysical Research: Solid Earth*, *126*(3), e2020JB020925.

Lipovsky, B. P., Meyer, C. R., Zoet, L. K., McCarthy, C., Hansen, D. D., Rempel, A. W., & Gimbert, F. (2019). Glacier sliding, seismicity and sediment entrainment. *Annals of Glaciology*, *60*(79), 182–192.

Liu, H., Ma, J., Xu, T., Yan, W., Ma, L., & Zhang, X. (2020). Vehicle detection and classification using distributed fiber optic acoustic sensing. *IEEE Transactions on Vehicular Technology*, *69*(2), 1363–1374. Retrieved 2021-12-01, from https://ieeexplore.ieee.org/document/8943448/ doi: 10.1109/TVT.2019.2962334

Liu, W., Cao, S., & Chen, Y. (2015). Seismic time–frequency analysis via empirical wavelet transform. *IEEE Geoscience and Remote Sensing Letters*, *13*(1), 28–32.

Lockner, D. A., & Beeler, N. M. (2002). Rock failure and earthquakes. In *International geophysics* (Vol. 81, pp. 505–537). Elsevier.

Lu, P., Qin, Y., Li, Z., Mondini, A. C., & Casagli, N. (2019, September). Landslide mapping from multi-sensor data through improved change detection-based Markov random field. *Remote Sensing of Environment*, *231*, 111235. Retrieved 2022-01-10, from https://linkinghub.elsevier.com/retrieve/pii/S0034425719302548 doi: 10.1016/j.rse.2019.111235

Ma, Y., & Hale, D. (2013). Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion. *Geophysics*, *78*(6), R223–R233.

MacAyeal, D. R., Okal, E. A., Aster, R. C., & Bassis, J. (2008). Seismic and hydroacoustic tremor generated by colliding icebergs. *Journal of Geophysical Research: Earth Surface*, *113*(F3).

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017, May). Implementation of a Multistation Approach for Automated Event Classification at Piton de la Fournaise Volcano. *Seismological Research Letters*, *88*(3), 878–891. Retrieved 2021-12-01, from https://pubs.geoscienceworld.org/srl/article/88/3/878-891/284054 doi: 10.1785/0220160189

Mahjoubi, S., Barhemat, R., & Bao, Y. (2020). Optimal placement of triaxial accelerometers using hypotrochoid spiral optimization algorithm for automated monitoring of high-rise buildings. *Automation in Construction*, *118*, 103273.

Malfante, M., Dalla Mura, M., Metaxian, J.-P., Mars, J. I., Macedo, O., & Inza, A. (2018, March). Machine Learning for Volcano-Seismic Signals: Challenges and Perspectives. *IEEE Signal Process. Mag.*, *35*(2), 20–30. Retrieved 2023-10-02, from http://ieeexplore.ieee.org/document/8310698/ doi: 10.1109/MSP.2017.2779166

Mani, I., & Zhang, I. (2003). knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets* (Vol. 126, pp. 1–7).

Martin, E. R., Huot, F., Ma, Y., Cieplicki, R., Cole, S., Karrenbach, M., & Biondi, B. L. (2018, March). A Seismic Shift in Scalable Acquisition Demands New Processing: Fiber-Optic Seismic Signal Retrieval in Urban Areas with Unsupervised Learning for Coherent Noise Removal. *IEEE Signal Processing Magazine*, *35*(2), 31–40. (Conference Name: IEEE Signal Processing Magazine) doi: 10.1109/MSP.2017.2783381

Martinez, K., Hart, J. K., Basford, P. J., Bragg, G. M., Ward, T., & Young, D. S. (2017). A geophone wireless sensor network for investigating glacier stick-slip motion. *Computers & Geosciences*, *105*, 103–112.

Matsumoto, H., Araki, E., Kimura, T., Fujie, G., Shiraishi, K., Tonegawa, T., … others (2021). Detection of hydroacoustic signals on a fiber-optic submarine cable. *Scientific reports*, *11*(1), 2797.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McNutt, S. R. (2005). Volcanic seismology. *Annu. Rev. Earth Planet. Sci.*, *33*(1), 461–491.

McNutt, S. R., & Roman, D. C. (2015). Chapter 59 - volcanic seismicity. In H. Sigurdsson (Ed.), *The encyclopedia of volcanoes (second edition)* (Second Edition ed., p. 1011-1034). Amsterdam: Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780123859389000596 doi: https://doi.org/10.1016/B978-0-12-385938-9.00059-6

Meinshausen, N., & Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, *7*(6).

Mejías, M., Ballesteros, B. J., Antón-Pacheco, C., Domínguez, J. A., Garcia-Orellana, J., Garcia-Solsona, E., & Masqué, P. (2012). Methodological study of submarine groundwater discharge from a karstic aquifer in the western mediterranean sea. *Journal of Hydrology*, *464*, 27–40.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, *28*, 92–122.

Meng, X., Dodson, A., & Roberts, G. W. (2007). Detecting bridge dynamics with gps and triaxial accelerometers. *Engineering Structures*, *29*(11), 3178–3184.

Métaxian, J.-P., Araujo, S., Mora, M., & Lesage, P. (2003). Seismicity related to the glacier of cotopaxi volcano, ecuador. *Geophysical Research Letters*, *30*(9).

Métaxian, J.-P., Biagioli, F., Trabattoni, A., Stutzmann, E., Lacanna, G., Risica, G., … others (2024). Using distributed acoustic sensing, seismic and infrasonic observation to track pyroclastic flows at stromboli volcano (italy). *European Geosciences Union*.

Meyer, Y. (1992). *Wavelets and operators: volume 1* (No. 37). Cambridge university press.

Michlmayr, G., Chalari, A., Clarke, A., & Or, D. (2017). Fiber-optic high-resolution acoustic emission (ae) monitoring of slope failure. *Landslides*, *14*(3), 1139–1146.

Minakami, T. (1974). Seismology of volcanoes in japan. In *Developments in solid earth geophysics* (Vol. 6, pp. 1–27). Elsevier.

Mita, A., & Yokoi, I. (2001). Fiber bragg grating accelerometer for buildings and civil infrastructures. In *Smart structures and materials 2001: Smart systems for bridges, structures, and highways* (Vol. 4330, pp. 479–486).

Mnih, V. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Molson, J., & Frind, E. (2012). On the use of mean groundwater age, life expectancy and capture probability for defining aquifer vulnerability and time-of-travel zones for source water protection. *Journal of Contaminant Hydrology*, *127*(1-4), 76–87.

Moschas, F., & Stiros, S. (2011). Measurement of the dynamic displacements and of the modal frequencies of a short-span pedestrian bridge using gps and an accelerometer. *Engineering structures*, *33*(1), 10–17.

Mosleh, A., Montenegro, P., Alves Costa, P., & Calçada, R. (2021). An approach for wheel flat detection of railway train wheels using envelope spectrum analysis. *Structure and Infrastructure Engineering*, *17*(12), 1710–1729.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phasc picking. *Nature communications*, *11*(1), 3952.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, *31*, 274–295.

Nayak, A., Ajo-Franklin, J., & the Imperial Valley Dark Fiber Team. (2021, July). Distributed Acoustic Sensing Using Dark Fiber for Array Detection of Regional Earthquakes. *Seismological Research Letters*, *92*(4), 2441–2452. Retrieved 2022-05-31, from https://pubs.geoscienceworld.org/srl/article/92/4/2441/595405/Distributed-Acoustic-Sensing-Using-Dark-Fiber-for doi: 10.1785/0220200416

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2021). Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 international joint conference on neural networks (ijcnn)* (pp. 1–8).

Ning, I. L. C., & Sava, P. (2018). High-resolution multi-component distributed acoustic sensing. *Geophysical Prospecting*, *66*(6), 1111–1122.

Nishida, Y., Ura, T., Sakamaki, T., Kojima, J., Ito, Y., & Kim, K. (2013). Hovering type auv "tuna-sand" and its surveys on smith caldera in izu-ogasawara ocean area. In *2013 oceans-san diego* (pp. 1–5).

Nishikawa, T., & Ide, S. (2014). Earthquake size distribution in subduction zones linked to slab buoyancy. *Nature Geoscience*, *7*(12), 904–908.

Nishimura, T., Emoto, K., Nakahara, H., Miura, S., Yamamoto, M., Sugimura, S., … Kimura, T. (2021). Source location of volcanic earthquakes and subsurface characterization using fiber-optic cable and distributed acoustic sensing system. *Scientific Reports*, *11*(1), 6319. Retrieved 2022-02-

21, from http://www.nature.com/articles/s41598-021-85621-8 doi: 10.1038/s41598-021-85621-8

Nziengui-Bâ, D., Coutant, O., Moreau, L., & Boué, P. (2023). Measuring the thickness and young's modulus of the ice pack with das, a test case on a frozen mountain lake. *Geophysical Journal International*, *233*(2), 1166–1177.

Olson, E. L., & Allen, R. M. (2005). The deterministic nature of earthquake rupture. *Nature*, *438*(7065), 212–215.

Ouellet, S. M., Dettmer, J., Lato, M. J., Cole, S., Hutchinson, D. J., Karrenbach, M., … Crickmore, R. (2024). Previously hidden landslide processes revealed using distributed acoustic sensing with nanostrain-rate sensitivity. *Nature Communications*, *15*(1), 6239. Retrieved 2024-11-11, from https://www.nature.com/articles/s41467-024-50604-6 doi: 10.1038/s41467-024-50604-6

Paitz, P., Lindner, N., Edme, P., Huguenin, P., Hohl, M., Sovilla, B., … Fichtner, A. (2023). Phenomenology of avalanche recordings from distributed acoustic sensings. *Journal of Geophysical Research: Earth Surface*, e2022JF007011.

Paleja, R., Mustafina, D., in 't panhuis, P., Park, T., Randell, D., van der Horst, J., & Crickmore, R. (2015). Velocity tracking for flow monitoring and production profiling using distributed acoustic sensing. *SPE Annual Technical Conference and Exhibition?*, D021S020R001.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

Papp, A., Wiesmeyr, C., Litzenberger, M., Garn, H., & Kropatsch, W. (2016). A real-time algorithm for train position monitoring using optical time-domain reflectometry. In *2016 ieee international conference on intelligent rail transportation (icirt)* (pp. 89–93).

Parker, T., Shatalin, S., & Farhadiroushan, M. (2014). Distributed acoustic sensing–a new tool for seismic applications. *first break*, *32*(2).

Patanè, D., Tusa, G., Yang, W., Astuti, A., Colino, A., Costanza, A., … et al. (2022). The urban seismic observatory of catania (italy): a real-time seismic monitoring at urban scale. *Remote Sensing*, *14*(11), 2583.

Pechmann, J. C., Nava, S. J., Terra, F. M., & Bernier, J. C. (2007). Local magnitude determinations for intermountain seismic belt earthquakes from broadband digital data. *Bulletin of the Seismological Society of America*, *97*(2), 557–574.

Peng, F., Duan, N., Rao, Y.-J., & Li, J. (2014). Real-time position and speed monitoring of trains using phase-sensitive otdr. *IEEE Photonics Technology Letters*, *26*(20), 2055–2057.

Peng, Z., Jian, J., Wen, H., Gribok, A., Wang, M., Liu, H., … Chen, K. P. (2020, September). Distributed fiber sensor and machine learning data analytics for pipeline protection against extrinsic intrusions and intrinsic corrosions. *Opt. Express*, *28*(19), 27277. Retrieved 2021-12-01, from https://www.osapublishing.org/abstract.cfm?URI=oe-28-19-27277 doi: 10.1364/OE.397509

Pérez, N., Granda, F. S., Benítez, D., Grijalva, F., & Lara, R. (2021). Toward real-time volcano seismic events' classification: A new approach using mathematical morphology and similarity criteria. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13.

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Provost, F., Hibert, C., & Malet, J.-P. (2017, January). Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier: SEISMIC SOURCES AUTOMATIC CLASSIFICATION. *Geophys. Res. Lett.*, *44*(1), 113–120. Retrieved 2021-

12-01, from http://doi.wiley.com/10.1002/2016GL070709 doi: 10.1002/2016GL070709

Provost, F., Malet, J.-P., Hibert, C., Helmstetter, A., Radiguet, M., Amitrano, D., ... others (2018). Towards a standard typology of endogenous landslide seismic sources. *Earth Surface Dynamics*, *6*(4), 1059–1088.

Prugger, A. F., & Gendzwill, D. J. (1988). Microearthquake location: A nonlinear approach that makes use of a simplex stepping procedure. *Bulletin of the Seismological Society of America*, *78*(2), 799–815.

Qamar, A. (1988). Calving icebergs: A source of low-frequency seismic signals from columbia glacier, alaska. *Journal of Geophysical Research: Solid Earth*, *93*(B6), 6615–6623.

Ran, Q., Hong, Y., Li, W., & Gao, J. (2018). A modelling study of rainfall-induced shallow landslide mechanisms under different rainfall characteristics. *Journal of Hydrology*, *563*, 790–801.

Recasens, A., Luc, P., Alayrac, J.-B., Wang, L., Strub, F., Tallec, C., ... et al. (2021). Broaden your views for self-supervised video learning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1255–1265).

Rice, J. R. (1979). *The mechanics of earthquake rupture* (Vol. 720). Division of Engineering, Brown University Providence.

Richardson, J. P., Waite, G. P., FitzGerald, K. A., & Pennington, W. D. (2010). Characteristics of seismic and acoustic signals produced by calving, bering glacier, alaska. *Geophysical Research Letters*, *37*(3).

Rigo, A., Souriau, A., Dubos, N., Sylvander, M., & Ponsolles, C. (2005, April). Analysis of the seismicity in the central part of the Pyrenees (France), and tectonic implications. *J Seismol*, *9*(2), 211–222. Retrieved 2023-10-02, from http://link.springer.com/10.1007/s10950-005-2775-1 doi: 10.1007/s10950-005-2775-1

Rilling, G., Flandrin, P., Goncalves, P., & et al. (2003). On empirical mode decomposition and its algorithms. In *Ieee-eurasip workshop on nonlinear signal and image processing* (Vol. 3, pp. 8–11).

Rimpot, J., Hibert, C., Malet, J.-P., Forestier, G., & Weber, J. (2024). Self-supervised learning strategies for clustering continuous seismic data. *European Geosciences Union*.

Rimpot, J., Hibert, C., Retailleau, L., Saurel, J.-M., Malet, J.-P., Forestier, G., ... Pelleau, P. (2024). Self-supervised learning of seismological data reveals new eruptive sequences at the mayotte submarine volcano. *Geophysical Journal International*, *240*(1), 1–12.

Roberts, R., Christoffersson, A., & Cassidy, F. (1989). Real-time event detection, phase identification and source location estimation using single station three-component seismic data. *Geophysical Journal International*, *97*(3), 471–480.

Romano, F., Trasatti, E., Lorito, S., Piromallo, C., Piatanesi, A., Ito, Y., ... Cocco, M. (2014). Structural control on the tohoku earthquake rupture process investigated by 3d fem, tsunami and geodetic data. *Scientific Reports*, *4*(1), 5631.

Röösli, C., Walter, F., Husen, S., Andrews, L. C., Lüthi, M. P., Catania, G. A., & Kissling, E. (2014). Sustained seismic tremors and icequakes detected in the ablation zone of the greenland ice sheet. *Journal of Glaciology*, *60*(221), 563–575.

Ross, Z. E., & Ben-Zion, Y. (2014, 08). Automatic picking of direct p, s seismic phases and fault zone head waves. *Geophysical Journal International*, *199*(1), 368-381. Retrieved from https://doi.org/10.1093/gji/ggu267 doi: 10.1093/gji/ggu267

Roux, P.-F., Marsan, D., Métaxian, J.-P., O'Brien, G., & Moreau, L. (2008). Microseismic activity within a serac zone in an alpine glacier (glacier d'argentiere, mont blanc, france). *Journal of*

Glaciology, *54*(184), 157–168.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43-49. doi: 10.1109/TASSP.1978.1163055

Santana, J., Van den Hoven, R., Van Liempd, C., Colin, M., Saillen, N., Zonta, D., … Van Hoof, C. (2012). A 3-axis accelerometer and strain sensor system for building integrity monitoring. *Sensors and Actuators A: Physical*, *188*, 141–147.

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Scarpa, R., Tilling, R. I., & McNutt, S. (1996). Seismic monitoring and eruption forecasting of volcanoes: a review of the state-of-the-art and case histories. *Monitoring and mitigation of volcano hazards*, 99–146.

Schellart, W. P., & Rawlinson, N. (2013). Global correlations between maximum magnitudes of subduction zone interface thrust earthquakes and physical parameters of subduction zones. *Physics of the Earth and Planetary Interiors*, *225*, 41–67.

Schimmel, M., & Paulssen, H. (1997). Noise reduction and detection of weak, coherent signals through phase-weighted stacks. *Geophysical Journal International*, *130*(2), 497–505.

Schmandt, B., Aster, R. C., Scherler, D., Tsai, V. C., & Karlstrom, K. (2013). Multiple fluvial processes detected by riverside seismic and infrasound monitoring of a controlled flood in the grand canyon. *Geophysical Research Letters*, *40*(18), 4858–4863.

Schmedes, J., Archuleta, R. J., & Lavallée, D. (2010). Correlation of earthquake source parameters inferred from dynamic rupture simulations. *Journal of Geophysical Research: Solid Earth*, *115*(B3).

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Scholz, C. H. (2002). *The mechanics of earthquakes and faulting* (2. ed. ed.). Cambridge Univ. Press.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.

Segall, P. (2013). Volcano deformation and eruption forecasting. *Geological Society, London, Special Publications*, *380*(1), 85–106.

Sergeant, A., Chmiel, M., Lindner, F., Walter, F., Roux, P., Chaput, J., … Mordret, A. (2020). On the green's function emergence from interferometry of seismic wave fields generated in high-melt glaciers: implications for passive imaging and monitoring. *The Cryosphere*, *14*(3), 1139–1171.

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. d., Campillo, M., & Baraniuk, R. (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature communications*, *11*(1), 3972.

Shaheen, S., Hicke, K., & Krebber, K. (2023). Blast-assisted subsurface characterisation using a novel distributed acoustic sensing setup based on geometric phases. *Sensors*, *24*(1), 30.

Sladen, A., Rivet, D., Ampuero, J. P., De Barros, L., Hello, Y., Calbris, G., & Lamare, P. (2019, December). Distributed sensing of earthquakes and ocean-solid Earth interactions on seafloor telecom cables. *Nat Commun*, *10*(1), 5777. Retrieved 2021-12-01, from `http://www.nature.com/articles/s41467-019-13793-z` doi: 10.1038/s41467-019-13793-z

Sobolevskaia, V., Ajo-Franklin, J., Cheng, F., Dou, S., Lindsey, N. J., & Wagner, A. (2024). Monitoring water level of a surficial aquifer using distributed acoustic sensing and ballistic surface waves. *Water Resources Research*, *60*(8), e2023WR036172.

Souriau, A., & Pauchet, H. (1998, May). A new synthesis of Pyrenean seismicity and its tectonic implications. *Tectonophysics*, *290*(3-4), 221–244. Retrieved 2023-10-02, from `https://`

linkinghub.elsevier.com/retrieve/pii/S0040195198000171 doi: 10.1016/S0040-1951(98)00017-1

Sparks, R. S. J. (2003). Forecasting volcanic eruptions. *Earth and Planetary Science Letters*, *210*(1-2), 1–15.

Spica, Z., Ajo-Franklin, J., Beroza, G., Biondi, B., Cheng, F., Gaite, B., … Zhu, T. (2022, September). *PubDAS: a PUBlic Distributed Acoustic Sensing datasets repository for geosciences* (preprint). Civil and Environmental Engineering. Retrieved 2022-12-20, from http://eartharxiv.org/repository/view/3574/ doi: 10.31223/X5D07S

Spica, Z. J., Castellanos, J. C., Viens, L., Nishida, K., Akuhara, T., Shinohara, M., & Yamada, T. (2022, January). Subsurface Imaging With Ocean-Bottom Distributed Acoustic Sensing and Water Phases Reverberations. *Geophysical Research Letters*, *49*(2). Retrieved 2023-08-02, from https://onlinelibrary.wiley.com/doi/10.1029/2021GL095287 doi: 10.1029/2021GL095287

Stajanca, P., Chruscicki, S., Homann, T., Seifert, S., Schmidt, D., & Habib, A. (2018). Detection of leak-induced pipeline vibrations using fiber—optic distributed acoustic sensing. *Sensors*, *18*(9), 2841.

Stanton, T., & Clay, C. (1986). Sonar echo statistics as a remote-sensing tool: Volume and seafloor. *IEEE Journal of Oceanic Engineering*, *11*(1), 79–96.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, *36*(2), 111–133.

Stork, A. L., Baird, A. F., Horne, S. A., Naldrett, G., Lapins, S., Kendall, J.-M., … Williams, A. (2020, September). Application of machine learning to microseismic event detection in distributed acoustic sensing data. *GEOPHYSICS*, *85*(5), KS149–KS160. Retrieved 2021-11-10, from https://library.seg.org/doi/10.1190/geo2019-0774.1 doi: 10.1190/geo2019-0774.1

Su, F., Anderson, J. G., Brune, J. N., & Zeng, Y. (1996). A comparison of direct s-wave and coda-wave site amplification determined from aftershocks of the little skull mountain earthquake. *Bulletin of the Seismological Society of America*, *86*(4), 1006–1018.

Sun, Q., Feng, H., Yan, X., & Zeng, Z. (2015). Recognition of a phase-sensitivity otdr sensing system based on morphologic feature extraction. *Sensors*, *15*(7), 15179–15197.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, *40*(12), 3358–3378.

Suriñach, E., Vilajosana, I., Khazaradze, G., Biescas, B., Furdada, G., & Vilaplana, J. M. (2005). Seismic detection and characterization of landslides and other mass movements. *Natural Hazards and Earth System Sciences*, *5*(6), 791–798.

Sylvander, M., Rigo, A., Sénéchal, G., Battaglia, J., Benahmed, S., Calvet, M., … Pauchet, H. (2022, January). Seismicity patterns in southwestern France. *Comptes Rendus. Géoscience*, *353*(S1), 79–104. Retrieved 2023-07-11, from https://comptes-rendus.academie-sciences.fr/geoscience/articles/10.5802/crgeos.60/ doi: 10.5802/crgeos.60

Take, W., Bolton, M., Wong, P., & Yeung, F. (2004). Evaluation of landslide triggering mechanisms in model fill slopes. *Landslides*, *1*, 173–184.

Takeo, A., Idehara, K., Iritani, R., Tonegawa, T., Nagaoka, Y., Nishida, K., … et al. (2010). Very broadband analysis of a swarm of very low frequency earthquakes and tremors beneath kii peninsula, sw japan. *Geophysical Research Letters*, *37*(6).

Talandier, J., Hyvernaud, O., Reymond, D., & Okal, E. A. (2006). Hydroacoustic signals generated by

parked and drifting icebergs in the southern indian and pacific oceans. *Geophysical Journal International*, *165*(3), 817–834.

Tanioka, Y., & Sataka, K. (1996). Fault parameters of the 1896 sanriku tsunami earthquake estimated from tsunami numerical modeling. *Geophysical research letters*, *23*(13), 1549–1552.

Teja, G. R., Harish, V., Khan, D. N. M., Krishna, R. B., Singh, R., & Chaudhary, S. (2014). Land slide detection and monitoring system using wireless sensor networks (wsn). In *2014 ieee international advance computing conference (iacc)* (pp. 149–154).

Tejedor, J., Macias-Guarasa, J., Martins, H. F., Martin-Lopez, S., & Gonzalez-Herraez, M. (2021a, March). A Multi-Position Approach in a Smart Fiber-Optic Surveillance System for Pipeline Integrity Threat Detection. *Electronics*, *10*(6), 712. Retrieved 2021-12-01, from https://www.mdpi.com/2079-9292/10/6/712 doi: 10.3390/electronics10060712

Tejedor, J., Macias-Guarasa, J., Martins, H. F., Martin-Lopez, S., & Gonzalez-Herraez, M. (2021b, March). A Multi-Position Approach in a Smart Fiber-Optic Surveillance System for Pipeline Integrity Threat Detection. *Electronics*, *10*(6), 712. Retrieved 2021-12-01, from https://www.mdpi.com/2079-9292/10/6/712 doi: 10.3390/electronics10060712

Thiran, J.-P., & Macq, B. (1996). Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Transactions on biomedical engineering*, *43*(10), 1011–1020.

Titos, M., Bueno, A., Garcia, L., Benitez, C., & Segura, J. C. (2020, May). Classification of Isolated Volcano-Seismic Events Based on Inductive Transfer Learning. *IEEE Geosci. Remote Sensing Lett.*, *17*(5), 869–873. Retrieved 2024-03-29, from https://ieeexplore.ieee.org/document/8798707/ doi: 10.1109/LGRS.2019.2931063

Tomek, I. (1976). Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.*.

Tonnellier, A., Helmstetter, A., Malet, J.-P., Schmittbuhl, J., Corsini, A., & Joswig, M. (2013). Seismic monitoring of soft-rock landslides: the super-sauze and valoria case studies. *Geophysical Journal International*, *193*(3), 1515–1536.

Trabattoni, A., Baillet, M., van den Ende, M., Rivet, D., Stutzman, E., Strumia, C., & Biagioli, F. (2024). Xdas: a python framework for distributed acoustic sensing. *EarthArXiv*.

Tribaldos, V., & Ajo-Franklin, J. B. (2021). Aquifer monitoring using ambient seismic noise recorded with distributed acoustic sensing (das) deployed on dark fiber. *Journal of Geophysical Research: Solid Earth*, *126*(4), e2020JB021004.

Trnkoczy, A. (2009). Understanding and parameter setting of sta/lta trigger algorithm. In *New manual of seismological observatory practice (nmsop)* (pp. 1–20). Deutsches GeoForschungsZentrum GFZ.

Tsuboi, S., Whitmore, P. M., & Sokolowski, T. J. (1999). Application of m wp to deep and teleseismic earthquakes. *Bulletin of the Seismological Society of America*, *89*(5), 1345–1351.

Turquet, A., Wuestefeld, A., Svendsen, G. K., Nyhammer, F. K., Nilsen, E., Persson, A. P.-O., & Refsum, V. (2024a). Automated avalanche monitoring, detection and classification system powered by distributed acoustic sensing. In *Proceedings, international snow science workshop*. Tromsø, Norway.

Turquet, A., Wuestefeld, A., Svendsen, G. K., Nyhammer, F. K., Nilsen, E., Persson, A. P.-O., & Refsum, V. (2024b). Automated snow avalanche monitoring and alert system using distributed acoustic sensing in norway. *Preprints*. Retrieved from https://doi.org/10.20944/preprints202412.0935.v1 doi: 10.20944/preprints202412.0935.v1

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

Vaezi, Y., & Van der Baan, M. (2015). Comparison of the sta/lta and power spectral density methods

for microseismic event detection. *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society*, *203*(3), 1896–1908.

Valagussa, A., Frattini, P., & Crosta, G. B. (2014). Earthquake-induced rockfall hazard zoning. *Engineering Geology*, *182*, 213-225. Retrieved from https://www.sciencedirect.com/science/article/pii/S0013795214001756 (Special Issue on The Long-Term Geologic Hazards in Areas Struck by Large-Magnitude Earthquakes) doi: https://doi.org/10.1016/j.enggeo.2014.07.009

van den Ende, M., Ferrari, A., Sladen, A., & Richard, C. (2021). Next-generation traffic monitoring with distributed acoustic sensing arrays and optimum array processing. *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 1104–1108.

Van den Ende, M., Lior, I., Ampuero, J.-P., Sladen, A., Ferrari, A., & Richard, C. (2021). A self-supervised deep learning approach for blind denoising and waveform coherence enhancement in distributed acoustic sensing data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(7), 3371–3384.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Van Herwijnen, A., & Schweizer, J. (2011). Monitoring avalanche activity using a seismic sensor. *Cold Regions Science and Technology*, *69*(2-3), 165–176.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Vilajosana, I., Suriñach, E., Abellán, A., Khazaradze, G., Garcia, D., & Llosa, J. (2008). Rockfall induced seismic signals: case study in montserrat, catalonia. *Natural Hazards and Earth System Sciences*, *8*(4), 805–812.

Walter, F., Gräff, D., Lindner, F., Paitz, P., Köpfli, M., Chmiel, M., & Fichtner, A. (2020). Distributed acoustic sensing of microseismic sources and wave propagation in glaciated terrain. *Nature communications*, *11*(1), 2436.

Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., Woelfel, I., Chang, J. C., ... Holland, A. A. (2020). The oklahoma geological survey statewide seismic network. *Seismological Research Letters*, *91*(2A), 611–621.

Wang, J., Bertasius, G., Tran, D., & Torresani, L. (2022). Long-short temporal contrastive learning of video transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 14010–14020).

Wang, S., Liu, F., & Liu, B. (2022). Semi-supervised deep learning in high-speed railway track detection based on distributed fiber acoustic sensing. *Sensors*, *22*(2), 413.

Wang, T., Bian, Y., Zhang, Y., & Hou, X. (2023). Classification of earthquakes, explosions and mining-induced earthquakes based on xgboost algorithm. *Computers & Geosciences*, *170*, 105242.

Wang, W.-M., Hao, J.-L., & Yao, Z.-X. (2013). Preliminary result for rupture process of apr. 20, 2013, lushan earthquake, sichuan, china. *Chinese Journal of Geophysics*, *56*(4), 1412–1417.

Wang, Y., Wang, P., Ding, K., Li, H., Zhang, J., Liu, X., ... Jin, B. (2019). Pattern Recognition Using Relevant Vector Machine in Optical Fiber Vibration Sensing System. *IEEE Access*, *7*, 5886–5895. Retrieved 2023-08-02, from https://ieeexplore.ieee.org/document/8598862/ doi: 10.1109/ACCESS.2018.2889699

Wang, Z., Lou, S., Liang, S., & Sheng, X. (2020). Multi-class disturbance events recognition based on emd and xgboost in φ-otdr. *IEEE Access*, *8*, 63551–63558.

Wanzenböck, J., Mehner, T., Schulz, M., Gassner, H., & Winfield, I. J. (2003). Quality assurance of hydroacoustic surveys: the repeatability of fish-abundance and biomass estimates in lakes

within and between hydroacoustic systems. *ICES Journal of Marine Science*, *60*(3), 486–492.

Ward, S. N. (1980). Relationships of tsunami generation and an earthquake source. *Journal of Physics of the Earth*, *28*(5), 441–474.

Ward, S. N. (2001). Landslide tsunami. *Journal of Geophysical Research: Solid Earth*, *106*(B6), 11201–11215.

Webb, S. C. (1998). Broadband seismology and noise under the ocean. *Reviews of Geophysics*, *36*(1), 105–142.

Wenner, M., Hibert, C., van Herwijnen, A., Meier, L., & Walter, F. (2021, January). Near-real-time automated classification of seismic signals of slope failures with continuous random forests. *Nat. Hazards Earth Syst. Sci.*, *21*(1), 339–361. Retrieved 2022-02-02, from https://nhess.copernicus.org/articles/21/339/2021/ doi: 10.5194/nhess-21-339-2021

Wesnousky, S. G. (2008). Displacement and geometrical characteristics of earthquake surface ruptures: Issues and implications for seismic-hazard analysis and the process of earthquake rupture. *Bulletin of the Seismological Society of America*, *98*(4), 1609–1632.

West, M. E., Larsen, C. F., Truffer, M., O'Neel, S., & LeBlanc, L. (2010). Glacier microseismicity. *Geology*, *38*(4), 319–322.

White, R., & McCausland, W. (2016). Volcano-tectonic earthquakes: A new tool for estimating intrusive volumes and forecasting eruptions. *Journal of Volcanology and Geothermal Research*, *309*, 139–155.

Whiteley, J., Chambers, J., Uhlemann, S., Wilkinson, P. B., & Kendall, J. (2019). Geophysical monitoring of moisture-induced landslides: A review. *Reviews of Geophysics*, *57*(1), 106–145.

Wiesmeyr, C., Coronel, C., Litzenberger, M., Döller, H. J., Schweiger, H.-B., & Calbris, G. (2021). Distributed acoustic sensing for vehicle speed and traffic flow estimation. In *2021 ieee international intelligent transportation systems conference (itsc)* (pp. 2596–2601).

Wiesmeyr, C., Litzenberger, M., Waser, M., Papp, A., Garn, H., Neunteufel, G., & Döller, H. (2020, January). Real-Time Train Tracking from Distributed Acoustic Sensing Data. *Applied Sciences*, *10*(2), 448. Retrieved 2021-12-01, from https://www.mdpi.com/2076-3417/10/2/448 doi: 10.3390/app10020448

Williams, E. F., Fernández-Ruiz, M. R., Magalhaes, R., Vanthillo, R., Zhan, Z., González-Herráez, M., & Martins, H. F. (2019). Distributed sensing of microseisms and teleseisms with submarine dark fibers. *Nature Communications*, *10*(1), 5778. Retrieved from https://www.nature.com/articles/s41467-019-13262-7 doi: 10.1038/s41467-019-13262-7

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, *30*(1), 79–82.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*(3), 408–421.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S., & Trujillo, J. (1998). A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bulletin of the Seismological Society of America*, *88*(1), 95–106.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern recognition*, *48*(9), 2839–2846.

Wu, H., Xiao, S., Li, X., Wang, Z., Xu, J., & Rao, Y. (2015). Separation and determination of the disturbing signals in phase-sensitive optical time domain reflectometry. *Journal of Lightwave*

Technology, *33*(15), 3156–3162.

Wuestefeld, A., & Wilks, M. (2019). How to twist and turn a fiber: Performance modeling for optimal das acquisitions. *The Leading Edge*, *38*(3), 226–231.

Wurman, G., Allen, R. M., & Lombard, P. (2007). Toward earthquake early warning in northern california. *Journal of Geophysical Research: Solid Earth*, *112*(B8).

Xie, T., Zhang, C.-C., Shi, B., Wang, Z., Zhang, S.-S., & Yin, J. (2023). Seismic monitoring of rockfalls using distributed acoustic sensing. *Engineering Geology*, *325*, 107285.

Xiong, C., Lu, H., & Zhu, J. (2017). Operational modal analysis of bridge structures with data from gnss/accelerometer measurements. *Sensors*, *17*(3), 436.

Yan, Y., Cui, Y., Tian, X., Hu, S., Guo, J., Wang, Z., … Liao, L. (2020). Seismic signal recognition and interpretation of the 2019 "7.23" shuicheng landslide by seismogram stations. *Landslides*, *17*, 1191–1206.

Yang, L., Fomel, S., Wang, S., Chen, X., Chen, W., Saad, O. M., & Chen, Y. (2023). Denoising of distributed acoustic sensing data using supervised deep learning. *Geophysics*, *88*(1), WA91–WA104.

Yoon, C. E., O'Reilly, O., Bergen, K. J., & Beroza, G. C. (2015). Earthquake detection through computationally efficient similarity search. *Science advances*, *1*(11), e1501057.

Young, C., Shragge, J., Schultz, W., Haines, S., Oren, C., Simmons, J., & Collett, T. S. (2022, April). Advanced Distributed Acoustic Sensing Vertical Seismic Profile Imaging of an Alaska North Slope Gas Hydrate Field. *Energy Fuels*, *36*(7), 3481–3495. Retrieved 2023-08-02, from https://pubs.acs.org/doi/10.1021/acs.energyfuels.1c04102 doi: 10.1021/acs.energyfuels.1c04102

Yu, C., Zhan, Z., Lindsey, N. J., Ajo-Franklin, J. B., & Robertson, M. (2019). The potential of DAS in teleseismic studies: Insights from the goldstone experiment. *Geophysical Research Letters*, *46*(3), 1320–1328. Retrieved 2024-11-10, from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018GL081195 doi: 10.1029/2018GL081195

Yu, X., Zhou, D., Lu, B., Liu, S., & Pan, M. (2015). Phase-sensitive optical time domain reflectometer for distributed fence-perimeter intrusion detection. In *Aopc 2015: Optical fiber sensors and applications* (Vol. 9679, pp. 157–161).

Yuan, S., Lellouch, A., Clapp, R. G., & Biondi, B. (2020). Near-surface characterization using a roadside distributed acoustic sensing array. *The Leading Edge*, *39*(9), 646–653.

Yuan, S., Liu, J., Young Noh, H., & Biondi, B. (2021, September). Urban system monitoring using combined vehicle onboard sensing and roadside distributed acoustic sensing. In *First International Meeting for Applied Geoscience & Energy Expanded Abstracts* (pp. 3235–3239). Denver, CO and virtual: Society of Exploration Geophysicists. Retrieved 2021-12-08, from https://library.seg.org/doi/10.1190/segam2021-3584136.1 doi: 10.1190/segam2021-3584136.1

Zali, Z., Mousavi, S. M., Ohrnberger, M., Eibl, E. P., & Cotton, F. (2024). Tremor clustering reveals pre-eruptive signals and evolution of the 2021 geldingadalir eruption of the fagradalsfjall fires, iceland. *Communications Earth & Environment*, *5*(1), 1.

Zeng, X., Bao, F., Thurber, C. H., Lin, R., Wang, S., Song, Z., & Han, L. (2022, March). Turning a Telecom Fiber-Optic Cable into an Ultradense Seismic Array for Rapid Postearthquake Response in an Urban Area. *Seismological Research Letters*, *93*(2A), 853–865. Retrieved 2022-05-31, from https://pubs.geoscienceworld.org/srl/article/93/2A/853/610193/Turning-a-Telecom-Fiber-Optic-Cable-into-an doi: 10.1785/0220210183

Zeng, X., & Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(1), 1–12.

Zhang, G., Song, Z., Osotuyi, A. G., Lin, R., & Chi, B. (2022). Railway traffic monitoring with trackside fiber-optic cable by distributed acoustic sensing technology. *Frontiers in Earth Science*, *10*, 990837.

Zhang, L., & Zhan, C. (2017). Machine learning in rock facies classification: An application of xgboost. In *International geophysical conference, qingdao, china, 17-20 april 2017* (pp. 1371–1374).

Zhao, Y., Li, Y., & Wu, N. (2021). Distributed Acoustic Sensing Vertical Seismic Profile Data Denoiser Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sensing*, 1–11. Retrieved 2023-10-02, from `https://ieeexplore.ieee.org/document/9293178/` doi: 10.1109/TGRS.2020.3042202

Zhong, R., Johnson Jr, R., & Chen, Z. (2020). Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (xgboost). *International Journal of Coal Geology*, *220*, 103416.

Zhu, T., Shen, J., & Martin, E. R. (2021, January). Sensing Earth and environment dynamics by telecommunication fiber-optic sensors: an urban experiment in Pennsylvania, USA. *Solid Earth*, *12*(1), 219–235. Retrieved 2022-02-21, from `https://se.copernicus.org/articles/12/219/2021/` doi: 10.5194/se-12-219-2021

Zhu, T., & Stensrud, D. J. (2019). Characterizing thunder-induced ground motions using fiber-optic distributed acoustic sensing array. *Journal of Geophysical Research: Atmospheres*, *124*(23), 12810–12823.

Zhu, W., & Beroza, G. C. (2019). Phasenet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273.

Zhu, W., Biondi, E., Li, J., Yin, J., Ross, Z. E., & Zhan, Z. (2023). Seismic arrival-time picking on distributed acoustic sensing data using semi-supervised learning. *Nature Communications*, *14*(1), 8192.

Zhu, X., & Goldberg, A. B. (2022). *Introduction to semi-supervised learning*. Springer Nature.

Zhu, X. J. (2005). Semi-supervised learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.

Zimmer, V. L., & Sitar, N. (2015). Detection and location of rock falls using seismic and infrasound sensors. *Engineering Geology*, *193*, 49–60.

Zou, X., Thiruvenkatanathan, P., & Seshia, A. A. (2014). A seismic-grade resonant mems accelerometer. *Journal of Microelectromechanical Systems*, *23*(4), 768–770.

Appendices

Appendix A

Paper: Real-Time Classification of Anthropogenic Seismic Sources from Distributed Acoustic Sensing Data: Application for Pipeline Monitoring

A1 Anthropogenic-Seismic Source Definition and Simulation Protocols

Appendix A1 provides a table detailing the experimental protocols for generating the seismic sources (Table A1.1). Because some sources have a similar acoustic signature (Falling objects, Manual excavation and Hammer impact),these have been grouped into a single class (Impact) for signal sources recognition.

Table A1.1: Anthropogenic-seismic source definition and simulation protocols.

Class	Sub-class	Protocol
Pedestrian walk	-	Multiple round trips of a person walking/running along the pipeline, each trip is followed by a 10 s break. (Weight of pedestrians: 50 – 80 kg)
Impact	Falling objects	One or two hammers/shovels/rocks/screwdrivers are dropped by an operator on the trench, are collected at least 5 s later, and are dropped back 5 s later. (Weight of screwdriver: 0.1 kg, hammer: 0.8 kg, shovel: 2 kg, rocks: 1 – 5 kg)
Impact	Manual excavation	An operator is either continuously (at least during 30 s) digging a hole using a pickaxe/shovel, either continuously (at least during 30 s) filing/compacting a hole using a shovel. (Weight of pickaxe: 1.5 kg, shovel: 2 kg)
Impact	Hammer impact	An operator is fastly/slowly hitting a pipeline buried in the trench. (Weight of hammer: 0.8 kg)
Backhoe	-	A backhoe is either entering/leaving the trench, either digging/filling a hole. (Weight of backhoe: 2,600 kg)

Continued on next page

Table A1.1 – continued from previous page

Class	Sub-class	Protocol
Compactor	-	An operator is compacting the dirt using a ramming machine. (Weight of compactor: 70 kg)
Water/air leakages	-	Before the measurement begins, the pipeline is filled with air/hot water under pressure (50 bars for air, 20 bars for hot water). At the beginning of the measurement, an operator remotely opens the electrovalve. At the end, the electrovalve is closed back and the measurement is stopped. The diameters of the nozzle are 3 and 5 mm.

A2 Features Used to Describe the Acoustic Signal

Appendix A2 provides an overview of all the features used for the machine learning process. The features can be grouped into three main groups: waveform attributes, spectral attributes and spectrogram attributes. Table A2.1 provides a contextual and mathematical description of each of them.

TABLE A2.1: Features used to describe the acoustic signal.

Number	Description	Formula
Temporal features - Waveform		
1	Ratio of the mean over the maximum of the envelop signal	-
2	Ratio of the median over the maximum of the envelop signal	-
3	Ratio between ascending and descending time	$\frac{t_{max}-t_i}{t_f-t_{max}}$ with t_{max}: time of the largest amplitude
4	Kurtosis of the raw signal (peakness of the signal)	$\frac{m_4}{\theta^4}$ with m_4: fourth moment, θ: standard deviation
5	Kurtosis of the envelop	see 4
6	Skewness of the raw signal	$\frac{m_3}{\theta^3}$ with m_3: third moment, θ: standard deviation
7	Skewness of the envelop	see 6
8	Number of peaks in the auto-correlation function	-
9	Energy in the first third part of the auto-correlation function	$\int_0^{T/3} C(\tau)d\tau$, with T: signal duration, C: auto-correlation function
10	Energy in the remaining part of the auto-correlation function	see 9
11	Ratio of 10 and 9	-

Continued on next page

Table A2.1 – continued from previous page

Number	Description	Formula
12-16	Energy of the signal filtered in 5–10 Hz, 10–30 Hz, 30–50 Hz, 50–75 Hz, 75-100 Hz	$\int_0^T y_f(t)dt$, with y_f: filtered signal in the frequency range $[f_1, f_2]$
17-21	Kurtosis of the signal filtered in 5–10 Hz, 10–30 Hz, 30-50 Hz, 50-75 Hz, 75-100 Hz	see 4
22	RMS between the decreasing part of the signal and $I(t) = Y_{max} - \frac{Y_{max}}{t_f - t_{max}} t$	$\sqrt{\overline{Y(t) - I(t)}^2}$, with Y: envelop of the signal spectral featurees
Temporal features - Spectral		
23	Mean of the DFT	DFT: discrete Fourier transform
24	Max of the DFT	-
25	Central frequency of the first quartile	-
26	Central frequency of the third quartile	-
27	Median of the normalized DFT	-
28	Variance of the normalized DFT	-
29	Number of peaks ($> 0.75 DFT_{max}$)	DFT_{max}: maximum of the DFT
30-33	Energy in $[0\frac{1}{4}]$ NyF, $[\frac{1}{4}\frac{1}{2}]$ NyF, $[\frac{1}{2}\frac{3}{4}]$ NyF, $[\frac{3}{4}1]$ NyF	$\int_{f_1}^{f_2} DFT(f)df$, with f_1, f_2: considered frequency ranges
34	Spectral centroid	$\gamma_1 = \frac{m_2}{m_1}$, with m_1, m_2: first and second moment
35	Gyration radius	$\gamma_2 = \sqrt{\frac{m_3}{m_2}}$, with m_3: third moment
36	Spectral centroid width	$\gamma_3 = \sqrt{\gamma_1^2 + \gamma_2^2}$
Temporal features - Spectrogram*		
37	Kurtosis of the maximum of all discrete Fourier transforms (DFTs)	$kurtosis[max_{t\in[0,T]}(SPEC(t,f))]$, with $SPEC(t,f)$: spectrogram as a function of time t
38	Kurtosis of the maximum of all DFTs as a function of time t	see 37
39	Mean ratio between the maximum and the mean of all DFTs	$mean(\frac{max(SPEC)}{mean(SPEC)})$
40	Mean ratio between the maximum and the median of all DFTs	see 39
41	Number of peaks in the curve showing the temporal evolution of the DFTs maximum	-
42	Number of peaks in the curve showing the temporal evolution of the DFTs mean	-

Continued on next page

Table A2.1 – continued from previous page

Number	Description	Formula
43	Number of peaks in the curve showing the temporal evolution of the DFTs median	-
44	Ratio between 41 and 42	-
45	Ratio between 41 and 43	-
46	Number of peaks in the curve of the temporal evolution of the DFTs central frequency	-
47	Number of peaks in the curve of the temporal evolution of the DFTs maximum frequency	-
48	Ratio between 46 and 47	-
49	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and mean frequency	-
50	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and median frequency	-
51	Mean distance between the 1st quartile and the median of all DFTs as a function of time	-
52	Mean distance between the 3rd quartile and the median of all DFTs as a function of time	-
53	Mean distance between the 3rd quartile and the 1st quartile of all DFTs as a function of time	-

* The spectrogram is the collection of the DFTs computed for signal windows of 1 s with an overlap of 90%. The spectrogram is represented as a 2D matrix representing the evolution of the frequency content (rows) through time (columns). Adapted from Provost et al. (2017).

Appendix B

Paper: A Real Scale Application of a Novel Set of Spatial and Similarity Features for Detection and Classification of Natural Seismic Sources from Distributed Acoustic Sensing Data

B1 The Nineteen Identified Events on DAS recording

Appendix B1 gives an overview of all the events recorded by BCSF-RENASS and visually inspected on the DAS recordings. Figure B1.1 presents the SR, EB, score map of the machine learning algorithm and detection map obtained using our processing chain for each event.

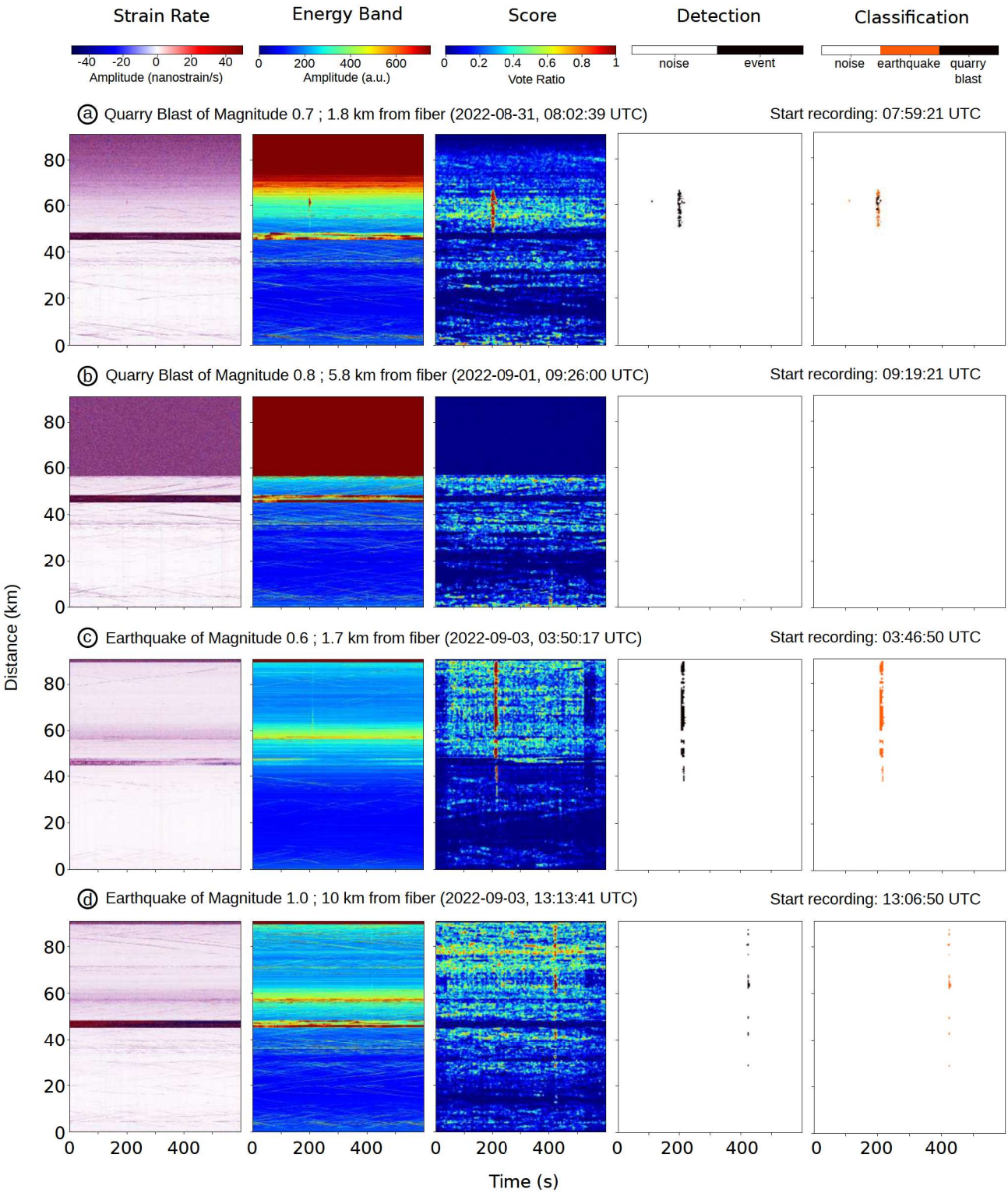
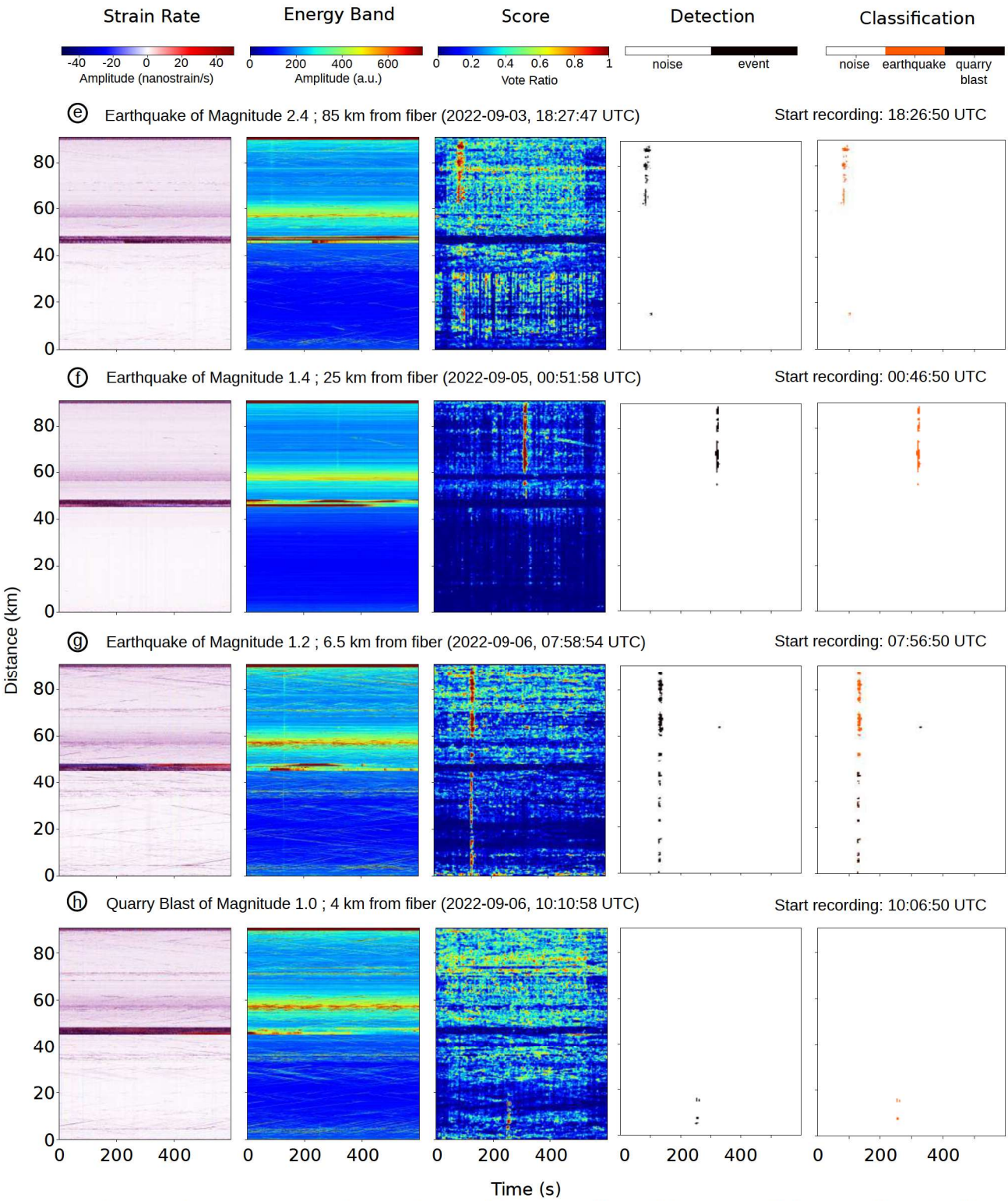
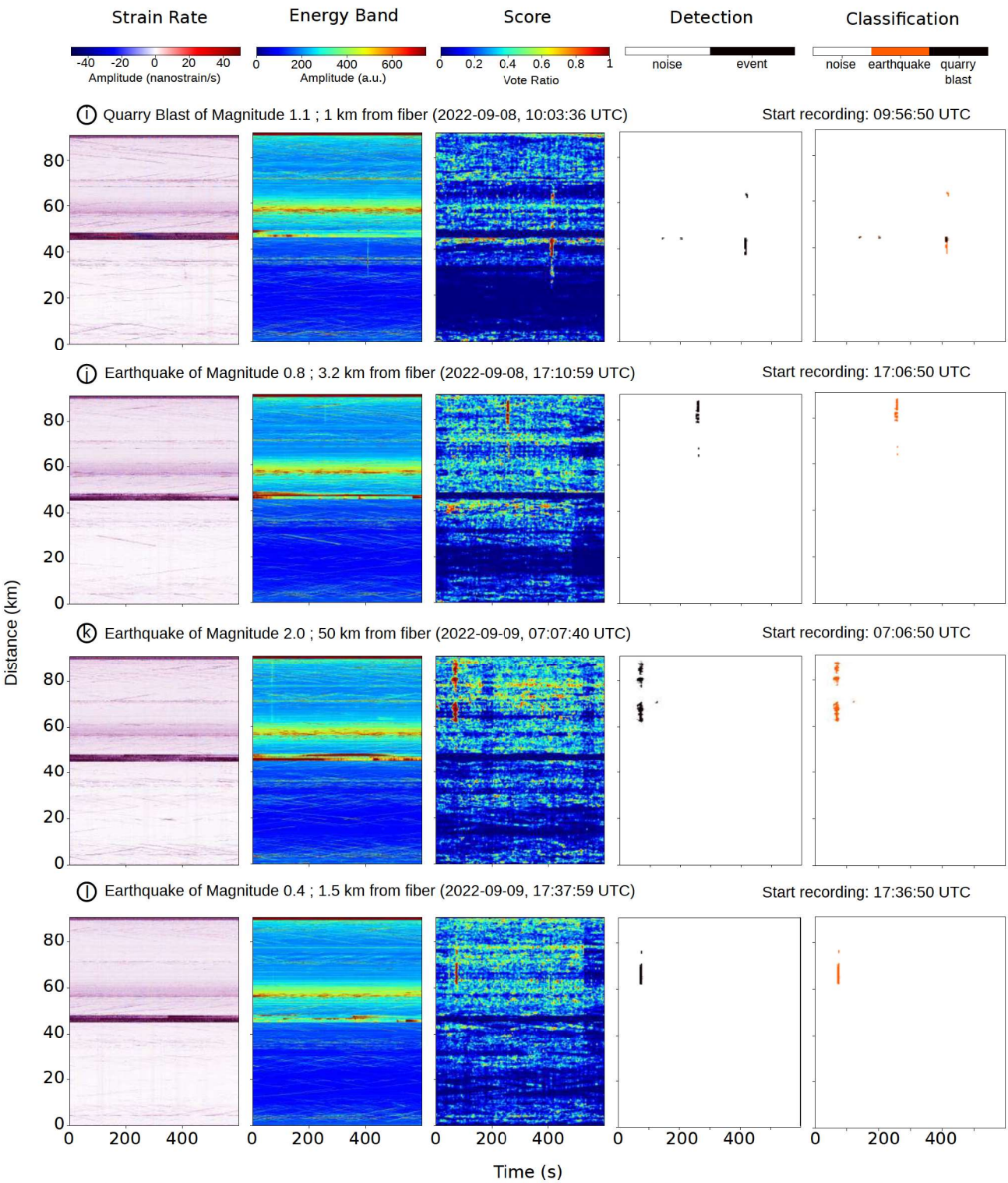


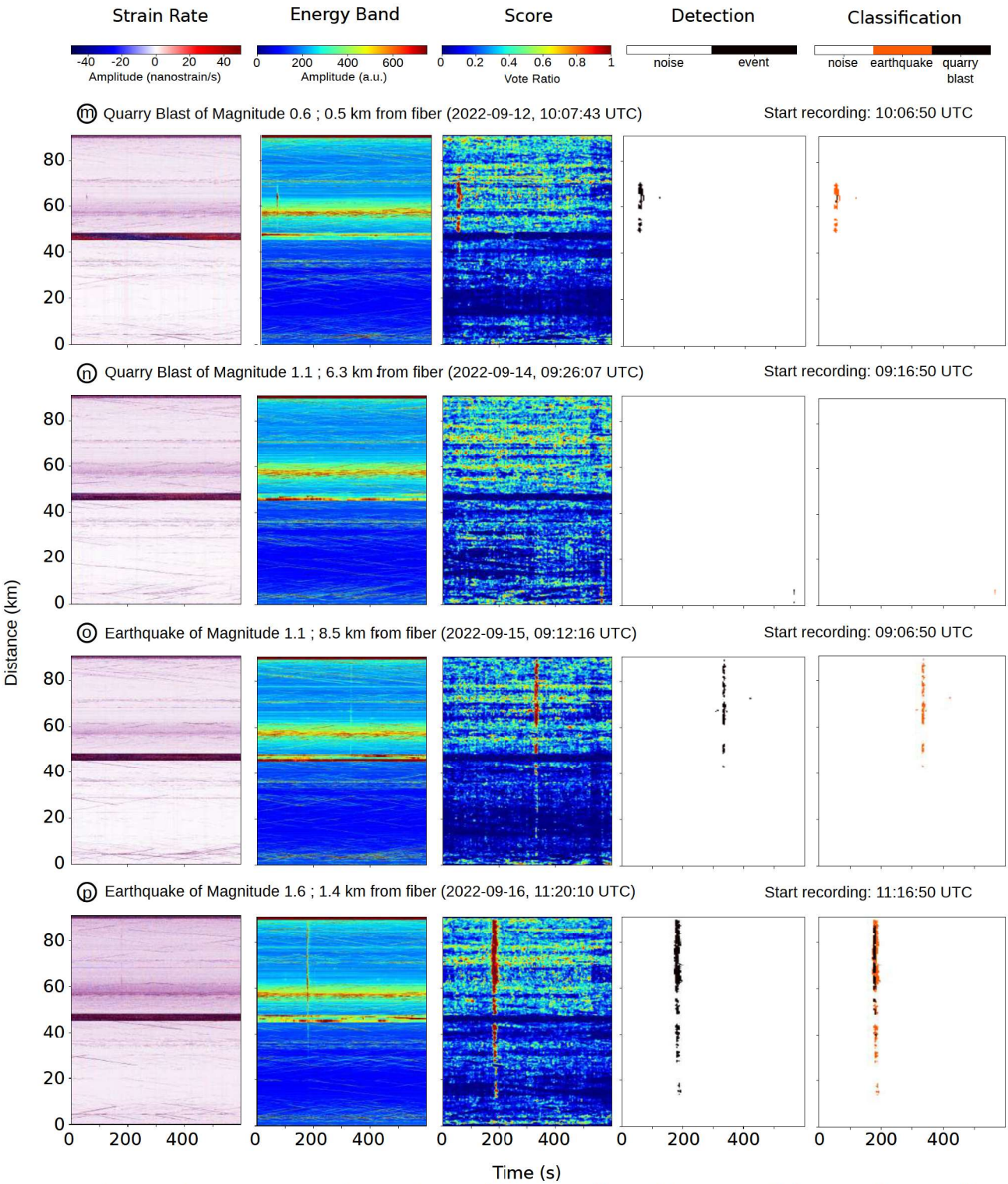
FIGURE B1.1: Strain Rate, Energy Band, Score Map and Detection Map of the 19 recorded events along the fiber. Strain Rate and Energy Band are computed directly from acquired data, whereas Score Map and Detection Map are obtained using our classification processing chain.



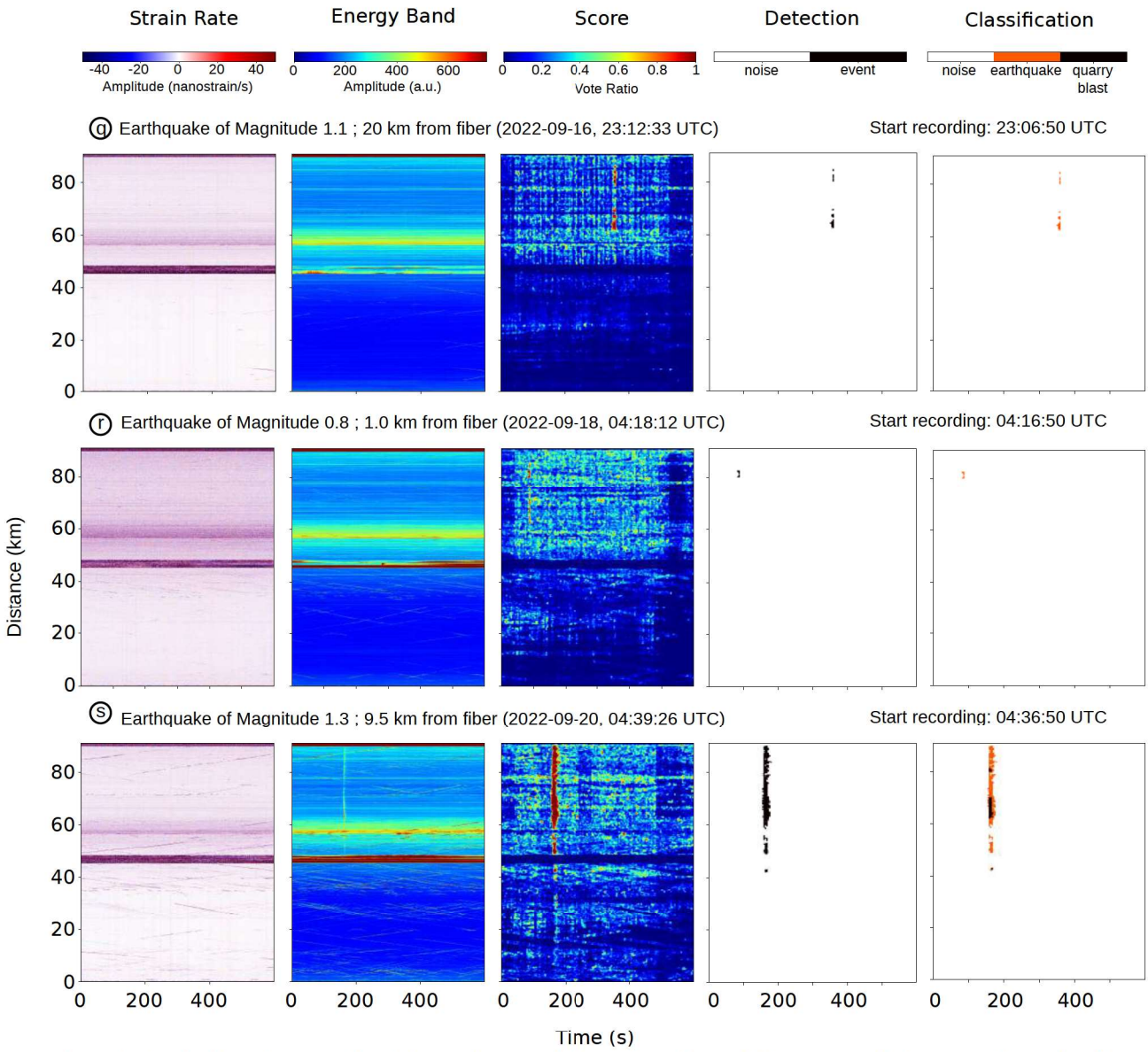
(continued) Strain Rate, Energy Band, Score Map and Detection Map of the 19 recorded events along the fiber. Strain Rate and Energy Band are computed directly from acquired data, whereas Score Map and Detection Map are obtained using our classification processing chain.



(continued) Strain Rate, Energy Band, Score Map and Detection Map of the 19 recorded events along the fiber. Strain Rate and Energy Band are computed directly from acquired data, whereas Score Map and Detection Map are obtained using our classification processing chain.



(continued) Strain Rate, Energy Band, Score Map and Detection Map of the 19 recorded events along the fiber. Strain Rate and Energy Band are computed directly from acquired data, whereas Score Map and Detection Map are obtained using our classification processing chain.



(continued) Strain Rate, Energy Band, Score Map and Detection Map of the 19 recorded events along the fiber. Strain Rate and Energy Band are computed directly from acquired data, whereas Score Map and Detection Map are obtained using our classification processing chain.

B2 Features Used to Describe the SR Signal

Appendix B2 gives an overview of all the features used for the machine learning process. The Table B2.1 gathers the temporal features related to waveform, the temporal features related to spectrum, the temporal features related to spectrogram, the spatial features and the similarity features. Spatial and similarity features were developed in this study specifically for DAS recordings.

TABLE B2.1: Features Used to Describe the SR Signal.

Number	Description	Formula
	Temporal features - Waveform	
	Averaging in window	
1	Ratio of the mean over the maximum of the envelop signal	-
2	Ratio of the median over the maximum of the envelop signal	-
3	Ratio between ascending and descending time	$\frac{t_{max}-t_i}{t_f-t_{max}}$ with t_{max}: time of the largest amplitude
4	Kurtosis of the raw signal (peakness of the signal)	$\frac{m_4}{\theta^4}$ with m_4: fourth moment, θ: standard deviation
5	Kurtosis of the envelop	see 4
6	Skewness of the raw signal	$\frac{m_3}{\theta^3}$ with m_3: third moment, θ: standard deviation
7	Skewness of the envelop	see 6
8	Number of peaks in the auto-correlation function	-
9	Energy in the first third part of the auto-correlation function	$\int_0^{T/3} C(\tau)d\tau$, with T: signal duration, C: auto-correlation function
10	Energy in the remaining part of the auto-correlation function	see 9
11	Ratio of 10 and 9	-
12-16	Energy of the signal filtered in 5–10 Hz, 10–30 Hz, 30–50 Hz, 50–75 Hz, 75-100 Hz	$\int_0^T y_f(t)dt$, with y_f: filtered signal in the frequency range $[f_1, f_2]$
17-22	Ratio between (12,13), (12,14), (12,15), (13,14), (13,15), (14,15)	-
23-27	Kurtosis of the signal filtered in 5–10 Hz, 10–30 Hz, 30-50 Hz, 50-75 Hz, 75-100 Hz	see 4
28	RMS between the decreasing part of the signal and $I(t) = Y_{max} - \frac{Y_{max}}{t_f-t_{max}}t$	$\sqrt{\overline{Y(t)-I(t)}^2}$, with Y: envelop of the signal spectral featurees
29	Maximum of envelope	-
	Temporal features - Spectral	
	Averaging in window	
30	Mean of the DFT	DFT: discrete Fourier transform
31	Max of the DFT	-
32	Frequency at the maximum of the DFT	-
33	Frequency of spectrum centroid	-
34	Central frequency of the first quartile	-
35	Central frequency of the third quartile	-
36	Median of the normalized DFT	-

Continued on next page

Table B2.1 – continued from previous page

Number	Description	Formula
37	Variance of the normalized DFT	-
38	Number of peaks ($> 0.75DFT_{max}$)	DFT_{max}: maximum of the DFT
39	Mean peaks value for peaks > 0.7	-
40-43	Energy in $[0\frac{1}{4}]$ NyF, $[\frac{1}{4}\frac{1}{2}]$ NyF, $[\frac{1}{2}\frac{3}{4}]$ NyF, $[\frac{3}{4}1]$ NyF	$\int_{f_1}^{f_2} DFT(f)df$, with f_1, f_2: considered frequency ranges
44	Spectral centroid	$\gamma_1 = \frac{m_2}{m_1}$, with m_1, m_2: first and second moment
45	Gyration radius	$\gamma_2 = \sqrt{\frac{m_3}{m_2}}$, with m_3: third moment
46	Spectral centroid width	$\gamma_3 = \sqrt{\gamma_1^2 + \gamma_2^2}$
	Temporal features - Spectrogram	
Averaging in window		
47	Kurtosis of the maximum of all discrete Fourier transforms (DFTs)	kurtosis[$max_{t\in[0,T]}$ (SPEC(t,f))], with SPEC(t,f): spectrogram as a function of time t
48	Kurtosis of the maximum of all DFTs as a function of time t	see 47
49	Mean ratio between the maximum and the mean of all DFTs	mean($\frac{max(SPEC)}{mean(SPEC)}$)
50	Mean ratio between the maximum and the median of all DFTs	see 49
51	Number of peaks in the curve showing the temporal evolution of the DFTs maximum	-
52	Number of peaks in the curve showing the temporal evolution of the DFTs mean	-
53	Number of peaks in the curve showing the temporal evolution of the DFTs median	-
54	Ratio between 51 and 52	-
55	Ratio between 51 and 53	-
56	Number of peaks in the curve of the temporal evolution of the DFTs central frequency	-
57	Number of peaks in the curve of the temporal evolution of the DFTs maximum frequency	-
58	Ratio between 56 and 57	-
59	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and mean frequency	-
60	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and median frequency	-
61	Mean distance between the 1st quartile and the median of all DFTs as a function of time	-
62	Mean distance between the 3rd quartile and the median of all DFTs as a function of time	-
63	Mean distance between the 3rd quartile and the 1st quartile of all DFTs as a function of time	-
	Spatial features	
Averaging in window		

Continued on next page

Table B2.1 – continued from previous page

Number	Description	Formula
64	Mean of the envelope of spatial trace	-
65	Mean of the raw spatial trace	-
66	Standard deviation of the envelope of spatial trace	-
67	Standard deviation of the raw spatial trace	-
68	Kurtosis of the envelope of spatial trace	see 4
69	Kurtosis of the raw spatial trace	see 4
70	Skewness of the envelope of spatial trace	see 6
71	Skewness of the raw spatial trace	see 6
72	Number of peaks of the auto-correlation function of spatial trace	-
73	Energy in the 1/3 around the origin of the auto-corr function of spatial trace	see 9
74	Energy in the last 2/3 of the autocorr function of spatial trace	see 9
75	Ratio of 74 and 73	-
Standard deviation in window		
76	Mean of the envelope of spatial trace	-
77	Mean of the raw spatial trace	-
78	Standard deviation of the envelope of spatial trace	-
79	Standard deviation of the raw spatial trace	-
80	Kurtosis of the envelope of spatial trace	see 4
81	Kurtosis of the raw spatial trace	see 4
82	Skewness of the envelope of spatial trace	see 6
83	Skewness of the raw spatial trace	see 6
84	Number of peaks of the auto-correlation function of spatial trace	-
85	Energy in the 1/3 around the origin of the auto-corr function of spatial trace	see 9
86	Energy in the last 2/3 of the autocorr function of spatial trace	see 9
87	Ratio of 86 and 85	-
	Similarity features	
Averaging in window		
88	Index of Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	-
89	Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	-
90	Similarity measurement of two consecutive stacks of temporal trace provided from DTW function	-
91	Integral of the DTW temporal distortion function*	-
92	Energy in the first third part of the auto-correlation function of the DTW temporal distortion function*	-

Continued on next page

Table B2.1 – continued from previous page

Number	Description	Formula
93	Energy in the remaining part of the auto-correlation function of the DTW temporal distortion function*	-
94	Ratio of 93 and 92	-
95	Difference of 93 and 92	-
Standard deviation in window		
96	Index of Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	-
97	Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	-
98	Similarity measurement of two consecutive stacks of temporal trace provided from DTW function	-
99	Integral of the DTW temporal distortion function*	-
100	Energy in the first third part of the auto-correlation function of the DTW temporal distortion function*	-
101	Energy in the remaining part of the auto-correlation function of the DTW temporal distortion function*	-
102	Ratio of 101 and 100	-
103	Difference of 101 and 100	-
Median in window		
104	Index of Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	-
105	Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	-
106	Similarity measurement of two consecutive stacks of temporal trace provided from DTW function	-
107	Integral of the DTW temporal distortion function*	-
108	Energy in the first third part of the auto-correlation function of the DTW temporal distortion function*	-
109	Energy in the remaining part of the auto-correlation function of the DTW temporal distortion function*	-
110	Ratio of 109 and 108	-
111	Difference of 109 and 108	-

* DTW temporal distortion function corresponds to the estimated stretch to apply to points of a trace A to obtain a trace closest to trace B.

Appendix C

Paper: Unsupervised Learning for the Comprehensive Exploration of Continuous-DAS Data

C1 Viella Site Photo and Work Description

Appendix C1 provides a photo of the Viella site (Figure C1.1), and a photo of the fiber optic field work, consisting in making a furrow with a hoe (Figure C1.2a) and then putting the fiber optic cable inside the furrow (Figure C1.2b).

FIGURE C1.1: *Viella landslide (taken by J.-P. Malet).*

FIGURE C1.2: *Viella field work.* ***(a)*** *shows making a furrow with a hoe,* ***(b)*** *shows putting the fiber optic cable inside the furrow.*

C2 t-SNE Vizualisation of the Produced Embeddings

Appendix C2 provides a t-SNE representation of the Pyrenees and Viella datasets with the EB thumbnails, using the human-engineered or image-BYOL latent space. These images help clarify the relationships between points in the t-SNE space.

1. Figure C2.1: Pyrenees dataset with the human-engineered latent space.
2. Figure C2.2: Pyrenees dataset with the image-BYOL latent space.
3. Figure C2.3: Viella dataset with the human-engineered latent space.
4. Figure C2.4: Viella dataset with the image-BYOL latent space.

Representation of t-SNE with classes encoded in color.

Figure C2.1: t-SNE representation of the Pyrenees dataset obtained with the human-engineered latent space. The thumbnails correspond to the EB representation of the windowed signal.

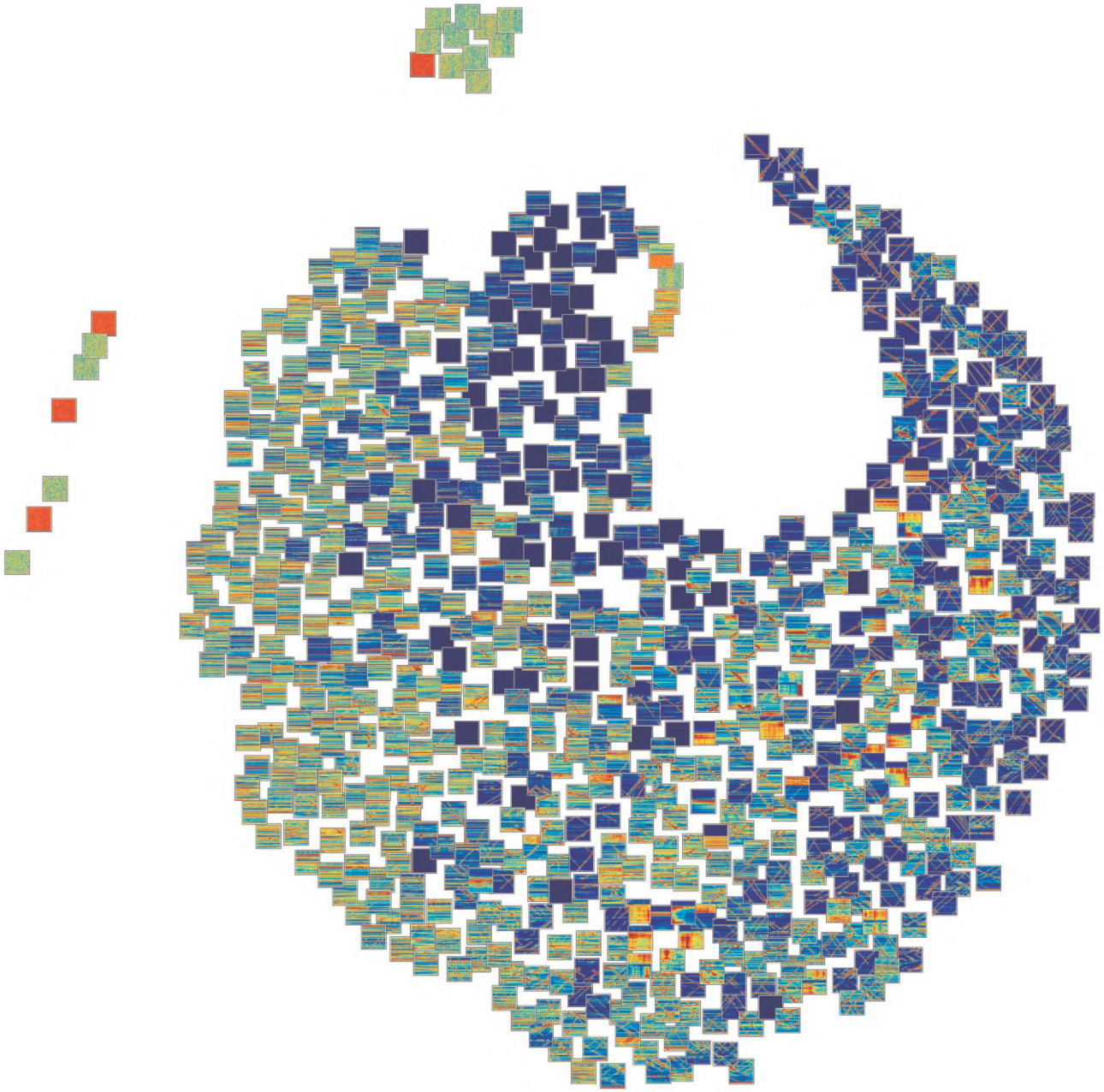


Figure C2.2: t-SNE representation of the Pyrenees dataset obtained with the image-BYOL latent space. The thumbnails correspond to the EB representation of the windowed signal.

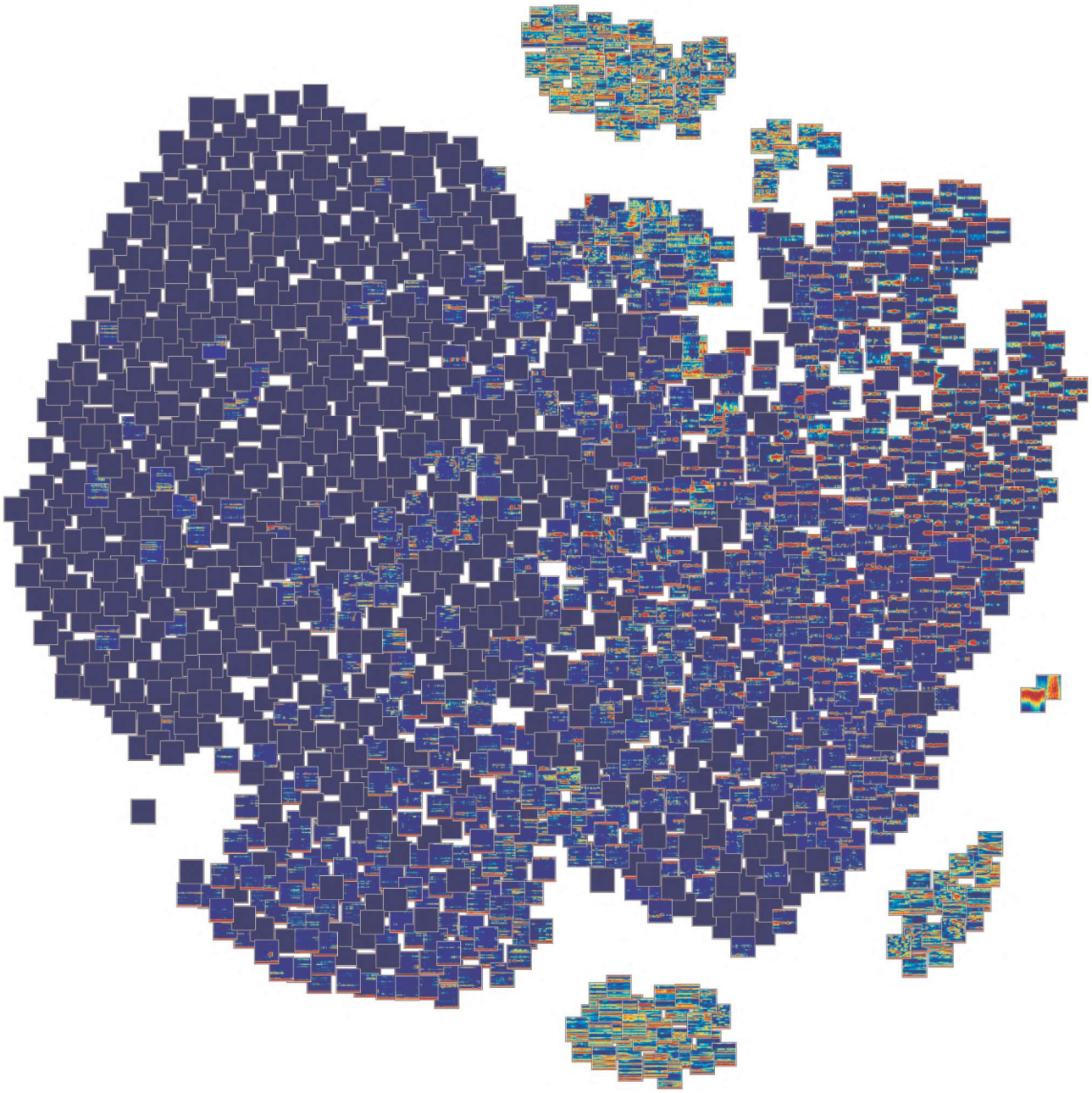


FIGURE C2.3: t-SNE representation of the Viella dataset obtained with the human-engineered latent space. The thumbnails correspond to the EB representation of the windowed signal.

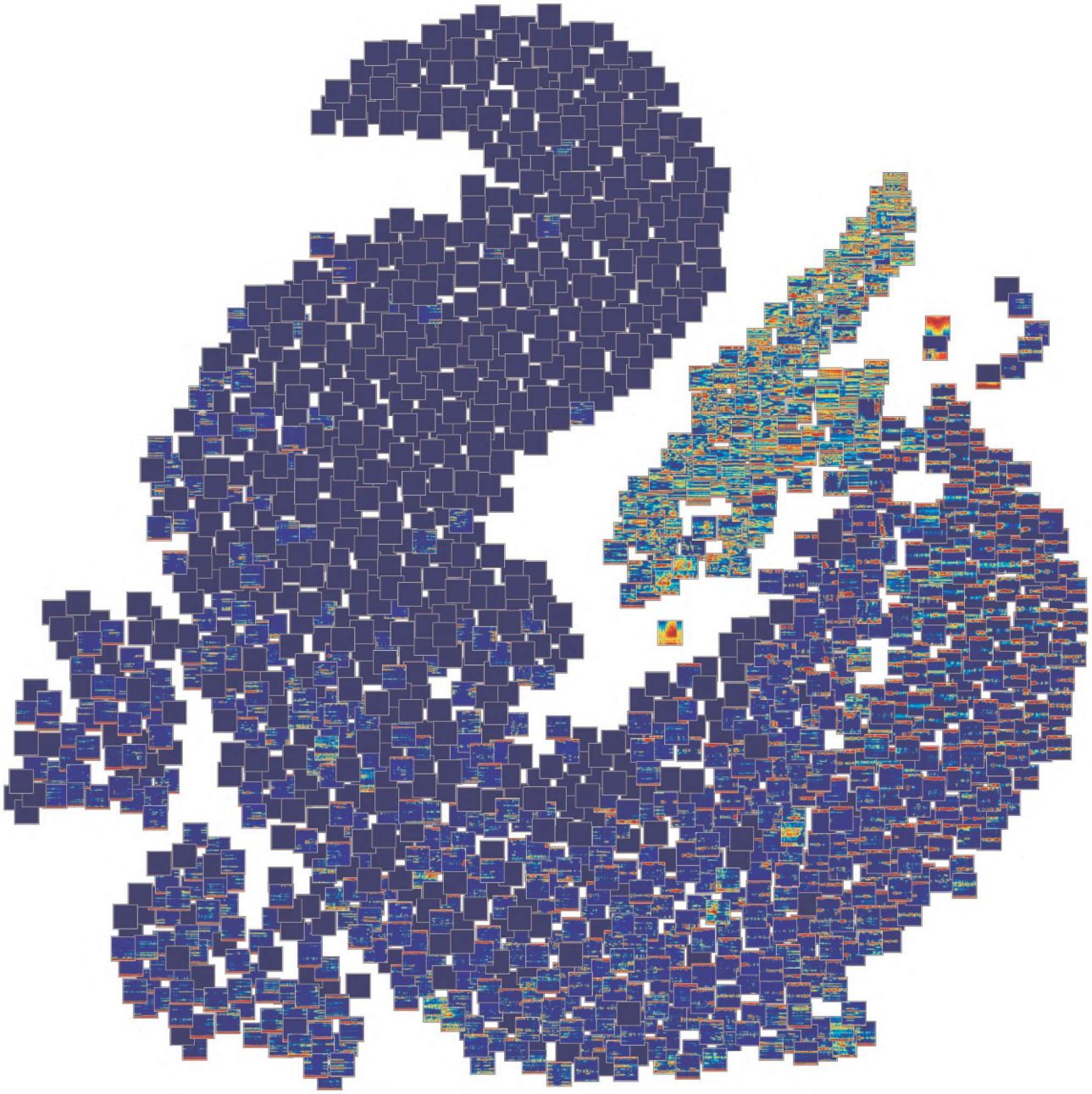


Figure C2.4: t-SNE representation of the Viella dataset obtained with the image-BYOL latent space. The thumbnails correspond to the EB representation of the windowed signal.

C3 Several Examples of Cataloged Classes Using the Clusters Produced with K-Means and Agglomerative Clustering

Appendix C3 provides several examples of data block cluster centers for each cataloged class. These examples are obtained using the human-engineered latent space or the image-BYOL latent space. The examples are provided for the Pyrenees and Viella datasets.

1. Figure C3.1: Pyrenees dataset with the human-engineered latent space.
2. Figure C3.2: Pyrenees dataset with the image-BYOL latent space.
3. Figure C3.3: Viella dataset with the human-engineered latent space.
4. Figure C3.4: Viella dataset with the image-BYOL latent space.

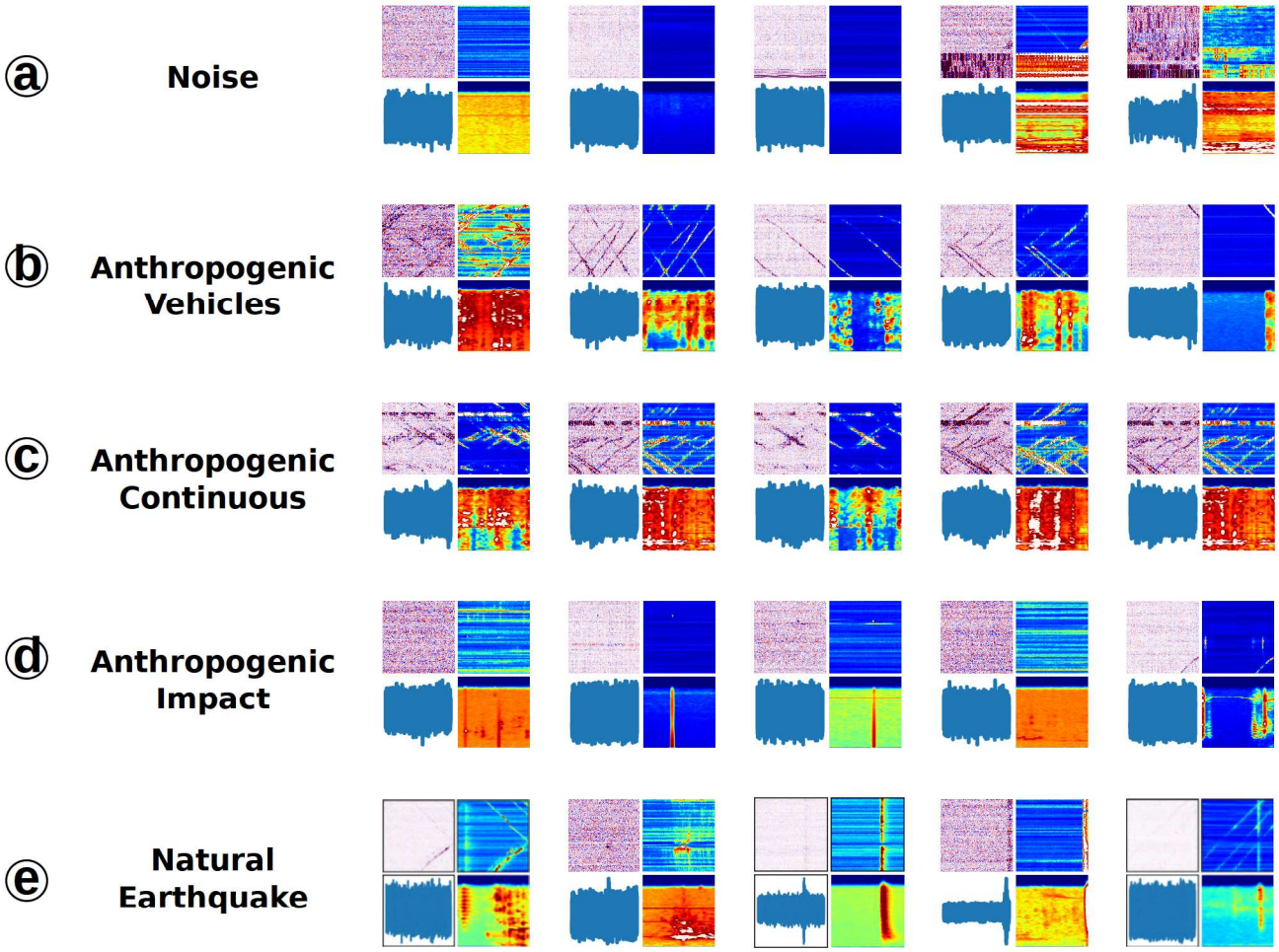


Figure C3.1: Cataloged cluster for the Pyrenees dataset, with clustering obtained using the human-engineered latent space. 5 classes are distinghuishable: ***(a)*** *noise,* ***(b)*** *anthropogenic vehicles,* ***(c)*** *anthropogenic continuous events,* ***(d)*** *anthropogenic impact events, and* ***(e)*** *earthquakes or quarry blasts.*

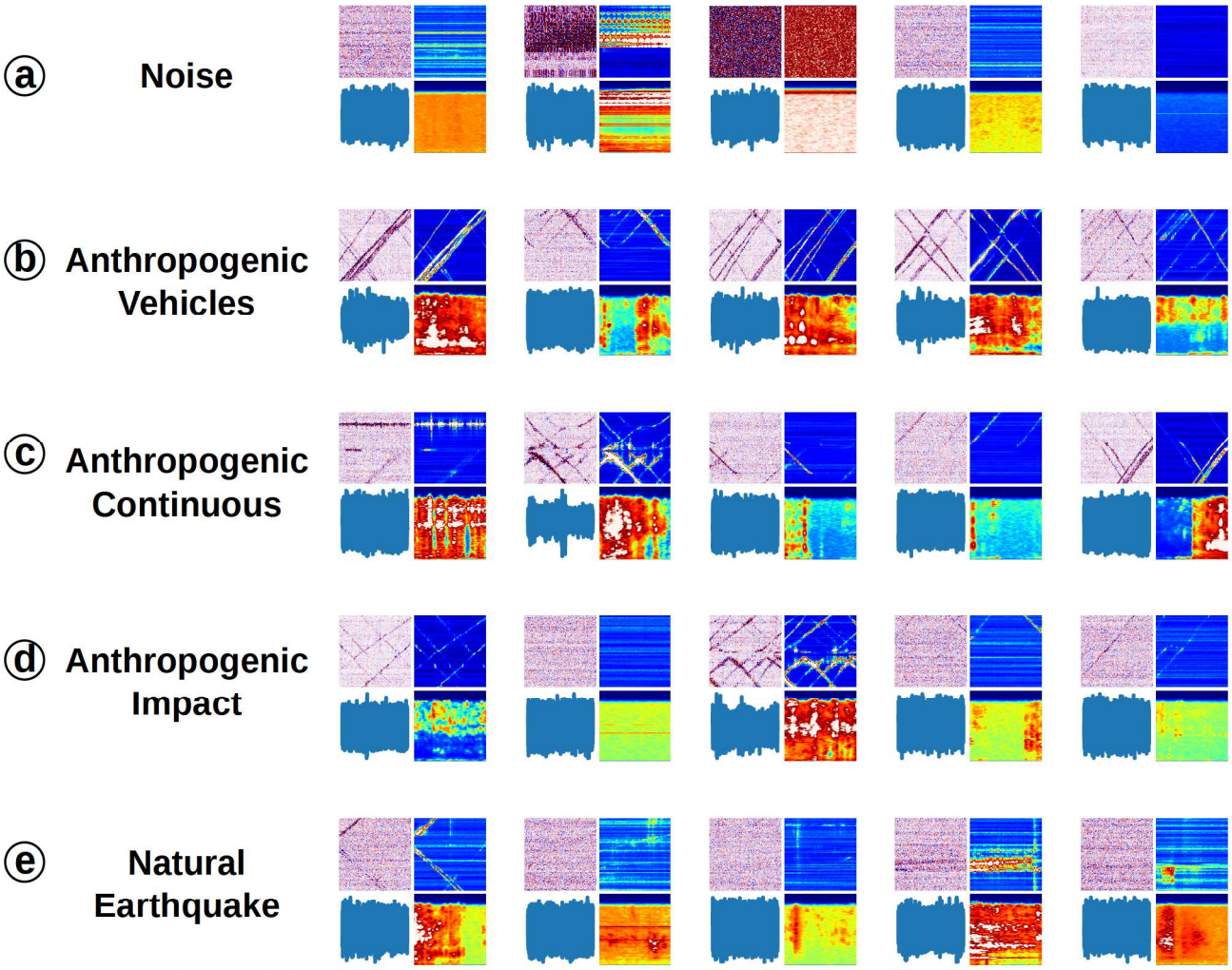


FIGURE C3.2: Cataloged cluster for the Pyrenees dataset, with clustering obtained using the image-BYOL latent space. 5 classes are distinghuishable: ***(a)*** *noise,* ***(b)*** *anthropogenic vehicles,* ***(c)*** *anthropogenic continuous events,* ***(d)*** *anthropogenic impact events, and* ***(e)*** *earthquakes or quarry blasts.*

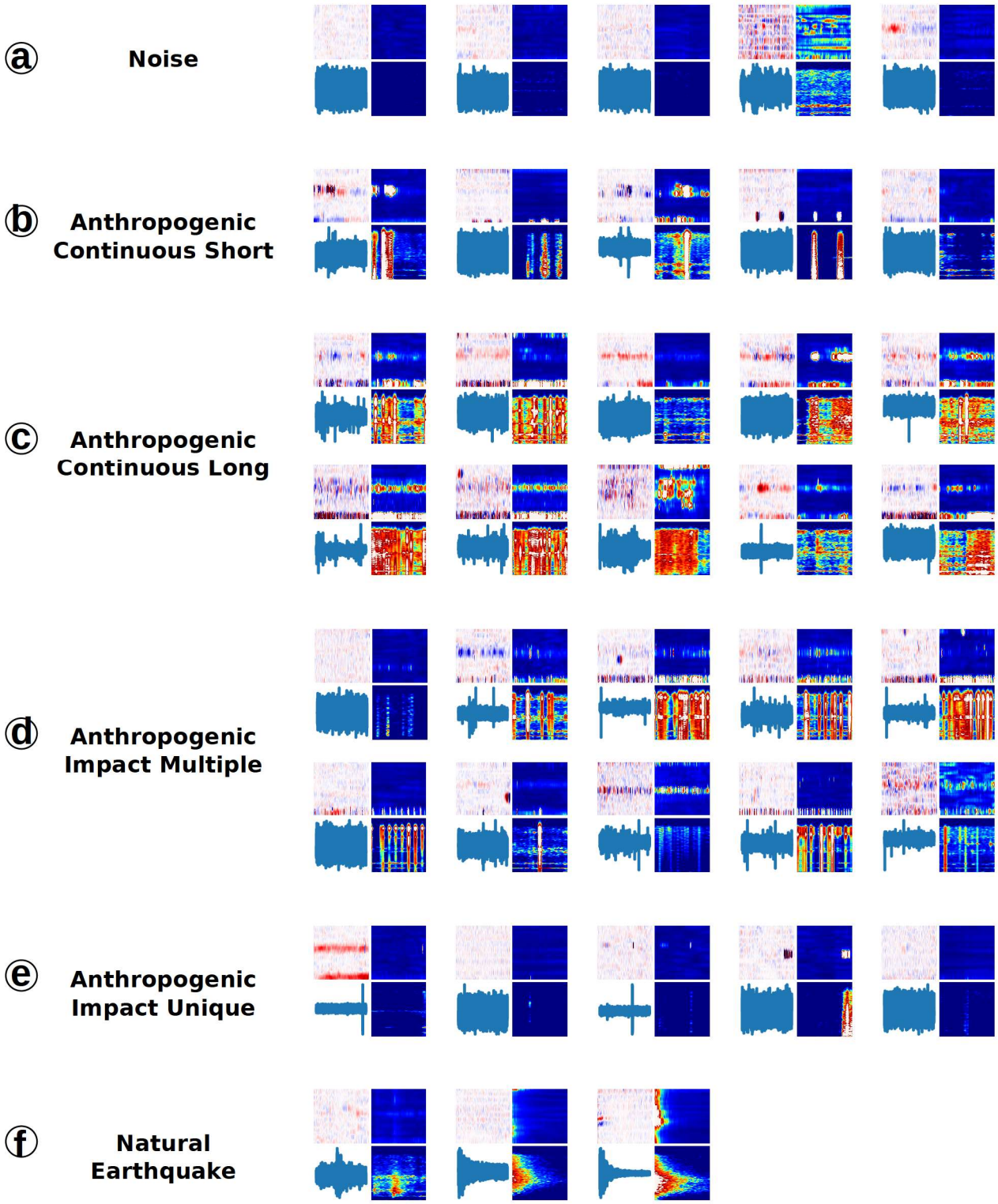


Figure C3.3: Cataloged cluster for the Viella dataset, with clustering obtained using the human-engineered latent space. 5 classes are distinghuishable: ***(a)*** *noise,* ***(b)*** *anthropogenic vehicles,* ***(c)*** *anthropogenic continuous events,* ***(d)*** *anthropogenic impact events, and* ***(e)*** *earthquake.*

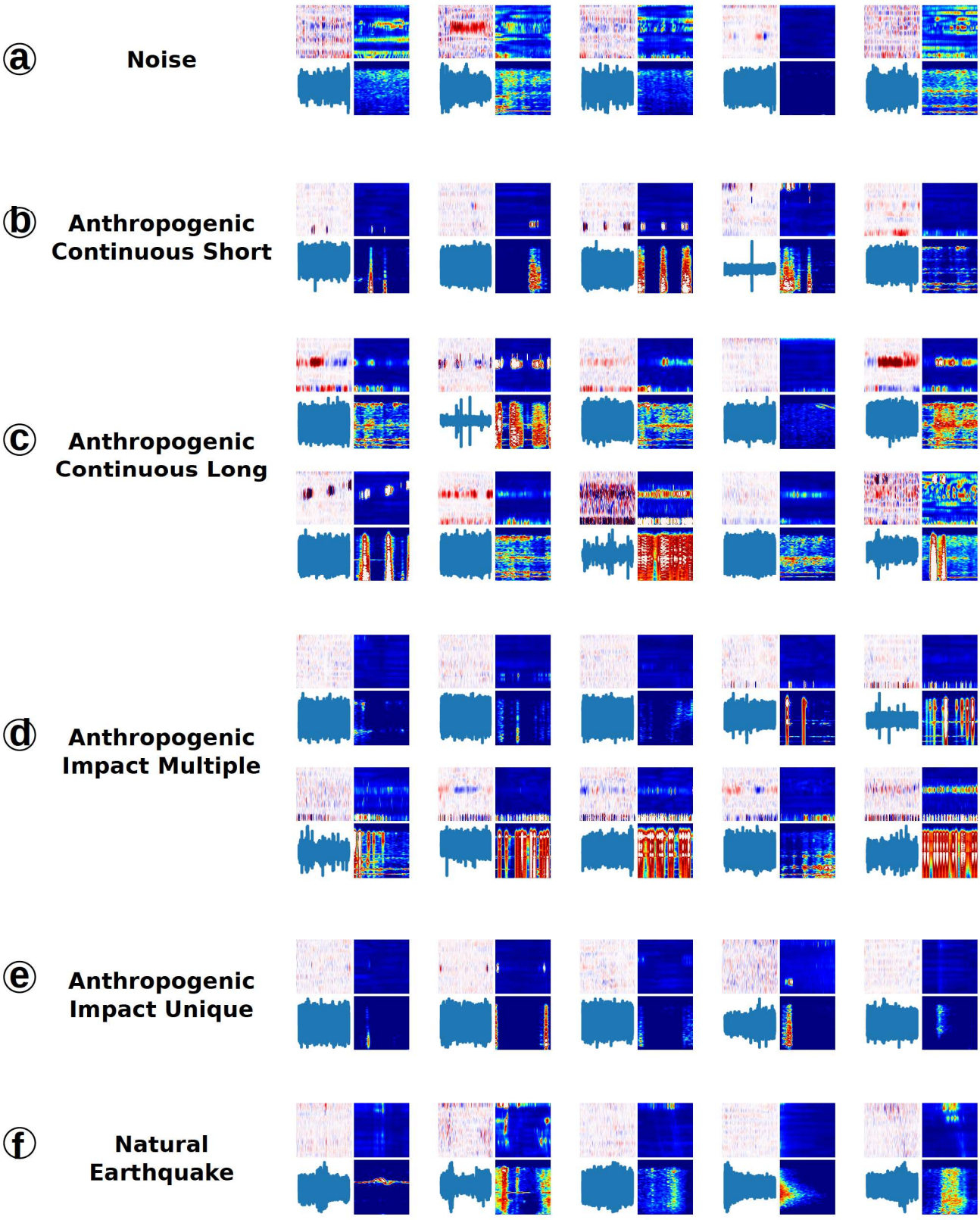


FIGURE C3.4: *Cataloged cluster for the Viella dataset, with clustering obtained using the image-BYOL latent space. 5 classes are distinghuishable:* ***(a)*** *noise,* ***(b)*** *anthropogenic vehicles,* ***(c)*** *anthropogenic continuous events,* ***(d)*** *anthropogenic impact events, and* ***(e)*** *earthquake.*

Résumé Etendu de Thèse de Doctorat

Camille Huynh

Soutenance du 11 Février 2025

1 Introduction

1.1 Composition et familles d'instruments sismologiques

Un instrument sismologique se compose généralement de deux éléments essentiels : un **capteur** et un système de traitement des données (appelé aussi **numériseur**). Le capteur a pour rôle de détecter les variations des propriétés physiques du sol, telles que les vibrations. Ces signaux, pouvant être sous différentes formes (optique, mécaniques, électriques, etc.) sont ensuite traités par le système de traitement des données, qui transforme ces signaux en informations exploitables, comme les amplitudes et les fréquences des vibrations du sol. Un instrument sismologique est défini comme passif s'il n'est pas alimenté par une source d'énergie externe, ou actif s'il nécessite une batterie ou une autre source d'énergie. Les sismomètres large bandes, les accéléromètres, les géophones ainsi que les **systèmes de détection acoustique distribuées** (Distrubuted Acoustic Sensing **(DAS)** en anglais) sont parmi les instruments sismologiques les plus utilisés, et entre tous dans la catégorie des instruments actifs. Dans ce résumé, nous nous intéressons en particulier aux systèmes de détection acoustique distribuées.

1.2 Fondamentaux sur la mesure par fibre optique

Dans un système de **détection acoustique distribuée**, le capteur et le numériseur sont dissociés. Le **capteur** est la fibre optique (FO), qui peut s'étendre sur plusieurs dizaines voire centaines de kilomètres, tandis que le **numériseur**, appelé interrogateur DAS, est connecté à une extrémité de la fibre optique (FO). Lors de la mesure, l'interrogateur émet des impulsions laser répétées. Ces impulsions se propagent dans la fibre et une fraction de la lumière est continuellement rétrodiffusée en raison des impuretés naturelles du verre. Lorsqu'une perturbation, comme une vibration, affecte la fibre, les impuretés se déplacent légèrement, modifiant la phase de la lumière rétrodiffusée. En analysant cette rétrodiffusion de Rayleigh, nous pouvons extraire le **taux de déformation** (Strain Rate, SR).

Le calcul du SR à partir des données optiques repose sur deux paramètres : la **longueur de jauge (Gauge Length, GL)** et le **temps de dérivation (Derivation Time, DT)**. La GL détermine la distance sur laquelle les déformations sont moyennées, affectant la résolution et le bruit : une GL plus courte détecte des déformations sur de courtes distances mais augmente le bruit, tandis qu'une GL plus longue améliore la stabilité du signal mais réduit la résolution. En pratique, GL est choisie pour dépasser la moitié de la plus courte longueur d'onde sismique anticipée (Dean et al., 2017). La DT, quant à lui, détermine la capacité à capturer des ondes sismiques à haute fréquence : un DT plus court permet de détecter des ondes plus rapides, mais avec un rapport signal sur bruit (SNR) moins favorable. En sismologie, DT détermine la fréquence maximale mesurable, avec une fréquence de coupure correspondant à l'inverse de DT. Le choix de la DT doit aussi prendre en compte que l'atténuation commence à distordre le signal à une fréquence plus faible appelé fréquence de coin.

1.3 Quelques exemples de sources sismogéniques

Les instruments FO-DAS sont sensible aux ondes sismiques émises par différentes sources. Dans l'état de l'art, nous pouvons citer des applications pour la détection des événements sismiques suivants :

- **Les tremblements de terre**, ou **séismes**, représentent les événements sismogéniques naturels les plus fréquemment observés. Selon leur profondeurs et leur puissance, ils peuvent être détectés jusqu'à plusieurs miliers de km de l'épicentre (Engdahl et al. 1998).
- **Les événements sismogéniques d'origine volcanique** peuvent s'exprimer sous plusieurs formes. En effet, les signaux sismiques peuvent provenir de sources différentes, allant de l'intrusion du magma dans la roche et sa fracturation pour les événements volcano-tectoniques (Aki et al. 1977), la libération de gaz et de magma par le volcan pour les explosions volcaniques, le mouvement continu du magma proche de la surface pour les tremors volcaniques ou encore la génération de signaux hydroacoustiques pour le volcanisme sous-marin.
- **Les événements sismogéniques d'origine environnementale** regroupent les autres sources sismogéniques, tels que celle générées par les glissements de terrain, les mouvements glaciaires et les vibrations causées par des facteurs météorologiques ou les changements climatiques. Pour l'étude des grands événements, des réseaux sismiques permanents ou semi-permanents programmables peuvent être utilisés (Aster et al. 2017 ; Whiteley et al. 2019). Pour des sites spécifiques, tels qu'un glissement de terrain, un glacier, une rivière ou un bassin versant, une surveillance in situ peut être réalisée à l'aide de réseaux de géophones portables (Teja et al. 2014 ; Martinez et al. 2017 ; Provost et al. 2018). Les instruments FO-DAS sont adaptés aux deux approches, car ils offrent des données continues en temps réel avec une haute résolution spatiale.
- **Les événements sismogéniques d'origine anthropique** contiennent tous les événements d'origine humaine. Ces sources peuvent être liés aux activités industrielles (extraction de blocs de pierre par tirs de carrière, fracturation hydraulique, géothermie, etc.) ou aux activités quotidiennes (circulation automobile, ferrée, déplacement de piétons, travaux de construction tiers, survol d'hélicoptères, etc.). Bien que non systématique, ces sources se distinguent souvent par la présence de fréquences harmoniques caractéristiques dans leur comportement fréquentiel.

1.4 Avantages et inconvénients identifiés de la mesure FO-DAS

Les différentes études sismiques menées avec l'instrument FO-DAS mettent en évidence son potentiel d'application, tout en soulignant également certaines de ses limites. Parmi les avantages, nous pouvons citer la **haute résolution spatiale des mesures fibre**, utile par exemple dans l'observation de la microsismicité dans différents contexte géophysiques (Williams et al. 2019, Klaasen et al. 2021), la **réutilisabilité de la fibre télécom pour de la mesure sismique** en particulier dans le cas d'observation post-séisme (Li et al. 2021), l'utilisation de fibre optique pour de la **surveillance sismique dans des zones difficiles d'accès**, par exemple pour observer le volcanisme sous-marin (Caudron et al. 2024), ou encore la **résistance de la fibre optique aux conditions extrêmes**, très utile pour l'observation des volcans pour la surveillance des coulées pyroclastiques (Métaxian et al. 2024).

Parmi les limites souvent relevées, la **sensibilité unidirectionnelle** est citée, en particulier pour la réalisation de profils sismiques verticaux pour la surveillance de puits (Kuvshinov 2016, Wuestefeld et al. 2019). Le **volume de donnée** générée représente également un challenge pour des mesures sur de longues périodes, par exemple lors de l'observation de glaciers (Hudson et al. 2021). Le **couplage de la fibre optique avec le sol** est également important afin d'assurer une bonne sensibilité des mesures sismiques, notamment parce que le **niveau de bruit est plus important** que lors de mesures avec géophones.

1.5 Les techniques de traitement de signal

Le **bruit instrumental** désigne les facteurs environnementaux externes qui perturbent le fonctionnement normal de l'instrument, notamment les fluctuations électriques et les vibrations sur l'équipement. Pour les données FO-DAS, sa réduction repose principalement sur la technique de **réduction de mode commun**, qui consiste à soustraire la moyenne du signal calculé le long de la fibre à chaque instant de mesure. Cette méthode permet d'atténuer les effets de bruit systématique et de mieux isoler les signaux d'intérêt.

Afin de détecter des événements, il existe différentes méthodes de traitement de signal. Parmi les plus utilisées, nous retrouvons les méthodes suivantes :

- **STA/LTA** : La technique STA/LTA consiste à comparer la moyenne du signal sur une fenêtre temporelle courte avec celle d'une fenêtre beaucoup plus longue, permettant ainsi de mettre en évidence des événements transitoires en détectant des changements dans le rapport STA/LTA. En l'absence d'événement, ce rapport est d'environ 1. Lors du début d'un événement, il devient supérieur à 1, et à la fin de l'événement, il tend à descendre en dessous de 1.
- **Data stacking** (ou méthode d'empilement des signaux) : La méthode d'empilement consiste à moyenner plusieurs traces sismiques du même événement provenant de canaux différents dans le cas des données FO-DAS. Cette méthode permet de réduire le bruit aléatoire dans la donnée et d'améliorer la détection d'événements faibles, comme les petits séismes ou les signaux environnementaux.
- **Filtrage** : Les techniques de filtrage permettent d'améliorer la qualité des données sismiques en isolant certaines plages de fréquences.
- **Template matching** : Le template matching permet de comparer un signal sismique avec un signal de référence afin d'identifier des motifs similaires.
- **Spectre** : Le spectre est une représentation du signal sismique qui montre comment son énergie ou sa puissance se répartit sur différentes fréquences. Il permet d'identifier les composants harmoniques et les motifs caractéristiques des sources sismogènes, et peut être utilisé pour distinguer différents événements tels que des séismes, des explosions ou du bruit anthropique.
- **Spectrogramme** : Le spectrogramme est une représentation de l'énergie ou de la puissance du signal suivant les axes temps-fréquence qui montre comment le contenu fréquentiel du signal évolue dans le temps. Il est utile pour détecter des signaux transitoires comme les séismes et suivre leur évolution en identifiant l'origine, la durée et le contenu fréquentiel des événements sismogènes.
- **Bande d'énergie (EB)** : La bande d'énergie (EB) d'un signal sismique correspond à l'intégration du spectre sur une plage de fréquences fixe. Cette analyse permet de détecter certaines sources sismiques caractérisées par un contenu fréquentiel particulier, ainsi que d'évaluer l'énergie produite par la source.

Les outils classiques d'exploration des données, tels que STA/LTA, le stacking, le filtrage, et l'analyse spectrale, sont utiles pour détecter et caractériser les événements sismogènes, mais nécessitent souvent une combinaison avec d'autres techniques et un ajustement fin des paramètres pour obtenir des résultats fiables. L'**intelligence Artificielle (IA)** permet d'intégrer ces outils de manière flexible et automatisée, en utilisant des paramètres définis par l'utilisateur, ce qui améliore la précision de l'analyse sismique tout en réduisant l'intervention manuelle.

1.6 L'Intelligence Artificielle

1.6.1 Généralités

L'intelligence artificielle (IA) repose sur des modèles statistiques capables d'apprendre à partir de données, en identifiant par exemple des motifs, pour effectuer des prédictions ou prendre des décisions. Ces modèles s'appuient sur un **espace latent**, une représentation intermédiaire construite à partir d'**attributs** (features en anglais). Ces attributs peuvent être soit définies manuellement, comme en **Machine Learning** (ML), soit apprises automatiquement, comme en **Deep Learning** (DL). Le DL, grâce à l'utilisation de réseaux neuronaux profonds, automatise la construction de l'espace latent en transformant progressivement les données brutes dans ses couches cachées pour extraire des attributs pertinentes pour la tâche visée. Cette approche nécessite cependant un grand volume de données pour l'entraînement et rend l'interprétation des résultats plus difficile.

L'intelligence artificielle (IA) regroupe des algorithmes adaptés à quatre grandes catégories de tâches : la prédiction, l'exploration, l'auto-apprentissage et les tâches basées sur l'interaction. Pour la **prédiction**, les algorithmes supervisés, comme les forêts aléatoires ou les machines à vecteurs de support, utilisent des données étiquetées pour effectuer des tâches de régression ou de classification. Les approches semi-supervisées et le transfert de connaissances permettent d'exploiter des données partiellement étiquetées ou des modèles pré-entraînés. L'**exploration** utilise des méthodes non supervisées, telles que le clustering (K-means, clustering hiérarchique, DBSCAN) ou la réduction de dimensionnalité (Analyse en Composante Principale, t-SNE), pour simplifier les représentations des données. L'**auto-apprentissage** (self-supervised learning en anglais) crée des pseudo-étiquettes pour extraire à partir de la donnée brute des caractéristiques complexes de façon autonome. Dans cette catégorie de modèle, nous pouvons citer BERT pour le traitement du langage naturel ou BYOL pour la construction d'espaces latents. Enfin, les tâches d'**interaction** sont apprises grâce à l'apprentissage par renforcement, dans lequel les modèles s'améliorent en interagissant avec leur environnement et en ajustant leurs actions en fonction des récompenses ou des pénalités reçues.

1.6.2 Exemples d'application en sismologie

L'IA trouve de nombreuses applications en sismologie et pour les données FO-DAS, notamment dans les domaines du prétraitement, de la classification et de l'exploration des données. En **prétraitement**, des approches comme les réseaux de neurones convolutifs (CNN) sont utilisées pour **réduire le bruit** et améliorer le rapport signal/bruit dans les données FO-DAS (Yang et al. 2023). Des techniques d'apprentissage auto-supervisé, telles que celles exploitant plusieurs câbles à fibre optique (Lapins et al. 2024), offrent également des solutions innovantes pour le débruitage. Pour la **détection des événements sismiques**, des modèles supervisés comme PhaseNet (Zhu et al. 2019) identifient automatiquement les phases P et S, tandis que des approches non supervisées, telles que le clustering flou, permettent de détecter des micro-séismes en présence de bruit de fond important (Chen et al. 2020).

L'**exploration des données** par IA s'illustre par des techniques non supervisées, qui identifient des motifs cachés dans les signaux sismiques. En volcanologie, ces méthodes sont appliquées pour classifier les phases d'activité volcanique (Hammer et al. 2012), étudier les tremblements liés à la viscosité du magma (Unglert et al. 2017), et regrouper les événements similaires pour surveiller les processus sismiques ou volcaniques (Yoon et al. 2015 ; Cesca et al. 2020). Des approches auto-supervisées permettent également d'identifier de nouvelles séquences dans les processus éruptifs (Rimpot et al. 2025). Dans le cas des glissements de terrain, le clustering aide à identifier des comportements sismiques spécifiques, parfois précurseurs d'effondrements (Seydoux et al. 2020). Pour les données FO-DAS, bien que les études soient encore limitées, des avancées sont réalisées, comme l'utilisation de ces instruments pour analyser les signaux acoustiques d'émissions de bulles volcaniques (Caudron et al. 2024), ouvrant de nouvelles perspectives pour l'analyse et la compréhension

des événements sismogéniques.

1.7 Les challenges et objectifs de la thèse

Appliquée à de la mesure sismique, l'instrument FO-DAS offre plusieurs avantages tels qu'une haute résolution spatiale, un coût réduit, une facilité de déploiement et une grande résilience dans des conditions environnementales difficiles. Cependant, des défis demeurent, notamment les volumes massifs de données générées, le bruit de fond élevé et la sensibilité à diverses sources de bruit, ce qui complique son utilisation dans certains contextes de surveillance. Des solutions permettent de corriger partiellement ces défauts, par exemple le traitement du signal pour réduire le bruit, l'analyse des formes d'onde et des fréquences pour identifier les événements, ou encore l'utilisation d'algorithmes d'intelligence artificielle pour la sélection des sources et la gestion du volume de données. Dans nos travaux, nous souhaitons adopter une approche Machine Learning (ML) afin de travailler avec des modèles facilement explicables. Les différents challenges soulevés permettent d'identifier trois questions scientifiques :

- **Comment les méthodes de traitement existantes peuvent-elles être adaptées pour gérer les données en temps réel tout en prenant en compte la nature spatiale distribuée des données FO-DAS dans la chaîne de traitement ?**
- **Comment intégrer les caractéristiques uniques des données FO-DAS dans les algorithmes d'apprentissage automatique, notamment l'analyse des formes d'onde spatiales des données sismiques ?**
- **Comment créer plus efficacement des ensembles de données pour l'entraînement des modèles de machine learning, surtout lorsqu'il s'agit de traiter de grands volumes de données continues couvrant plusieurs saisons ou années ?**

Afin de répondre à ces questions, nous formulons trois hypothèses :

- **H0** : Nous préférons utiliser les techniques de ML plutôt que de DL pour l'interprétabilité des modèles.
- **H1** : Nous avons choisi d'adopter une approche de flux de données, en construisant nos chaînes de traitement autour de ce concept.
- **H2** : Bien que nous nous concentrions sur les applications en temps réel, nous privilégions les questions concernant les avantages de l'utilisation de nouveaux attributs dans le processus de classification, plutôt que la réduction du temps de calcul de ces attributs. Nous permettons également le traitement des données sur des systèmes de calcul haute performance (HPC).
- **H3** : Pour l'étiquetage à long terme, nous visons à éviter l'étiquetage manuel événement par événement. Cependant, nous autorisons une intervention sur quelques événements ou groupes d'événements dans le processus d'étiquetage afin de garantir la qualité des données d'entraînement.

1.8 Plan de thèse

- **Solution 1 : Chaîne de traitement pour la détection et classification des événements en temps réel**
  Nous développons une chaîne de traitement complète des données FO-DAS pour la détection et la classification des événements en temps réel. Nous utilisons des techniques de ML issues de travaux réalisés en sismologie conventionnelle. La chaîne inclut en plus un traitement de la cohérence spatiale des données, utilisant des champs aléatoires de Markov pour prendre en compte les dépendances spatiales. Nous testons la chaîne de traitement sur un jeu de données

du centre d'essai FEBUS Optics, contenant des événements anthropiques. Cela est abordé dans la Section 2.

— **Solution 2 : Intégration des attributs spatiaux dans le processus ML**
Dans cette étape, nous incorporons les attributs spatiaux des données FO-DAS et les relations entre les capteurs virtuels dans le modèle ML. Nous appliquons cette méthode sur un jeu de données DAS acquis sur le terrain. Le jeu de données consiste en des données provenant d'une fibre optique de 91 km dans les Hautes-Pyrénées, contenant à la fois des événements sismogènes naturels et des événements anthropiques. Cela est discuté dans la Section 3.

— **Solution 3 : Optimisation de l'étiquetage des ensembles de données**
Nous concevons une méthode pour simplifier la création d'ensembles de données d'entraînement pour l'acquisition FO-DAS. Cela implique le regroupement des données FO-DAS similaires, réduisant ainsi le besoin d'un étiquetage manuel événement par événement. Nous testons cette méthode sur un jeu de données acquis dans les Hautes-Pyrénées et sur le glissement de terrain de Viella, comprenant des événements sismogènes naturels, des événements anthropiques liés à l'agriculture quotidienne, et d'autres événements anthropiques. Cela est présenté dans la Section 4.

2 Classification de Données Acquis par Fibres Optique à l'Aide d'Attributs Issues de Travaux Menés en Sismologie Conventionnelle

2.1 Introduction

L'utilisation des capteurs distribuées acoustiques à fibre optique (Distributed Acoustic Sensing, DAS) pour la surveillance sismique a ouvert de nouvelles perspectives pour l'observation des phénomènes naturels et anthropiques. Cependant, l'analyse automatique des données FO-DAS reste un défi pour plusieurs raisons : d'abord en raison du volume des données générées, puisqu'on dispose virtuellement d'un grand ensemble de capteurs ponctuels, mais aussi parce que les données sismiques acquises avec les instruments FO-DAS sont plus bruitées que leur équivalent acquis avec des géophones. Dans ce contexte, il est nécessaire de développer une chaîne de traitement spécifique. Un des éléments-clé concerne l'utilisation d'attributs adaptés dans le but d'appliquer des algorithmes de Machine Learning (ML).

Dans cette section, nous présentons notre chaîne de traitement, qui utilise des attributs ayant permis de classer des signaux sismiques provenant de sismomètres conventionnels (Hibert et al. 2014 ; Hibert, Provost et al. 2017 ; Hibert et al. 2019 ; Maggi et al. 2017 ; Provost et al. 2017 ; Chmiel et al. 2021 ; Domel et al. 2023). Nous introduisons également une étape de pré-traitement avant l'utilisation du ML afin de préparer la donnée et une étape de post-traitement suite au ML utilisant un algorithme de Markov Random Field (MRF) pour améliorer la classification. Cette dernière étape utilise la redondance spatiale des signaux sismiques pour maintenir la cohérence de la classification et réduire l'impact du bruit. Nous évaluons les performances de cette chaîne de traitement sur des données collectées au centre d'essai FEBUS Optics dans un environnement contrôlé, en particulier des événements sismiques générés au-dessus une tranchée de 22 mètres de long.

2.2 Données

Les données utilisées pour cette étude ont été collectées dans un environnement contrôlé au centre d'essais FEBUS Optics. Il s'agit d'une tranchée de 22 m dans laquelle est installée un câble à fibre optique effectuant des aller-retours dans différentes configurations : différentes profondeurs, différents types de fibres (mono-mode, multi-mode), différents couplage entre la fibre et le câble (loose/tight), et avec ou sans protection pour la gaine de câble.

Pour notre simulation, les signaux sont mesurées et filtrées par un filtre passe-bas avec une fréquence de coupure de 100 Hz et un sous-échantillonnage. La GL est choisie égale à 5 m avec un échantillonnage spatial de 80 cm. La base de données comprend six types de sources sismiques d'événements anthropiques, constitué chacun d'un certain nombre d'événements :

- **Mouvement piéton** : Marche d'une personne le long de la fibre. 15 occurrences.
- **Impact ou chute d'objets** : Chute d'objets ou impact généré avec un marteau sur un bloc en métal. 48 occurrences.
- **Compacteur** : Fonctionnement d'un compacteur à proximité de la fibre. 5 occurences.
- **Pelle mécanique** : Mouvements d'une pelle mécanique. 10 occurences.
- **Fuite d'eau/d'air** : Simulation d'une fuite à haute pression. 11 occurences.
- **Bruit** : Absence d'événement sismique d'intérêt.

Parce que chacun de ces événements est mesuré par différentes types de fibres répartis dans la tranchée, et que chacun des événements mesurés est également découpé par fenêtre temporelle de 4 s avec un recouvrement de 50%, le nombre de signaux constituant la base de donnée (appelé unité de

signal) est bien plus importante que le nombre d'occurrence simulé (appelé nombre d'événement). La Table 1 donne ces valeurs.

Table 1 – Occurrences d'événements pour chaque classe considérée dans la base de données.

Type d'événement	Nombre d'événements*	Nombre de sous-événements†	Nombre d'unités de signal‡
Marche de piéton	15	244	254 129
Impacts ou objets tombants	48	1,017	344 609
Pelle mécanique	10	195	220 313
Compacteur	5	114	57 516
Fuites	11	239	548 185

* Un événement regroupe tous les signaux émis par une même source sismique.
† Un sous-événement regroupe tous les signaux mesurés sur une même portion de fibre de 22 m. ‡ Une unité de signal correspond à une fenêtre de 4 s prise en un point de la fibre.

2.3 Méthodes

La chaîne de traitement développée se compose de trois étapes principales : **pré-traitement**, **classification**, et **post-traitement**. La Figure 1 représente la chaîne de traitement avec les étapes de classification et de post-traitement.

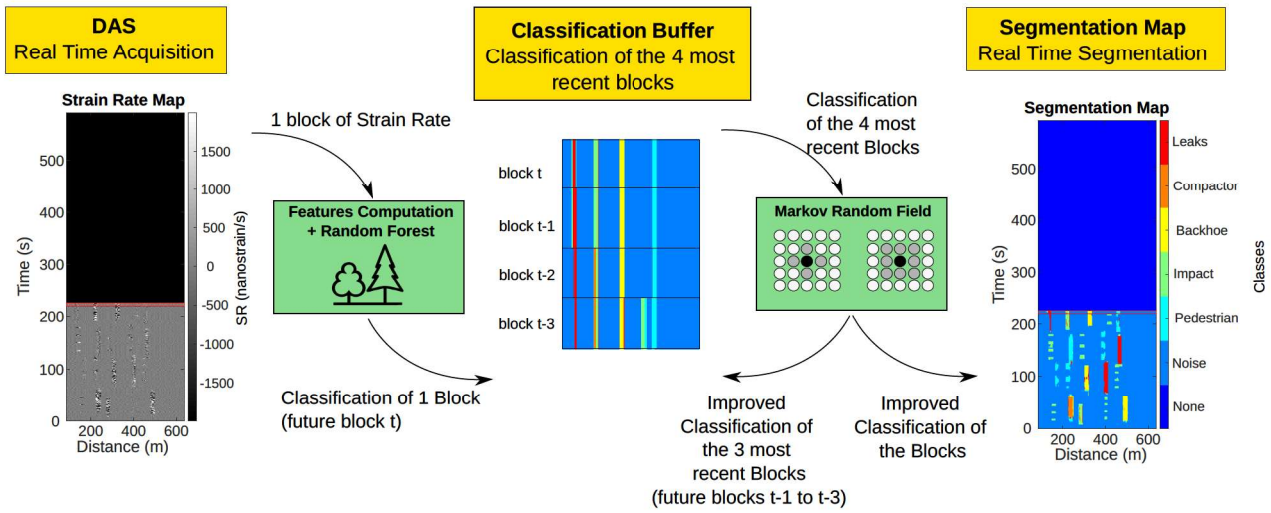


Figure 1 – Chaîne de traitement.

2.3.1 Pré-traitement

Le pré-traitement vise à améliorer la qualité des données brutes en éliminant les artefacts instrumentaux et en découpant les signaux mesurées à l'aide de fenêtres temporelles adaptées pour l'analyse en flux. Deux opérations importantes sont effectuées :

– **Suppression du mode commun** : Cette étape élimine les perturbations globales affectant l'ensemble de la fibre, comme les vibrations induites sur le système d'acquisition ou des artefacts dus au système d'acquisition.

— **Découpage en flux** : Les données sont divisées en fenêtres temporelles de taille fixe, afin d'imiter l'acquisition des données FO-DAS en temps quasi réel. Chaque fenêtre est analysée indépendamment par l'algorithme de classification.

2.3.2 Classification

La classification repose sur l'extraction d'attributs temporels dérivés de la littérature en sismologie conventionnelle (Bessason et al. 2007, Provost et al. 2017, Hibert et al. 2019). Ces 53 attributs peuvent être regroupés en 3 catégories :

— **Descripteurs temporels** : Par exemple la durée du signal, l'amplitude maximale, l'énergie cumulée.

— **Descripteurs fréquentiels** : Le contenu spectral, par exemple la fréquence dominante dans le signal, l'énergie entre différentes bandes de fréquences.

— **Descripteurs de l'évolution fréquentielle en fonction du temps** : Le changement de contenu spectral en fonction du temps, par exemple la variation de la fréquence maximale, la moyenne ou la médiane en fonction du temps.

Ces attributs sont utilisées pour construire un espace latent représentant chacune des fenêtres de données. Cet espace est ensuite utilisé en entrée de l'algorithme de classification supervisée appelé forêt aléatoire (Breiman, 2001) (Random Forest (RF) en anglais), qui attribuent une classe à chaque fenêtre. RF se base sur l'utilisation d'un grand nombre d'arbre de décision utilisés en parallèle, et dans lequel chaque arbre est entraîné sur un jeu de donnée réduit et avec des attributs sélectionnés aléatoirement. RF est capable de fonctionner malgré la présence d'attributs peu significatifs dans le processus de classification, ainsi que de quantifier l'importance de chacun des attributs. RF est aussi utilisé pour sa capacité à estimer un score de classification pour chaque donnée présentée en entrée. Le jeu de donnée est divisé en un jeu d'entraînement et un jeu de test, chacun contenant 50% du dataset initial, et le jeu de donnée d'entraînement est augmenté avec l'algorithme SMOTE (Chawla et al. 2002) afin d'équilibrer le nombre d'unité de signal dans chaque classe.

2.3.3 Post-traitement

Le post-traitement vise à intégrer la cohérence spatiale des données FO-DAS en considérant l'existence d'une redondance d'information entre des canaux adjacents. L'algorithme Markov Random Field (MRF) permet de considérer deux paramètres : le score de classification fourni en sortie de RF et la composition de la classification des canaux adjacents.

2.4 Résultats

Les résultats obtenus montrent une précision globale de classification de 87 %, avec une amélioration de 4 % grâce à l'intégration de la redondance spatiale via le MRF (précision initiale de 83 %). Les performances varient légèrement selon les classes, les fuites d'eau et d'air étant les plus facilement détectées en raison de leur signature acoustique distincte, tandis que les mouvements de piétons et les impacts mécaniques sont plus difficiles à différencier. Ces résultats illustrent l'efficacité de l'approche dans un environnement contrôlé. Les tests effectués sur des données simulées montrent également que le système est capable de traiter les signaux en flux, confirmant son potentiel pour une application en temps réel.

2.5 Discussion et Conclusion

Les résultats obtenus au centre d'essais FEBUS Optics sont prometteurs et montrent que les attributs issues de la sismologie conventionnelle peuvent être adaptées à l'analyse des données FO-

DAS. Cependant, plusieurs limitations doivent être soulignées :

1. **Environnement contrôlé** : La tranchée de 22 mètres et les conditions de test simplifiées ne reflètent pas la complexité des environnements réels, où les événements sismiques se superposent souvent à des bruits environnementaux variés.
2. **Non-exhaustivité des attributs** : Les 53 attributs utilisées ne capturent que partiellement la richesse des signaux acquis avec les instruments FO-DAS. Seuls la forme du signal temporel et les relations de redondance d'information avec les canaux adjacents sont considérés. Le forme du signal spatial n'est pas étudiée dans cette étude.

Ces limites mettent en évidence la nécessité de développer des attributs supplémentaires pour représenter la dimension spatiale des signaux FO-DAS et mieux capturer les relations entre canaux. Une approche testée dans des conditions réelles avec une fibre de plus grande longueur est également essentielle pour valider la robustesse de la chaîne de traitement. En conclusion, cette section constitue une première étape importante dans le développement puis le déploiement d'une chaîne de traitement supervisée pour les données FO-DAS. Les résultats obtenus fournissent une base solide pour des travaux futurs, qui porteront en particulier sur l'intégration d'attributs spatiaux et l'évaluation dans des environnements plus complexes.

3 Intégration des Spécificités de la Donnée DAS dans le Processus de Classification

3.1 Introduction

L'analyse des données sismiques collectées à l'aide des instruments FO-DAS présente des défis uniques, principalement en raison de la densité et de la redondance des données générées. Contrairement aux réseaux sismiques conventionnels, où chaque station agit comme un capteur individuel, les données FO-DAS exploitent les fibres optiques comme des réseaux de capteurs uniformément distribués, introduisant une dimension spatiale potentiellement intéressante à exploiter. Bien que les attributs utilisées dans l'étude précédente (Section 2) ait montré son efficacité dans un contexte contrôlé, elles ne capturent pas pleinement les relations spatiales contenues dans les données FO-DAS.

Cette section propose d'intégrer les spécificités des données FO-DAS dans la chaîne de traitement en enrichissant l'étape de classification avec de nouveaux attributs. Nous testons cette approche sur une fibre optique de 91 km déployée dans les Pyrénées, en utilisant des données acquises sur trois semaines et annotées à l'aide du catalogue mis à disposition en ligne sur le site du BCSF-RENASS.

3.2 Données

La fibre optique utilisée pour cette étude s'étend sur une longueur de 91 km dans la région des Hautes-Pyrénées, en France. Cette fibre est un câble de télécommunication standard, enterré à des profondeurs variables, longeant et traversant diverses zones à comportement sismiques particuliers, telles que des routes, des zones urbaines, des champs agricoles, et des formations rocheuses. Cette configuration offre un cadre idéal pour tester l'efficacité des nouveaux attributs dans des environnements divers et complexes.

Les données ont été enregistrées à l'aide d'un interrogateur DAS produit par FEBUS Optics, le FEBUS A1-R DAS, avec la configuration suivante : la GL est fixée à 10 m, et la donnée est acquise avec une fréquence d'échantillonnage de 400 Hz durant trois semaines consécutives, entre le 30 août et le 20 septembre 2022. L'échantillonnage spatial est de 4.8 m, ce qui correspond à l'équivalent de 18 958 canaux le long de la fibre optique. Au total, le volume de ce jeu de donnée est de 40 To.

Les premières observations de ce jeu de donnée indique la présence d'un nombre important d'événements anthropiques le long de la fibre, ce qui complique l'identification visuelle des séismes de faible magnitude. Afin de constituer notre jeu de données de tremblements de terre et de tirs de carrières, nous avons utilisé le service internet mis à disposition par le "Bureau Central Sismologique Français" (BCSF) et le "Reseau National de Surveillance Sismique" (RENASS) qui répertorie les événements sismiques détectés et catalogués par les réseaux de sismomètres présents dans cette zone géographique. Avec l'aide du service fourni par le service internet du BCSF-RENASS, nous avons pu identifier visuellement 13 tremblements de terre de faible magnitude ($0.4 < Mw < 2.4$), ainsi 6 tirs de carrière. Ces 19 événements, contenu dans des enregistrements FO-DAS de 10 minutes chacun, représentent un volume de 206 Go. Ils constituent notre jeu de donnée.

3.3 Méthodes

3.3.1 Identification visuelle des événements

Afin d'identifier visuellement les événements sismiques sur les enregistrements FO-DAS, nous avons utilisé la représentation appelée représentation en bande d'énergie (Energy Band (EB) en anglais). Pour ce jeu de donnée, nous calculons l'intégrale de la transformée de Fourier du signal sismique sur l'ensemble de a plage de fréquence disponible (jusqu'à 100 Hz) et sur une durée de 2 s avec un taux de recouvrement de 75%.

3.3.2 Chaîne de traitement

La chaîne de traitement développée dans cette Section conserve les trois étapes principales décrites à la Section 2 (**pré-traitement**, **classification**, **post-traitement**), mais enrichit l'étape de classification avec des attributs conçus afin de prendre en compte les spécificités spatiales et les relations canal-à-canal des données FO-DAS. La Figure 2 représente la chaîne de traitement avec les étapes de classification et de post-traitement.

Pré-traitement

Le pré-traitement comprend :

- **Suppression du mode commun** : Cette étape élimine les perturbations globales affectant l'ensemble de la fibre, comme les vibrations induites sur le système d'acquisition ou des artefacts dus au système d'acquisition.
- **Découpage en flux** : Les données sont divisées en fenêtres de taille en temps et en distance adapté (t_{win}, d_{win}). Chacune de ces fenêtre sert de base pour le système de classification.

Extraction des attributs et classification

En complément des 57 attributs issues de la sismologie conventionnelle et utilisé dans la Section 2, nous avons introduit 54 nouveaux attributs pour exploiter les spécificités des données FO-DAS. Ces nouveaux attributs entrent dans les deux catégories suivantes :

- **Attributs spatiaux** : Par exemple la moyenne, écart-type et pic de l'amplitude sur des fenêtres spatiales définies.
- **Attributs de similarité** : Analyse de la corrélation croisée entre les canaux adjacents pour capturer les relations trace-à-trace, ainsi que de la fonction de déformation dynamique du temps (Dynamic Time Warping (DTW) en anglais).

Ces 111 attributs permettent de construire un espace latent enrichi plus exhaustif pour caractériser la donnée FO-DAS.

Nous utilisons le modèle de ML supervisé XGBoost (Chen et al. 2016) pour classifier les fenêtres de données dans cet espace latent. XGBoost se base sur l'utilisation d'un grand nombre d'arbre de décision utilisés séquentiellement (boosting), et dans lequel chaque arbre est entraîné sur des données mal classifiées par l'arbre précédent. XGBoost est capable de fonctionner malgré la présence d'attributs peu significatifs dans le processus de classification, ainsi que de quantifier l'importance de chacun des attributs. XGBoost est aussi utilisé pour sa capacité à estimer un score de classification pour chaque donnée présentée en entrée. Les performances de l'algorithme entraîné sont évalués en utilisant la validation croisée de type "leave-one-out" (Leave-One-Out Cross-Validation (LOOCV) en anglais) : nous ne gardons pour le test qu'un événement (parmi les 19 du jeu de donnée) et nous entraînons l'algorithme avec le reste ; et nous répétons cela pour chacun des événements. Chacun des algorithmes entraînés peut être évalué en calculant le score F1.

Post-traitement

Un premier post-traitement consiste à extraire une première carte de segmentation de la donnée en appliquant un seuil sur la carte de score. Le modèle de MRF introduit à la Section 2 a été utilisé suite à ce premier post-traitement pour intégrer des informations de cohérence spatiale et réduire les erreurs de classification.

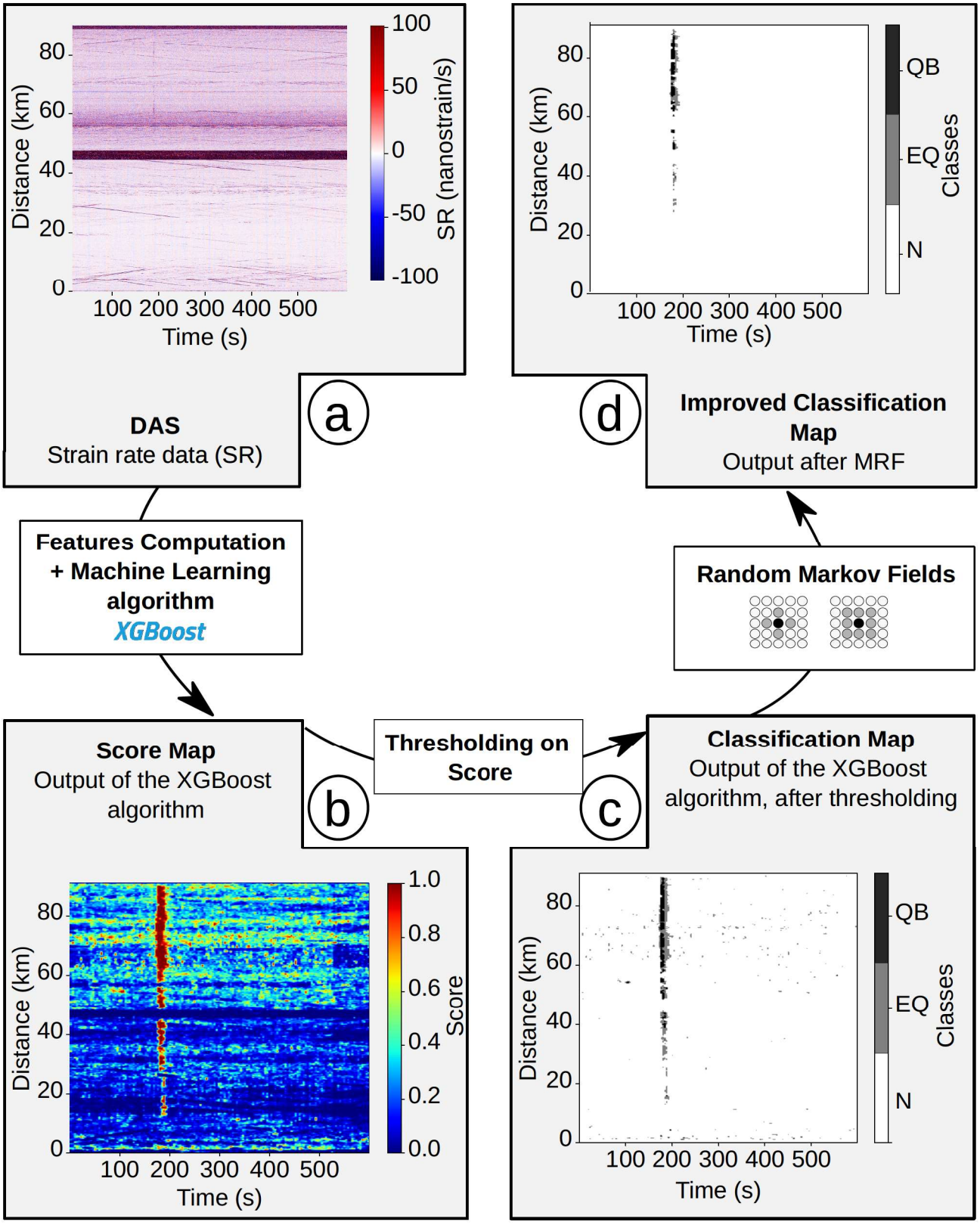


FIGURE 2 – Chaîne de traitement. Les données SR **(a)** sont introduites dans un modèle d'apprentissage automatique, produisant une carte de scores par classe **(b)**. Les points avec une valeur de score supérieure à 0,95 sont conservés et produisent une carte de classification **(c)**. Un algorithme de post-traitement, basé sur les Champs de Markov Aléatoires (MRF), est appliqué sur la carte de classification pour réduire le bruit **(d)**. N fait référence à la classe de bruit, EQ aux tremblements de terre, et QB aux explosions de carrière. Pour simplifier la représentation, la carte de scores représentée ici correspond à la somme des classes non bruitées.

3.4 Résultats

3.4.1 Choix de la taille de fenêtre pour le découpage en flux de la donnée FO-DAS

Afin de choisir la dimension des fenêtres pour la construction des données en flux (t_{win}, d_{win}), nous comparons le score F1 pour plusieurs couples de taille de fenêtre (t_{win}, d_{win}). En effectuant pour l'ensemble des couples formés avec les longueurs de fenêtres [100, 250, 500, 750, 1000, 1250, 1500, 2000] m et durées de fenêtres [4, 8, 12, 16, 20] s, nous trouvons que le meilleur couple (t_{win}, d_{win}) est (1000 m, 8 s), avec lequel nous obtenons un score F1 de 0.69.

3.4.2 Performance de la classification

En utilisant la méthode de validation croisée LOOCV, et parmi les 19 événements repérés, le système est capable de détecter l'ensemble des 13 tremblements de terre ainsi que 3 des 6 tirs de carrières. Contrairement aux 3 tirs de carrières détectés, les 3 tirs de carrières non détectés sont localisés à plus de 4 km de la fibre optique.

3.4.3 Impact des nouveaux attributs

L'ajout des attributs spatiaux et de similarités permet une meilleure distinction entre les signaux sismiques naturels et les bruits anthropiques (Figure 3). Par exemple, les événements liés aux véhicules mobiles, fréquemment rencontrés le long de la fibre, ont pu être différenciés des tremblements de terre grâce à l'analyse de leur signature spatiale et de similarité.

3.4.4 Limites identifiées

Dans ces travaux, plusieurs limites ont été observées :

- Le jeu de donnée utilisé ne dispose que de peu d'événements sismiques catalogués par le BCSF-RENASS (13 tremblements de terre, 6 tirs de carrières) sur une durée relativement longue (3 semaines).
- L'utilisation d'un algorithme ML supervisé nécessite l'utilisation d'un jeu de données annoté limite l'applicabilité du système à des environnements où les catalogues sismiques sont disponibles ou pouvant être facilement mis en place, avec le risque cependant de mal labelliser certains séismes de faible magnitudes.
- L'influence de la géométrie du réseau de fibres optiques et de son rapport signal sur bruit (SNR ratio en anglais) varie en fonction de la position de fibre.

3.5 Conclusion

L'intégration des attributs spécifiques aux données FO-DAS (spatiaux, similarités) a permis d'améliorer de façon importante les performances de la classification supervisée dans un environnement réel. Cependant, la dépendance aux données annotées reste une contrainte majeure, limitant l'évolutivité de la méthode à des scénarios de surveillance continue. Pour surmonter ces limites, l'utilisation de méthodes non supervisées peuvent aider à regrouper les données similaires entre elles.

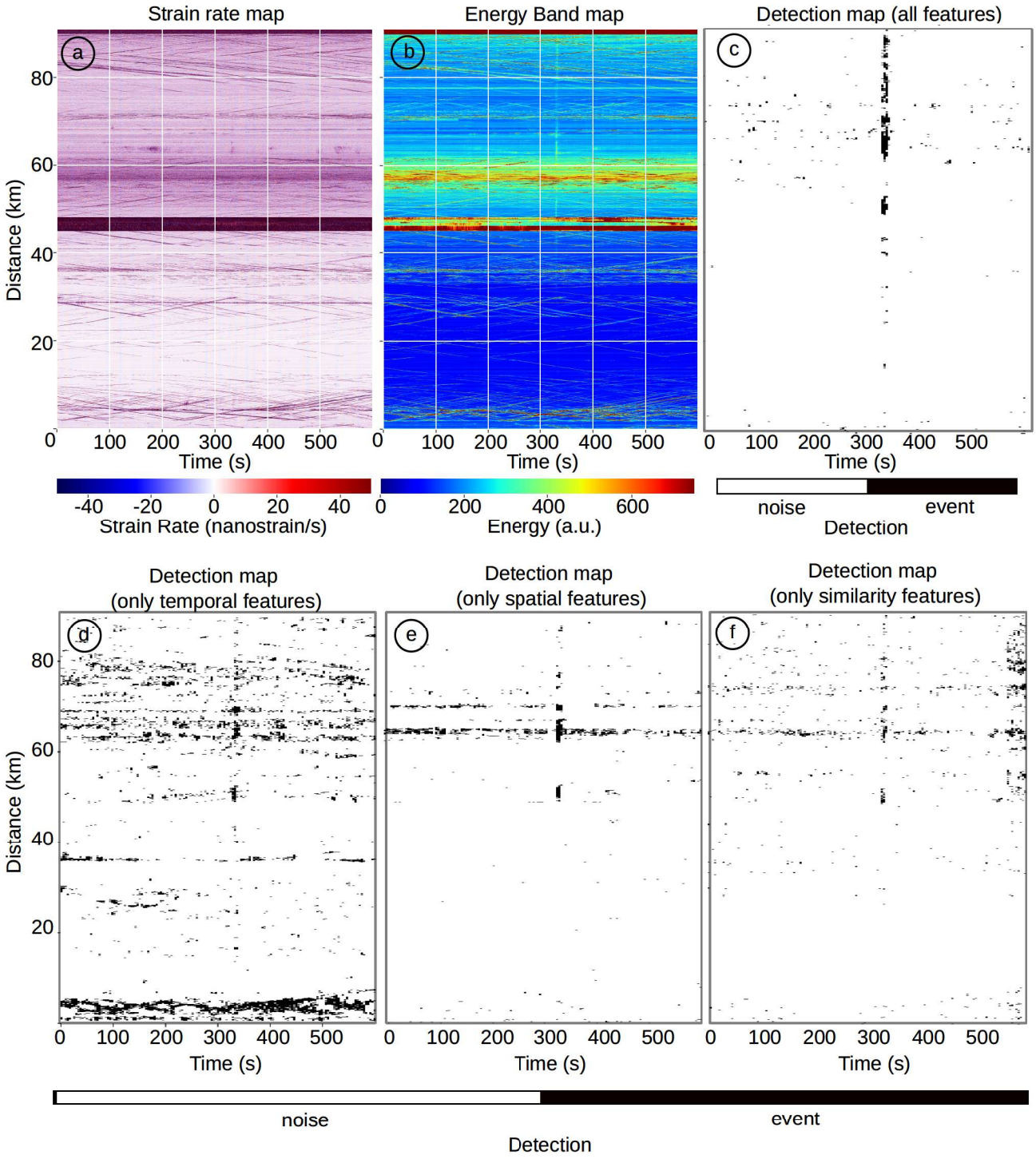


FIGURE 3 – Représentation de la carte de détection d'un tremblement de terre de magnitude Mw = 1.1 survenu près d'Argelès-Gazost, le 15 septembre 2022 à 9 :12 :16 UTC. La carte de détection est obtenue à partir de la carte de scores avec un seuil de 0,95. **(a)** représente le SR, **(b)** représente le EB, **(c)** la carte de détection obtenue avec un classificateur utilisant toutes les caractéristiques, **(d)** la carte de détection obtenue avec un classificateur entraîné uniquement sur les caractéristiques temporelles, **(e)** la carte de détection obtenue avec un classificateur entraîné uniquement sur les caractéristiques spatiales, **(f)** la carte de détection obtenue avec un classificateur utilisant uniquement les caractéristiques de similarité.

4 Exploration des Données FO-DAS à l'Aide des Attributs Précédemment Conçues et Comparaison avec des Méthodes d'Apprentissage Auto-Supervisé

4.1 Introduction

Dans les sections précédents, nous avons développé des méthodes de classification pour les données FO-DAS en utilisant des caractéristiques dérivées de la sismologie conventionnelle, suivies de la création de nouvelles caractéristiques spécifiques aux données FO-DAS. Bien que ces approches se soient révélées efficaces, elles reposaient sur un apprentissage supervisé, nécessitant un étiquetage précis des données. Pour pallier temporairement ce défi, nous avons utilisé soit des données générées à partir d'événements contrôlés réalisés dans nos expériences (Section 2), soit nous avons croisé les données FO-DAS avec des événements catalogués provenant de services externes comme BCSF-RENASS (Section 3). Cependant, ces approches dépendent au final des annotations externes, qui ne sont pas toujours disponibles. En l'absence de ces annotations, l'examen manuel des données de chaque événement devient nécessaire, un processus à la fois chronophage et susceptible de biais humains.

Cette section explore une approche alternative en utilisant des méthodes non supervisées pour analyser les données FO-DAS, regrouper les données similaires en clusters, puis effectuer un étiquetage cluster par cluster. Nous nous appuyons sur les représentations de l'espace latent développées dans les Sections 2 et 3, que nous comparons à une représentation générée à l'aide d'une méthode d'apprentissage auto-supervisé appelée BYOL. Les données d'entrée pour BYOL se composent de quatre représentations couramment utilisées dans l'analyse des données sismiques : la bande d'énergie, le rapport STA/LTA, le spectrogramme et la densité spectrale de puissance. Ces représentations sont ensuite transformées au format image pour faciliter leur manipulation dans BYOL. Des algorithmes de clustering sont appliqués à chaque représentation de l'espace latent pour organiser les données en clusters, et chaque cluster est étiqueté manuellement. Enfin, nous évaluons les deux méthodes de représentation en comparant le comportement des classes extraites par chaque approche et en évaluant la faisabilité du clustering pour grouper efficacement les données similaires afin de faciliter un étiquetage manuel rapide.

4.2 Données

4.2.1 Base de données Hautes-Pyrénées

Cette base de données inclut 19 enregistrements d'événements sismiques effectués sur 10 minutes chacun. Ces événements incluent des tremblements de terre ($Mw > 0.4$), des explosions de carrière, ainsi que des bruits anthropiques, tels que des véhicules en mouvement, et capturés sur une fibre optique de 91 km déployée dans les Hautes-Pyrénées. La donnée est acquise en utilisant comme paramètre une longueur de jauge (GL) de 10 m, un pas d'échantillonnage spatial de 4.8 m et une fréquence d'échantillonnage de 200 Hz. Les 19 enregistrements représentent un volume de donnée total de 206 Go.

4.2.2 Base de données Viella

La base de données Viella contient 44 jours d'enregistrements continus, collectés à l'aide d'une fibre optique de 800 m déployée sur le glissement de terrain de Viella et à proximité de l'éboulis. Ces enregistrements incluent des activités agricoles, des passages de véhicules, et des événements sismiques naturels, offrant un contexte riche et varié pour tester les méthodes proposées. Le câble de fibre optique a été enterré manuellement à une profondeur de 2 à 5 cm, et compacté manuellement.

La donnée est acquise en utilisant comme paramètre une longueur de jauge (GL) de 10 m, un pas d'échantillonnage spatial de 2.4 m et une fréquence d'échantillonnage de 400 Hz. Entre le début de la mesure et la fin de la mesure, la fibre a été coupé en plusieurs endroits, réduisant le nombre de canaux de 312 à 68. Le volume total de stockage des données Viella est de 3 To.

En parallèle de ces données acquis avec l'instrument FO-DAS, 8 géophones ont également été déployés. Ces géophones enregistrent la donnée avec une fréquence d'échantillonnage de 500 Hz. La durée d'acquisition avec les géophones est de 20 à 25 jours, selon la capacité de la batterie embarquée.

4.3 Méthodes

La chaîne de traitement développée se compose de trois étapes principales : **pré-traitement**, **représentation de la donnée**, et **clustering** (Figure 4).

4.3.1 Pré-traitement

Les données des deux bases ont été pré-traitées pour supprimer les bruits globaux (suppression du mode commun) et découpé en fenêtres de taille en temps et en distance adapté (t_{win}, d_{win}). Pour la base Pyrénées, chaque fenêtre a pour dimension (10 min, 1000 m), tandis que pour la base Viella, des fenêtres de dimension (1 min, 100 m) ont été utilisées. Au total, nous disposons de 17 480 fenêtres de données pour la base de données Pyrénées dont 243 contiennent un tremblement de terre et 52 un tir de carrière. Pour la base de donnée Viella, 142 727 fenêtres de données ont été acquises.

4.3.2 Représentation des données

Deux représentations de données ont été testées :

- **Attributs construits à la main.** Cette méthode repose sur 111 attributs décrivant les comportements temporels, fréquentiels et spatiaux des signaux FO-DAS, ainsi que les relations de similarité entre les canaux adjacents. Ces 111 attributs sont développés à la section précédente (Section 3).
- **Attributs construit avec l'approche image-BYOL.** Les signaux FO-DAS sont transformés en images combinant des spectrogrammes, des cartes d'énergie, des ratios STA/LTA, et des spectres locaux (Figure 5). Ces images sont ensuite traitées par l'algorithme BYOL pour extraire un espace latent optimisé.

Description de l'utilisation de BYOL. BYOL (Bootstrap Your Own Latent) est un algorithme d'apprentissage auto-supervisé conçu pour apprendre une représentation utile à partir de données non étiquetées (Grill et al., 2020). L'algorithme repose sur une approche d'apprentissage contrastif, où le modèle est entraîné à produire des caractéristiques similaires pour des versions augmentées du même input et des caractéristiques différentes pour des versions augmentées d'inputs différents. Cette méthodologie incite le modèle à apprendre un espace latent significatif qui capture la structure sous-jacente des données sans nécessiter d'étiquettes explicites. Dans notre travail, nous utilisons le modèle BYOL pour apprendre un espace latent basé sur 512 attributs extraits des images.

Cependant, BYOL est fortement dépendant de la qualité du processus d'augmentation des données. L'algorithme nécessite un ensemble diversifié de versions augmentées des données d'entrée pour apprendre efficacement un espace latent robuste et généralisable. Dans notre étude, nous considérons plusieurs transformations classiques d'images, y compris le retournement, la rotation, les variations de couleurs et le recadrage aléatoire. Des techniques comme le retournement, la rotation et le recadrage aléatoire interagissent directement avec le contenu des images, rendant le modèle moins sensible aux variations de position et d'orientation des formes caractérisant des types d'événements spécifiques. En revanche, les variations de couleurs introduisent du bruit dans les données

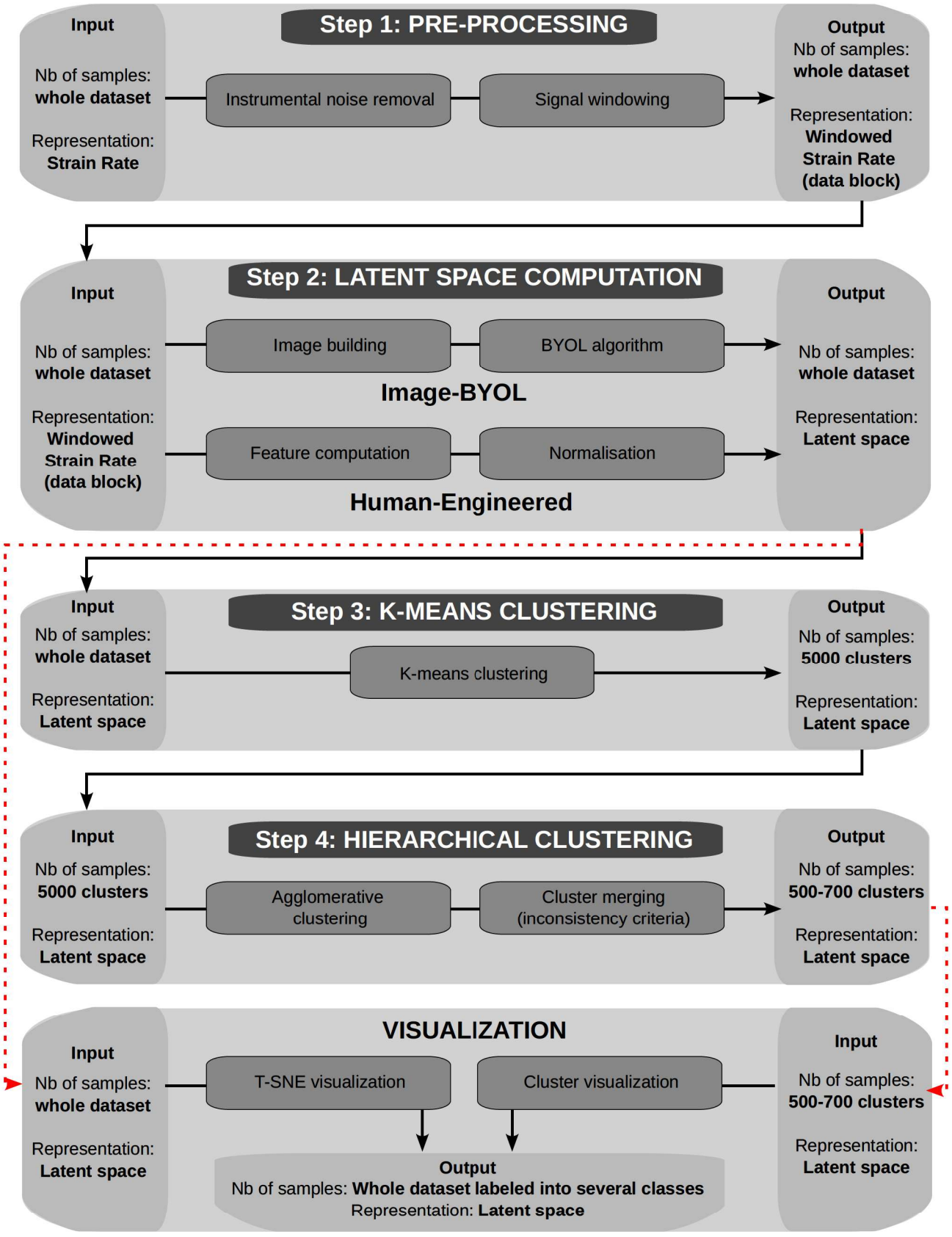


FIGURE 4 – Aperçu de la chaîne de traitement.

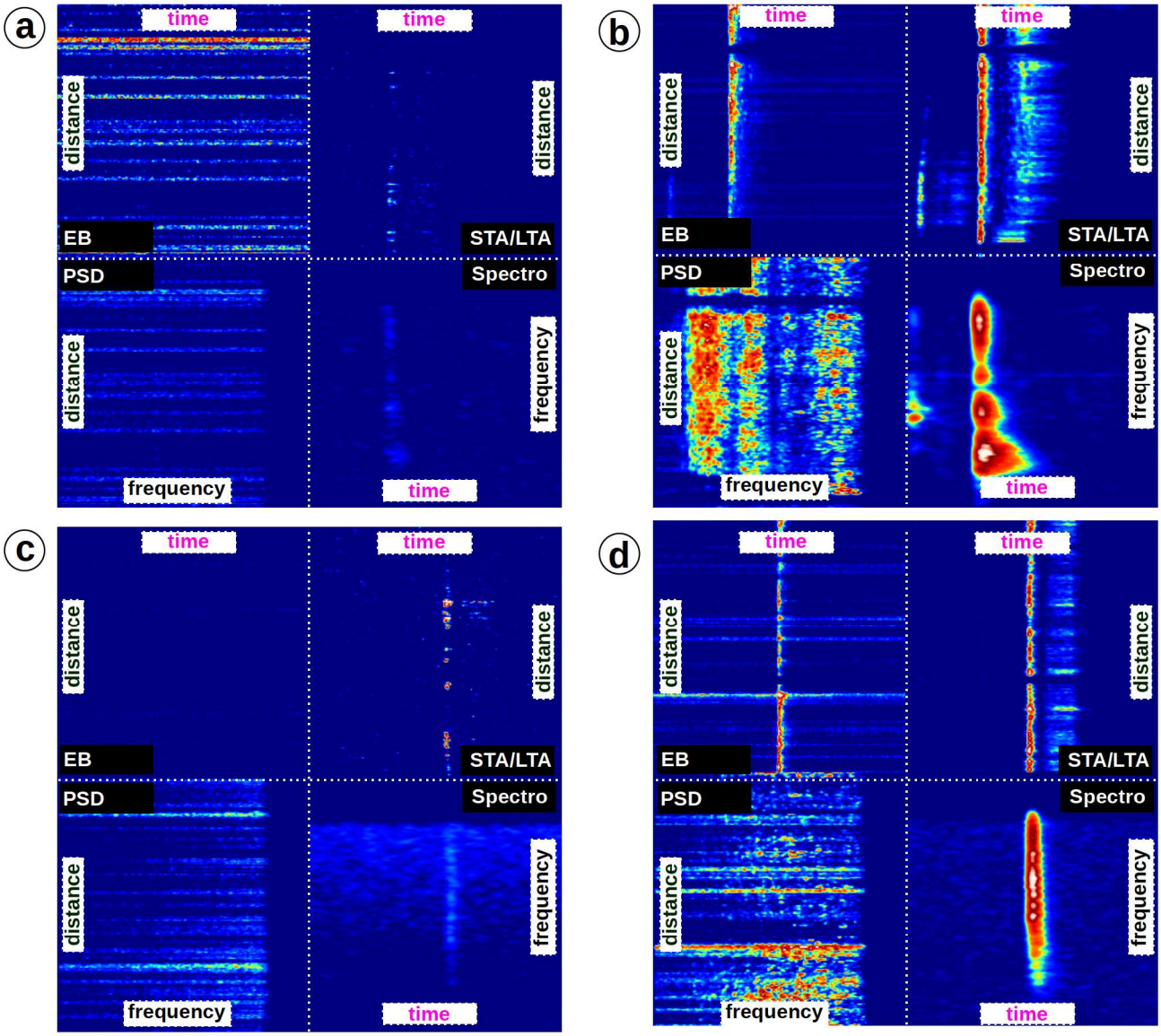


Figure 5 – Représentation en image du même événement, mesuré en deux points différents de la fibre optique. **(a)** et **(b)** représentent une explosion de carrière mesurée à deux points distants de 6 km. **(c)** et **(d)** représentent un tremblement de terre mesuré à deux points distants de 31 km. Données issues du jeu de données des Pyrénées.

d'entrée en modifiant des éléments tels que la teinte, la saturation, le contraste et la luminosité. Cette technique permet de rendre les caractéristiques moins sensibles aux effets causés par la qualité de l'instrumentation ou par l'influence de la localisation de l'événement par rapport à la fibre optique ou à l'interrogateur.

4.3.3 Clustering et analyse

Un algorithme de clustering à deux niveaux est appliqué pour réduire le volume des données :

- **K-Means.** Cette étape regroupe les données de chacun des jeux de données dans 5000 clusters. Ce nombre élevé permet de simplifier l'analyse des données sans perdre d'informations significatives. Pour la base de données Pyrénées, nous réduisons la quantité de donnée à utiliser, de 17 480 fenêtres de données à 5000 clusters. Pour Viella, nous passons de 142 727 fenêtres de données à 5000 clusters.
- **Clustering hiérarchique.** Ces clusters sont ensuite organisés dans une structure arborescente, reliant les clusters les plus similaires en eux. Dans cet arbre, chaque feuille est un des 5000 clusters produits par K-Means, et chaque nœud est un regroupement de deux clusters (initiaux ou formés en un autre nœud. Il est possible de couper cet arbre en suivant plusieurs critères, comme le critère d'inconsistance : ce critère compare la distance entre le nœud parent et les nœuds enfants. Le cluster formé par le nœud parent est considéré inconsistant si la distance du nœud parent est bien plus important que la moyenne des distances des nœuds enfants. Cela permet de former des clusters constitués d'un plus faible effectif de données que les autres mais contenant des données significativement différentes que celles contenus dans d'autres clusters plus peuplés. Nous fixons ce critère d'inconsistence à 1.152 après plusieurs essais afin d'obtenir en sortie entre 500 et 700 clusters. Cela permet d'obtenir, pour la base de donnée Pyrénées, 582 clusters avec les attributs construits à la main et 570 clusters avec l'approche image-BYOL. Pour la base de données Viella, nous obtenons respectivement 612 et 640 clusters.

Les clusters obtenus sont par la suite analysés pour leur homogénéité interne et leur pertinence par rapport aux événements réels, en comparant leurs contenus aux catalogues de référence et en identifiant les motifs périodiques ou anomalies.

4.4 Résultats

4.4.1 Impact du temps et de la position sur fibre dans l'espace latent

Afin de visualiser la distribution des données dans l'espace latent, nous utilisons l'algorithme t-SNE pour réduire les 111 dimensions formés avec les attributs manuels et les 512 dimensions construites à partir des attributs produit par l'approche image-BYOL vers 2 dimensions.

Les résultats montrent que l'influence du temps n'affecte pas significativement la localisation des points dans la représentation t-SNE pour le dataset Pyrenees, tandis que l'influence spatiale est marquée, particulièrement avec l'espace latent construit à la main, qui forme deux clusters distincts. En revanche, l'espace latent image-BYOL montre une distribution plus uniforme des points, avec une séparation moins claire basée sur la position ous le temps.

Pour le dataset Viella, l'influence du temps et de la position est plus prononcée. Certaines périodes de temps (par exemple, les premiers 11 jours ou après 37 jours) se différencient bien dans l'espace t-SNE. Certains points se regroupent également selon la distance le long de la fibre, particulièrement dans les premiers 100 m. Cela reflète les conditions environnementales variées, notamment le bruit dans la ferme pour la première section de la fibre. Cette analyse démontre que l'augmentation des données et la qualité de la construction de l'image dans l'espace latent image-BYOL permettent une meilleure réduction de l'impact de la distance, ce qui se traduit par une distribution plus homogène des points.

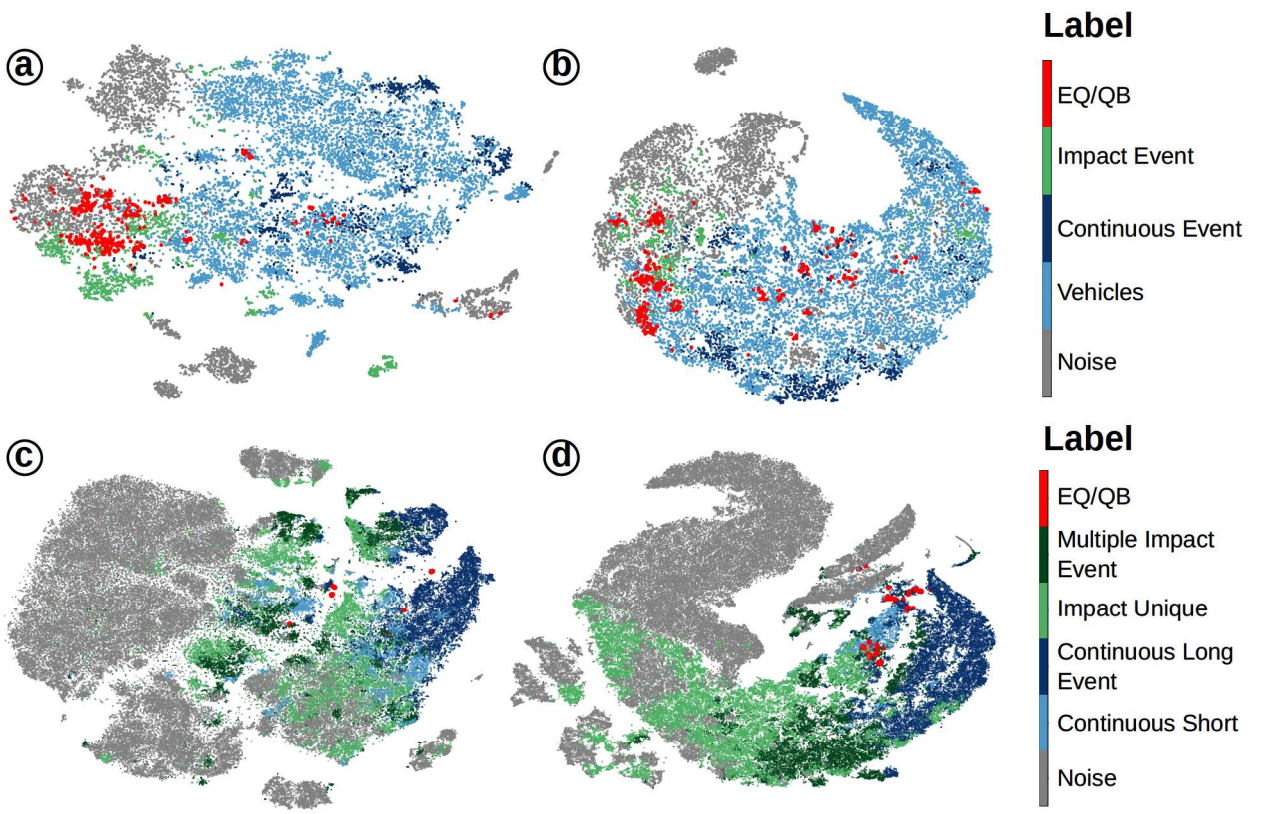


FIGURE 6 – Représentation *t-SNE* pour le jeu de données des Pyrénées **(a,b)** et de Viella **(c,d)**. Les couleurs représentent les classes attribuées manuellement. **(a,c)** sont obtenues avec l'espace latent basé sur des caractéristiques humaines, tandis que **(b,d)** sont obtenues avec l'espace latent issu de l'approche image-*BYOL*. Pour des raisons de visualisation, la classe des tremblements de terre est représentée par des points agrandis à une taille de 5 par rapport aux autres pour le jeu de données des Pyrénées, et à une taille de 150 pour le jeu de données de Viella.

4.4.2 Analyse et labellisation des clusters

Nous avons examiné visuellement le contenu de chacun des clusters formés à partir de données FO-DAS, afin d'identifier les principales classes d'événements sismiques. Pour le jeu de données des Pyrénées, nous en avons identifié cinq : véhicules, événements continus (signal constant d'au moins 10 s), événements d'impact (moins de 10 s), séismes ou explosions de carrière, et bruit. Pour le jeu de données de Viella,nous en avons identifié six : événements continus courts, événements continus longs, impacts multiples, impacts uniques, séismes et bruit. Les clusters ont été étiquetés en fonction de leur population majoritaire pour des événements communs (véhicules en mouvement, bruit), tandis que pour les événements plus rares (événements continus, impacts, tremblement de terre, tirs de carrière), un seuil de 25% de blocs de données appartenant à la classe en question a été utilisé.

Pour visualiser la distribution des classes, nous avons utilisé la représentation t-SNE basées sur les espaces latents construits à l'aide de caractéristiques calculées manuellement et de l'approche image-BYOL (Figure 6). Ces représentations montrent que les classes fréquentes sont bien séparées, tandis que les classes rares, comme les séismes et les explosions de carrière, apparaissent dans de petits groupes isolés souvent mêlés à des classes plus fréquentes. Cette observation ne reflète pas une inefficacité des caractéristiques utilisées, mais plutôt une limitation inhérente à la représentation t-SNE, qui privilégie les grandes tendances dans un espace réduit.

4.4.3 Validation

Base de données Hautes-Pyrenees : Comparaison avec une catalogue de référence pour la détection de tremblement de terre

La validation des classes d'événements sismogéniques définies manuellement pour le jeu de données des Pyrénées a été réalisée en comparant la proportion de fenêtres de données classifiés comme séisme ou explosion de carrière à chaque intervalle d'une minute avec la médiane des données sur une période de 10 minutes. Une détection est considérée réussie si cette proportion dépasse la médiane augmentée de deux fois l'écart médian absolu (MAD), une mesure robuste moins sensible aux valeurs extrêmes. En utilisant le catalogue de référence BCSF-RENASS pour valider ces résultats, 9 des 19 événements répertoriés ont été détectés avec l'espace latent basé sur des attributs calculées manuellement, contre 11 sur 19 avec l'approche image-BYOL. Toutefois, plusieurs événements, dont certains de magnitude supérieure à 1.0, n'ont été détectés par aucune des méthodes, notamment ceux situés à plus de 10 km de la fibre. Des différences de détection entre les deux approches ont été observées : certains événements éloignés (>25 km) sont détectés uniquement par image-BYOL, tandis que d'autres proches de la fibre sont capturés uniquement par les attributs manuels. Ces résultats mitigés s'expliquent par la présence de clusters impurs, contenant plusieurs classes d'événements, ce qui complique la fiabilité des critères de décision et peut engendrer des faux positifs, notamment dans des régions à faible activité sismique ou à fort bruit de fond.

Base de données Viella : Analyse de la périodicité pour les événements anthropiques

Le jeu de données de Viella, en raison de son enregistrement continu, permet une analyse temporelle approfondie des résultats de détection, notamment en explorant les motifs périodiques sur des bases journalières et hebdomadaires.

La classe "bruit" révèle une périodicité journalière marquée, caractérisée par des périodes de silence accru la nuit et une activité plus intense en journée. Parmi les classes anthropiques, les événements continus longs présentent une périodicité particulièrement forte, avec une proportion accrue d'événements pendant les heures de forte activité, tandis que les classes des événements continus courts et des impacts multiples montrent une périodicité moins prononcée, avec des hausses légères durant la journée. À l'inverse, la classe des impacts uniques ne présente pas de motif périodique distinct.

Ces observations soulignent des tendances intéressantes : la périodicité des événements continus longs suggère qu'ils proviennent majoritairement d'activités anthropiques, telles que des travaux agricoles, caractéristiques du site de Viella. En revanche, l'absence de périodicité pour les impacts uniques laisse penser que ces événements sont sporadiques et pourraient inclure des sources non anthropiques.

Base de données Viella : Comparaison avec un catalogue de référence pour les tremblements de terre

Au cours de la période de mesure, 20 séismes ont été confirmés visuellement sur les données FO-DAS en utilisant les catalogues sismiques de l'Observatoire Midi-Pyrénées (OMP) et du BCSF-RENASS. Tous les événements de magnitude $M_w \geq 2.0$ ont été détectés, y compris un séisme situé à 143 km avec $M_w = 2.8$, bien que la détection soit moins cohérente pour les magnitudes plus faibles. L'approche basée sur l'espace latent catalogué a identifié 16 événements, dont 8 séismes catalogués (le plus faible avec $M_w = 1.5$ à 13 km) et 3 événements non catalogués. En comparaison, la méthode image-BYOL a détecté 55 événements, dont 9 séismes catalogués (le plus faible avec $M_w = 1.4$ à 28 km) et 5 événements non catalogués. Bien que la méthode image-BYOL ait détecté davantage d'évé-

nements, elle a généré plus de faux positifs et manquait de critères de filtrage, tandis que l'approche basée sur l'espace latent catalogué, avec moins de fausses alertes, a fourni des résultats plus fiables pour identifier les séismes potentiels.

Base de données Viella : Comparaison avec des données géophones pour la détection de tremblement de terre

Pour évaluer les performances de la chaîne de traitement appliqué à l'instrumentation FO-DAS, une analyse complémentaire des données provenant des nœuds sismiques installés le long du câble de fibre optique a été réalisée. En raison d'une rupture de fibre dès le premier jour de mesure, seules les données des nœuds 1 à 6 ont été utilisées, avec une restriction temporelle jusqu'au 1er janvier en raison de contraintes d'alimentation des nœuds. Après suppression de la réponse instrumentale des géophones par déconvolution et segmentation des données en blocs de 1 minute, la méthode image-BYOL a été appliquée sur des images combinant traces sismiques et spectrogrammes. Cette analyse a permis de détecter les 21 événements répertoriés par le BCSF-RENASS ou l'OMP, mais a également généré plus de faux positifs que l'approche image-BYOL appliquée aux données FO-DAS, soulignant la nécessité d'améliorations pour réduire ces faux positifs, en particulier pour les événements de faible intensité.

4.5 Discussion

4.5.1 Forces des méthodes proposées

- L'approche non supervisée réduit considérablement l'effort d'annotation en regroupant des millions de segments individuels en plusieurs centaines de clusters.
- Les représentations avec attributs manuels et attributs issues de l'approche image-BYOL sont d'efficacité comparables. L'approche image-BYOL est intéressante car la chaîne de traitement est entièrement indépendante des efforts de construction manuelle des attributs et permet une évaluation complète de la performance de la chaîne de traitement.

4.5.2 Limites et perspectives

Malgré les résultats prometteurs, plusieurs défis restent à relever :

- **Impureté des clusters.** Certains clusters contiennent des signaux mixtes, nécessitant des itérations supplémentaires pour améliorer leur homogénéité.
- **Coût de calcul.** L'approche image-BYOL nécessite des ressources GPU, ce qui peut limiter son adoption à grande échelle.
- **Règles pré-définies.** L'identification des événements rares repose encore sur des règles pré-définies, ce qui peut introduire des biais et des erreurs.

4.6 Conclusion

Cette section a mis en lumière le potentiel des approches non supervisées pour analyser les données FO-DAS, en comparant deux représentations distinctes. Les résultats montrent que les méthodes proposées permettent une classification efficace des événements fréquents et une réduction significative de l'effort d'annotation. Cependant, des améliorations sont nécessaires pour gérer les événements rares et optimiser l'homogénéité des clusters. Ces travaux ouvrent la voie à des systèmes d'analyse plus autonomes et robustes pour les applications utilisant l'instrumentation FO-DAS, avec des implications prometteuses pour la surveillance sismique et environnementale.

5 Conclusions et Perspectives

5.1 Résumé des résultats

Cette thèse met en place des chaînes de traitement complètes pour la classification et l'exploration automatique des données mesurées par instrument FO-DAS. Après un état de l'art détaillé sur les applications de l'instrumentation FO-DAS en sismologie et les techniques de traitement des données (Section 1), la **Section 2** présente une chaîne de traitement composée de trois étapes : pré-traitement, classification et post-traitement. Basée sur un système de traitement en flux de données (mis en place à l'étape du pré-traitement), un algorithme de classification appelé forêt aléatoire utilisant 53 attributs développés pour la sismologie conventionnelle, et une étape de prise en compte des canaux adjacents appelé champ de Markov aléatoire, cette méthode a atteint une précision de classification de 87% pour six classes d'événements mesuré dans un environnement contrôlé au centre d'essai FEBUS Optics. Cependant, les tests étaient limités par des conditions de mesure idéale sans perturbation sismique externe et à l'aide d'une faible longueur de fibre (22 m). La **Section 3** a élargi la classification avec 111 nouvelles caractéristiques tenant compte de la dimension spatiale des données FO-DAS. Utilisant une fibre de 91 km déployée dans les Hautes-Pyrénées, le système a détecté des séismes de magnitudes Mw>0.4 et des tirs de carrière malgré un rapport SNR variable. Cependant, la dépendance à des annotations précises limite la généralisation de l'approche. Pour pallier ce problème, la **Section 4** a exploré une méthode de clustering non supervisé pour la labellisation de données, réduisant l'effort de labellisation de plus de 140 000 fenêtres individuelles à environ 600 clusters pour le jeu de données Viella (44 jours d'enregistrement continu). Cette approche a permis d'identifier des événements fréquents (mouvements de véhicules, activités agricoles) et rares (séismes locaux) dans le jeu de données, mais a révélé des défis liés à la labellisation des clusters liés à l'impureté de ceux-ci et à l'utilisation de règles pré-définies pour gérer les faux positifs et négatifs. Dans l'ensemble, ces travaux ont cherché à **limiter l'influence de la subjectivité humaine** dans la labellisation, tout en soulignant les défis persistants dans la détection des événements rares et l'amélioration de la généralisabilité des modèles.

5.2 Contributions apportées par la thèse

Cette thèse permet de faire le lien entre la sismologie conventionnelle et celle basée sur l'utilisation de l'instrument FO-DAS en combinant des techniques classiques de traitement du signal (analyse des formes d'onde temporelles, spectres, spectrogrammes) avec de nouveaux outils adaptés aux données FO-DAS, comme l'utilisation d'attributs spatiaux et la similarité entre canaux. Ces approches permettent d'extraire des informations pertinentes pour la **classification des événements sismiques** et d'exploiter efficacement les vastes volumes de données générés par les instruments FO-DAS. Un apport majeur réside dans le **regroupement des événements sismiques similaires** à l'aide de méthodes de clustering, réduisant le temps d'annotation des données et rendant la méthode applicable aussi bien à des événements fréquents qu'à des phénomènes rares tels que les séismes locaux ou les activités anthropiques.

Malgré des résultats prometteurs, plusieurs **limitations** ont été identifiées, notamment la dépendance à des données spécifiques au site et les difficultés à généraliser les modèles à d'autres environnements. Les travaux se sont concentrés sur des périodes d'acquisition limitées, ce qui restreint l'entraînement des modèles supervisés, en particulier pour les événements rares. L'utilisation de techniques d'apprentissage par transfert ou la création de modèles pré-entraînés sur des données multi-sites pourraient répondre à ces défis. À terme, une généralisation des modèles contribuerait à déployer rapidement des systèmes de surveillance sismique à grande échelle, ouvrant la voie à des applications pratiques pour la détection des risques naturels et industriels.

5.3 Perspectives

5.3.1 Amélioration de la méthode

Plusieurs pistes d'amélioration peuvent être mis en place pour le traitement des données FO-DAS, notamment par l'exploration d'une méthode potentiellement prometteuse de **re-clustering itératif**, en complément de l'approche exploratoire développée dans la Section 4. Cette approche consisterait à affiner les clusters impurs en réappliquant le processus de regroupement sur les données mal classifiées, permettant ainsi de créer des clusters plus homogènes au fil des itérations. Inspirée de domaines tels que la segmentation d'images (Bensaid et al. 1996) et les systèmes de conduite autonome (Kruber et al. 2018), cette méthode pourrait être adaptée à la classification des données FO-DAS pour mieux identifier les événements sismiques, y compris ceux de très faible magnitude. Bien que cette méthode n'ait pas été implémentée dans le cadre de cette thèse, elle représente une piste intéressante pour améliorer la précision des classifications tout en réduisant la subjectivité liée à l'utilisation de règles prédéfinies.

Un autre axe d'étude, dans le but de réduire les besoins en données d'entraînement tout en maintenant des performances élevés, est l'utilisation d'algorithmes de **transfer learning**. Cette méthode, qui permet d'exploiter les connaissances acquises sur un site différent ou par un type de capteur différent , pourrait répondre aux défis liés à la généralisation des modèles. Cependant, des obstacles restent à surmonter, notamment l'influence des variations entre sites (conditions environnementales, placement des capteurs) et les différences intrinsèques entre les sismomètres conventionnels et les instruments FO-DAS, comme leur réponse en fréquence variable et leur rapport SNR souvent plus faible.

5.3.2 Partage de données et collaboration en sismologie FO-DAS

Le partage des données FO-DAS reste un enjeu majeur, en raison de l'existence de multiples formats de fichiers propriétaires et du manque de standardisation des métadonnées. Certaines initiatives, telles que PubDAS (Spica et al. 2022), favorisent l'accès à des **bases de données ouvertes**, tandis que d'autres efforts sont menés pour proposer des **bibliothèques de traitement standardisées**, comme DASPy (Hu et al. 2024) ou encore XDAS (Trabattoni et al. 2024). Ces outils permettent de faciliter le prétraitement, la compression et l'analyse des données, tout en encourageant une collaboration accrue entre chercheurs. Notre contribution s'inscrit dans cette démarche, notamment par la mise à disposition des données acquises dans les Hautes-Pyrénées (DOI : 10.57932/5b1302d6-57cd-44e4-81ac-5d585a7f8951.) et de la chaîne de traitement développée dans la Section 3 (répertoire GitLab : "EOST/seis-learning-spatial").

5.3.3 Ressources de calcul

Concernant les ressources de calcul, plusieurs solutions adaptées aux besoins spécifiques des applications FO-DAS ont été explorées. L'utilisation des **ressources de calcul locale** à l'interrogateur DAS est une solution pour des applications sur le terrain avec une latent presque inexistante, cependant les ressources de calcul sont limités. Durant la thèse, j'ai pu travailler avec plusieurs ingénieurs de FEBUS Optics afin d'implémenter le système dans un DAS en vue de démonstration au centre d'essai de FEBUS auprès d'entreprises clientes. Les ressources informatiques limités nous ont obligé de n'utiliser que les attributs les plus importants dans le processus de classification. L'utilisation de **stations de travail portables (workstation en anglais)** semble être un compromis idéal pour les applications sur le terrain nécessitant une faible latence, comme les tâches de classification en temps réel. En revanche, les **systèmes de calcul haute performance (HPC)** sont plus adaptés à l'exploration de données sur de longues périodes ou à l'entraînement de modèles complexes, bien que l'échange de données volumineuses entre l'interrogateur DAS et le HPC introduise des délais importants. Des algorithmes de **compression des données** FO-DAS, comme ceux proposés

par Dong et al. 2022 (compression à 40% et sans perte), offrent des perspectives intéressantes pour réduire les temps de transfert et optimiser le stockage des données.

Camille HUYNH

Real-time seismic monitoring using DAS fiber-optic instrumentation and machine learning: towards autonomous classification of natural and anthropogenic events

Résumé

Ces dernières années, une nouvelle technologie basée sur l'utilisation de fibres optiques est apparue pour surveiller les événements acoustiques naturels ou anthropogéniques : la détection acoustique distribuée (Distributed Acoustic Sensing - DAS). Cette technologie innovante permet de mesurer les vibrations sismiques à très haute résolution spatiale sur des distances allant de quelques dizaines de mètres à plusieurs centaines de kilomètres. Bien que ces données soient plus volumineuses et plus complexes à traiter que celles des sismomètres traditionnels, elles offrent des perspectives prometteuses, notamment pour l'analyse des champs d'ondes générés par les tremblements de terre, la détection des glissements de terrain, la surveillance de divers événements anthropogéniques (tels que les déplacements de piétons, les mouvements de véhicules, ou les signaux sismiques provenant des activités humaines), les événements de faible amplitude ou très localisés, et la localisation précise de l'origine de ces événements sismiques. L'objectif de cette thèse est de développer et de tester des chaînes d'analyse de données automatisées en utilisant des approches basées sur l'IA pour détecter, classer et analyser les données DAS à fibre optique en temps quasi réel. L'objectif est axé sur la surveillance locale et régionale de zones spécifiques afin de permettre la détection et l'identification en temps réel d'événements naturels tels que les tremblements de terre et les glissements de terrain.

Mots clés : Distributed Acoustic Sensing (DAS), Apprentissage machine, Sismologie, Espace latent, Tremblements de terre, Sources sismiques anthropiques, Apprentissage supervisée, Apprentissage non-supervisée

Abstract

In recent years, alongside traditional seismometer-based approaches, a new technology based on the use of optical fibers has emerged for monitoring natural or anthropogenic acoustic events: Distributed Acoustic Sensing (DAS). This innovative technology enables the measurement of seismic vibrations at very high spatial resolution over distances ranging from tens of meters to several hundred kilometers. Although these data are larger and more complex to process than those from traditional seismometers, they offer promising perspectives, particularly for analyzing the wavefields generated by earthquakes, detecting landslides, monitoring various anthropogenic events (such as pedestrian movements, vehicle movements, or seismic signals from human activities), low-amplitude or highly localized events, and precisely locating the origin of these seismic events. The goal of this thesis is to develop and test automated data analysis chains using AI-based approaches to detect, classify and analyze near-real-time fiber-optics DAS data. The objective is focused on local and regional monitoring of specific areas to enable the real-time detection and identification of natural events such as earthquakes and landslides.

Keywords: Distributed Acoustic Sensing (DAS), Machine Learning, Seismology, Latent Space, Earthquakes, Anthropogenic Seismic Sources, Supervised Learning, Unsupervised Learning