Application of Data Analysis and Machine Learning Techniques to Improve Baseline Volcano and Mountain Hazards Monitoring

by

Juan Camilo Anzieta

M.Sc., Universidad San Francisco de Quito, 2017
B.Sc., Escuela Politénica Nacional, 2012

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Earth Sciences
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2024

Declaration of Committee

Name: **Juan Camilo Anzieta**

Degree: **Doctor of Philosophy**

Thesis title: **Application of Data Analysis and Machine Learning Techniques to Improve Baseline Volcano and Mountain Hazards Monitoring**

Committee: **Chair:** Diana Allen
Professor, Earth Sciences

Glyn Williams-Jones
Supervisor
Professor, Earth Sciences

Nilima Nigam
Committee Member
Professor, Mathematics

Uwe Glässer
Committee Member
Professor, Computing Science

Gwenn Flowers
Examiner
Professor, Earth Sciences

Gregory Waite
External Examiner
Professor, Geological and Mining Engineering and Sciences,
Michigan Technological University

Abstract

In this work, several machine learning and data analysis tools were applied to various datasets consisting of seismo-acoustic recordings. Due to the complexities and uncertainties of each dataset, several criteria were proposed for the adequate application of these techniques for different tasks. The first was to assess the quality of two open-access volcano seismic datasets, one considered noisy (Cotopaxi volcano, Ecuador), and the other clean (Llaima volcano, Chile). By applying a catalogue cleaning procedure, metrics and benchmarks were defined to rapidly assess each dataset's quality. After analyses, the Cotopaxi dataset showed numerous mislabelled events, which was confirmed by performing a blind test from experts' assessment. In contrast, Llaima's catalogue yielded few mislabelled events, validating its cleaner status. A second task consisted of expanding a dataset of volcanic explosions from nearly 10 years (2006-2016) of continuous acoustic recordings from a seismo-acoustic network on the flanks of Tungurahua volcano, Ecuador. The original explosions were identified by human inspection using an amplitude threshold and omitted small to medium sized explosions. To expand this catalogue, a series of successive steps combining traditional and novel data analysis with (un)supervised machine learning tools was applied to continuous recordings from one station. This led to more than 29,000 new explosions being detected that were grouped and linked to changes in Tungurahua's activity. Finally, a third dataset consisting of several months (March-July 2021) of novel continuous acoustic recordings from the Squamish River (Mount Cayley, BC.), was explored. To use these recordings as a potential tool for natural hazards monitoring in the region, an acoustic activity baseline was established by harnessing the extreme weather conditions caused by the 2021 Western North American heat wave. Recordings of local weather parameters, especially water level gauges, were used to relate part of the acoustic measurements to the river's discharge rate using rapid data analysis tools. Anomalous signals deviating from this baseline are proposed as a means to identify future natural hazards. This work showcases the flexibility -yet care- with which machine learning and data analysis tools can be applied to monitor volcano and mountain hazards, and lays the groundwork for future developments.

Keywords: Volcano Seismology and Infrasound; Signal Detection and Classification; Machine-Learning Applications; Data Analysis for Geophysical Signals

Dedication

To my mother Natacha Reyes, to whom I owe all that I am and all that I will ever be. Gracias mamita.

"For what does a person benefit if he gains the whole world and lacks his soul? Or what will a person give to regain his soul?..."

Matthew 16:26

Acknowledgements

Just like every human effort, this document is the result of the joint collaboration (direct or indirect) of several people besides that of the main author. While this work summarizes almost 5 years of formal doctoral work, in reality it is also partly the distillation of several more years (more than a decade!) of cooperation between different institutions and individuals. As such, I will try to mention several different actors that have contributed to these studies, with the –not only risk but– certainty of missing some. For that, sorry to all of you who are not explicitly mentioned, but be aware that I will forever be in debt to you, and am willing to pay it in the form of food or drinks if you confront me for my forgetfulness. I will try to credit the different people's ordering first roughly related to the chapters' appearances and then in a general fashion.

First and foremost I would like to thank my supervision committee: Glyn Williams-Jones, Nilima Nigam, and Uwe Glässer. Thank you for your invaluable support and comments that guided me through all these years; I hope I was able to reflect all the insights you shared with me during this time, and that I can in turn share it in the future with others as well. Thanks to my examiners Gwenn Flowers and Gregory Waite who also helped improve the final version of this document. I want to kindly thank the personnel from the Instituto Geofísico de la Escuela Politécnica Nacional for sharing their data and thoughts for the second and third chapters of this thesis. Specifically, I would like to thank Mario Ruiz, Daniel Pacheco, Mónica Segovia and Andrea Córdova for their collaboration for the first chapter. I would like to thank my decades long friend Hugo Ortiz, who has shared ideas and support through the years, and specially helped shaping the third chapter of this thesis. As for the fourth chapter, I am in debt with many people that helped in the instrumentation and logistics aspects for data acquisition, and the improvement of several ideas. I would like to thank people from SFU's electronics shop, Anthony Slater and Gian Moreira for their help into trying to adapt different instrumentation to British Columbia's context. Thanks to Daniel Stein, Braden Walsh, Sarah Aufrère, Benjamin Coughlan and Malcolm Pidsosny for hours of support and driving to deploy instruments at Mount Cayley, BC, and special thanks to Jake Anderson for instrumentation support. Thanks also to all the people from SFU who supported this endeavour outside of the 'academic content' but made all other institutional affairs possible: Matt Plotnikoff, Elilan Ganesathas, Tarja Väisänen, Lorena Muñoz, Alyssa Hamilton-Messenger and many more.

Special thanks to all my friends with whom we shared Glyn as our incredible supervisor but also several discussions, jokes, adventures, and fun times during these years: you were the best group I could wish to share this journey with and you helped me through this time in many more ways than just work. Also a shout-out to friends from the glaciology group with whom we shared the research lab, and also offered their support (and inspiration) even to the last minute.

Finally, thanks to my family for allowing me to 'escape reality' for almost 5 years and waiting for me so kindly. Special thanks to my companion Antonina who backed me during this trip and without whom it would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The ongoing development of nations worldwide leads to a growing number of people searching for new living spaces and resources. Meanwhile, the Earth's dynamic geological processes are constantly reshaping our planet as various regions experience constant natural phenomena, including the growth and erosion of volcanoes and mountain ranges. These processes give rise to natural hazards that, added to the periodic effect of seasons and climate change, affect various groups of people around the world. For instance, as of 2015 more than 8% of the world's population lived within 100 km of a volcano with at least one significant eruption in the Holocene (the current geological epoch starting approximately 11,000 years ago), and more than 1 billion people (14.3%) lived within 100 km of a volcano with any activity in the Holocene (*Freire et al.*, 2019). In addition to the increasing proximity of communities to volcanoes, volcanic activity is a persistent phenomenon globally with around 40-50 continuing eruptions any given time (*Smithsonian Institution*, 2024). This means that the potential impacts of volcanic hazards to populations are not expected to decrease in the near future. Impacts associated to volcanic activity include phenomena with varying impact ranging from ash fall and lava flows, to volcanic tsunamis (e.g., *Connor et al.*, 2015).

Added to the hazards of active volcanoes, seasonal cycles contribute to the weathering, erosion and/or degradation of volcanoes, mountain ranges and glaciers that in turn produce other hazards such as landslides, avalanches, debris flows, lahars, etc. Since these hazards can occur in various mountain settings, as well as on volcanoes that are not necessarily erupting, they are much more prevalent than those directly related to volcanic activity, and thus cause economic damage and deaths at a higher rate. For example, landslides cause thousands of deaths yearly even considering that recorded values are under-registered (*Petley*, 2012). Furthermore, climate change has increased the occurrence of weather-triggered landslides or debris flows (*Gariano and Guzzetti*, 2016; *Chiarle et al.*, 2021).

These two facts: 1) Increasing population in the need of new places to live, and 2) natural hazards that, at best, remain constant over time, make it so that more people and infrastructure are at risk of potential loss and are a continuous concern for local and national governments. On the other hand, as exposure to natural hazards increases, technology advances and new instruments and methods become more accessible in cost and operability allowing for more effective risk mitigation initiatives. In particular, the cost of seismic and infrasonic recording systems is decreasing continuously with alternatives to traditional systems based on simple processing boards for digitizing signals and/or the use of low cost MEMS (Micro-ElectroMechanical systems) sensors (e.g., Raspberry Shake seismic and acoustic sensors, *Raspberry Shake*, 2024, Gem Infrasound Logger, *Anderson et al.*, 2018). Systems such as these allow for denser monitoring networks, or rapid temporal deployments without severely sacrificing the quality of the signals recorded (*Anderson et al.*, 2018; *Anthony et al.*, 2019). These advances often imply reduction of additional equipment related to power or digitizing capabilities and allow sensor installations on previously restrictive areas where even one new instrument is a significant improvement for hazard monitoring.

These new sensors, their increasing installation numbers, and the specifics involved in the whole system come with different complications at different stages of implementation. For instance, when monitoring is at its onset and sensors have just started to acquire data from the geological processes, meaningful conclusions are difficult to extract. This means that at first, a remote and unknown area requires careful examination of the regular signal contents to enable eventual distinction of normal versus anomalous behaviour. Therefore, there is a need to properly/objectively establish a *baseline* and perform anomaly detection. On the other extreme case, once the networks start growing larger, the amount of data requires a paradigm that allows efficient information extraction, so that it can adequately keep up with the stream of baseline information as well as novel signals.

While initially a system requires manual treatment of the data, at some point the tasks become unfeasible without large human resources, and a system has to be built to handle ever increasing amounts of data. This means that the designed system needs to be **robust and scalable**: i.e., it performs consistently for similar datasets, and reasonably for increasingly larger datasets –or at least that can be easily modified to cope for them. The amount of data also requires that logistical challenges, such as transmission of data in remote areas, is done efficiently. This can be done by executing appropriate in-situ processing, and transmitting only the relevant information –via satellite or local low-bandwidth radios– instead of full signals. Since different sensor specifications allow for specific subsets of techniques to be applied, it is relevant to identify what information can be extracted from limited scenarios, and how it compares with ideal contexts. Additionally, systems based solely on human revision could suffer from methodological uncertainties, procedural errors, or subjective biases that go beyond the techniques or amount of data analyzed.

As a means to address these issues, the research and adoption of machine learning tools has become increasingly popular both at the level of natural hazard monitoring in general (*Jain et al.*, 2023; *Abraham et al.*, 2024), as well as for volcanic (*Malfante et al.*, 2018; *Khan et al.*, 2019; *Carniel and Guzmán*, 2020), earthquake (*Mousavi and Beroza*, 2023), and mass movement events such as landslides and debris flows (*Tehrani et al.*, 2022; *Ponziani et al.*, 2023) in particular. Nevertheless, all systems begin with some kind of human prior assessment, scrutinized data or predefined input, and are subject to potential unknown errors or knowledge gaps that if not corrected in time, can propagate surreptitiously even after the development of fully automated procedures. The success of a system then depends on the difficulties or particularities of its implementation, but especially on the nature, quantity, and quality of the available information to address a task. The degree of previous knowledge and the extension/size of a given dataset determines what kind of techniques or procedures can be applied to gain insights, and what tasks for hazards monitoring they can expound. Thus, here I define as a ***locked*** dataset one which is already given in a definite manner: there is a finite number of given observations and no natural expansion on the data itself can be achieved; all objects within are already catalogued, i.e., they have been assigned a label or class. Nevertheless, they are important because they constitute the input to create automated systems or even synthetic data. Conversely, an ***unlocked*** dataset is that which possesses streams of unconstrained data which can be further explored. For unlocked datasets, a given catalogue may or may not exist, but in principle new objects of interest can be extracted depending on a project's goal.

In this work, three different tasks related to monitoring of volcano and mass movement events were specified for the analysis of three different types of data sets, and the general conceptualization is summarized in Figure 1.1.

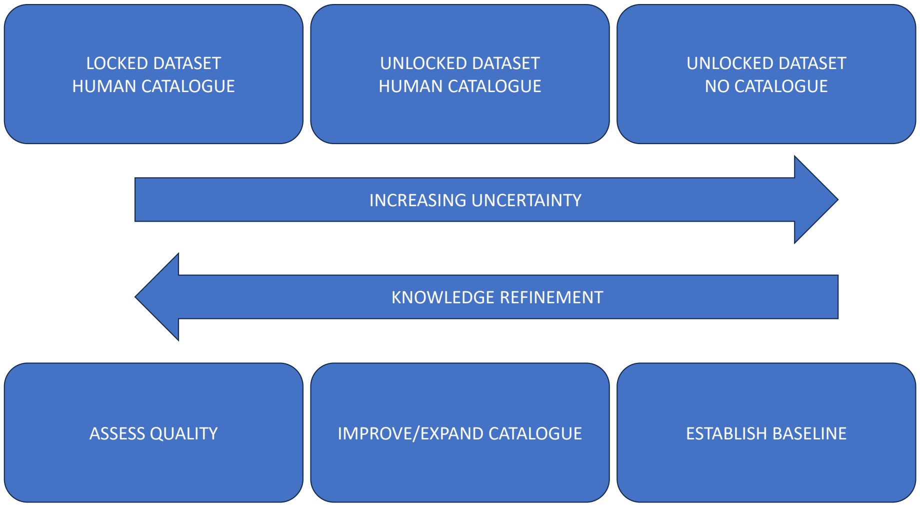

Figure 1.1: Diagram of applicable tasks depending on the nature and knowledge of given datasets.

Figure 1.1 shows the relationship between the uncertainty of different types of datasets with the knowledge refinement (i.e., amount of advancement or detail to be achieved) and tasks enacted upon each. To achieve the tasks, several data analysis and machine learning approaches were developed and applied to diverse seismic and acoustic datasets.

1.1 Data and Methods

1.1.1 Seismic and Acoustic Data

Seismology is a branch of geophysics and Earth Sciences which focuses on the study of the Earth's vibrations, in order to gain knowledge of the subsurface (*Shearer*, 2009). The measurement of the Earth's vibrations is done with *seismometers* that record ground motion based on principles like the inertia of suspended masses and springs inside an instrument responding to external movement (*Shearer*, 2009). Modern instruments output different magnitudes, e.g., voltages that are digitized into counts, which can be converted into a physical magnitude related to motion (i.e., displacement, velocity or acceleration). While these measurements occur in continuous time, the instruments record at given sampling frequencies (usually tens to hundreds of samples per second), and create a *time series* of the vibrations detected at the instrument's location. Acoustics on the other hand, study sound -mechanical waves propagating through the atmosphere- using *microbarometers* or *microphones.* Acoustic infrasound sensors measure pressure transients in the air in frequencies usually lower than human hearing (i.e., 0.01 to 20Hz vs 20Hz to 20kHz) because waves propagate longer distances in those frequencies (*Dannemann et al.*, 2023). These acoustic infrasound sensors are capable of recording explosions or large mass movements and make them well-suited for monitoring superficial ground activity at volcanic and mountain regions (*Allstadt et al.*, 2018). Thus, acoustics are a complementary tool to seismology that can enhance the understanding of different natural or man-made phenomena, and to describe the coupling between seismic and acoustic waves (*Dannemann et al.*, 2023). It is worth mentioning that seismic and acoustic signals are similar in the sense that both record perturbations propagating through a given medium (Figure 1.2).

Different phenomena can give rise to different *waveforms* in the time series recordings, and they contain information concerning the nature of the source that generated them, the path between the source and the sensor, and the site characteristics of the sensor (besides the instrument's response to the signal itself) (*Shearer*, 2009). Sometimes, recognizing different types of signals is straightforward for a trained analyst, or aided by a *spectrogram* of the waveform (Figure 1.3).

However, most signals are varied and complex, their characteristics may not be unique to a given class, they may overlap in time, they may be obscured by noise, or subtle differences may be of interest that may require meticulous analysis. Most importantly, given that a standard seismic or acoustic recording can contain millions of samples in a single day (e.g.,

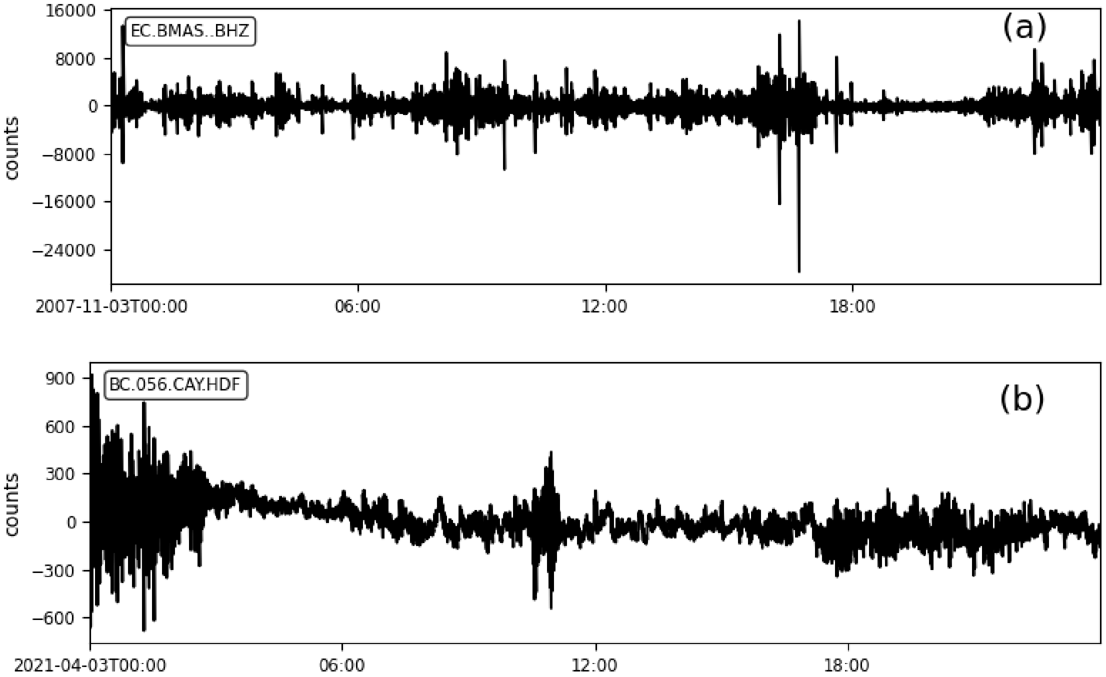

Figure 1.2: a) Example of a day's worth of raw seismic data and b) a day's worth of raw acoustic data from Tungurahua volcano and Mount Cayley, respectively.

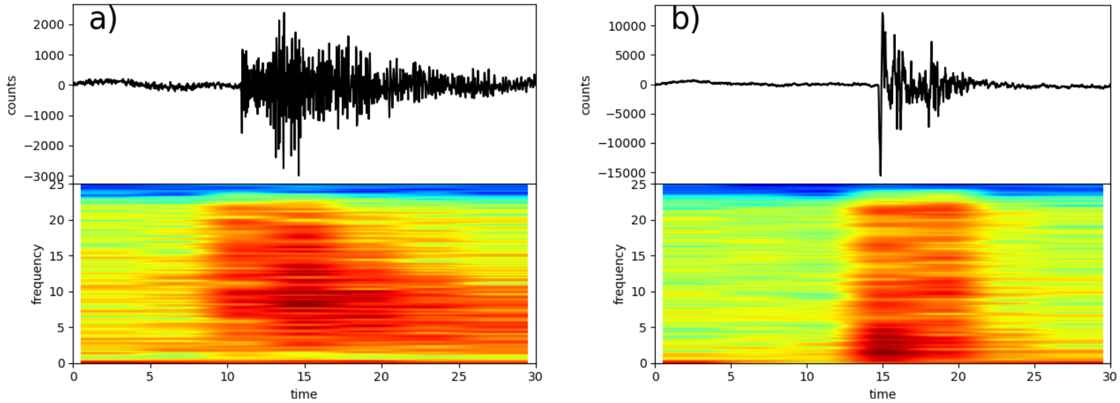

Figure 1.3: a) Example of a raw seismic volcano-tectonic signal with its spectrogram below. b) Example of a raw acoustic explosion signal with its spectrogram below. Both taken at Tungurahua volcano.

4,320,000 samples for an instrument recording at $50Hz$), human operators can seldom afford the time to analyze the waveforms "manually" from an ever growing number of instruments. Thus, automated systems comprised of computers using data analysis and machine learning tools are constantly being deployed and improved for better natural hazards monitoring.

1.1.2 Methods

Machine Learning

Many definitions exist for Machine Mearning (ML), however most revolve around roughly the same idea: "Machine Learning is the science of programming computers so that they can **learn from data**" (*Géron*, 2019). Since the use of data is required, many concepts and approaches overlap between Machine Learning and Statistics and interchangeable terminology is used depending on the context in which they are explained (*Wasserman*, 2004; *James et al.*, 2013). For example, supervised learning (a Computer Science term) is equivalent to class-prediction (a Statistics term), nevertheless, this variety has helped enrich the related fields involved (*Berk*, 2016).

A simplified way to understand (most) of machine learning is by using the following equation 1.1 as an analogy:

$$Y = f_X(x) \tag{1.1}$$

where f is a function that a computer will *learn* based on a given *training* set X, that will generate an output Y given an input x. The nature of each term can be wide-ranging. For example, supervised learning is a task where given a dataset X of observations with their respective labels, f learns to predict the label or class Y, of an input instance x. Unsupervised learning, or clustering, can be thought of as when the dataset X has no predefined labels for the observations but *class memberships* need to be assigned to each based on some similarity criteria. If the outputs Y are not classes, but numbers, then the task falls on the regression task, and similarly if the outputs $Y \sim X$ are new observations, this is called *generative learning.* An even more abstract task called *reinforcement learning*, consists on finding f such that it learns to take "optimal" actions Y based on rewards and policies as its "training set" X (*Géron*, 2019).

In general, machine learning application steps can be summarized as follows:

1. Identify the kind of task to be addressed

2. Obtain/Organize/Preprocess available data compatible with the goal

3. Select appropriate methods, models and algorithms for the dataset and task at hand
4. Train, tune, and validate the tools
5. Implement, apply or deploy the tools (and share them!)

In the natural hazards context, two interrelated tasks are of major relevance: the detection and the classification of signals of interest. These tasks can range from the identification of predefined/known types of waveforms, to the formation of new unknown classes or families of events, to the detection of anomalous signals. These tasks have been and continue to be explored for volcano (e.g., *Malfante et al.*, 2018; *Carniel and Guzmán*, 2020), as well as mass movement (e.g., *Thüring et al.*, 2015; *Chmiel et al.*, 2021; *Liu et al.*, 2021) monitoring. Since seismovolcanic events have been defined for a long time and are identified routinely in observatories (*Chouet and Matoza*, 2013), most of the ML techniques applied in volcano monitoring deal with supervised learning. Unsupervised techniques are less common and used in cases of large data sets where none or little prior human classifications have been performed (*Malfante et al.*, 2018), or with the aim of finding interesting families of events or patterns potentially related to volcanic activity (e.g., citealpAnzieta2019, Zali2024). Neverheless, before the application of many machine learning techniques, data generally needs to be preprocessed in one way or another, usually to avoid the *curse of dimensionality* that can lead to *overfitting* systems (systems that did not learn to *generalize* outside of their training dataset) with poor performance (*Géron*, 2019). For a brief description of some popular Machine Learning tools, and data analysis algorithms applied in the forthcoming literature see Appendix A.1.

Feature extraction: Seismic or acoustic recordings are continuous, and as such, there is no predefined data size of interest for analysis (i.e., some events of interest may last for longer or shorter periods, which implies time windows have varying sizes even for events that belong to the same class or type). For instance, some events might last a few seconds (e.g., explosions), while others can extend for minutes, hours or even days (e.g., volcanic tremor). For this reason, when trying to apply ML techniques in seismic or acoustic recordings, a usual strategy is to extract features from different events so that they can be properly compared (*Cortes et al.*, 2016). There is still no consensus on which are the best features to be used and different strategies are still being studied (e.g., *Malfante et al.*, 2018; *Khan et al.*, 2019), but comprehensive feature selection schemes can yield superior performances compared to ones based on "domain expertise" (*Toney et al.*, 2022).

Event Detection: In volcano monitoring, much of the work is dedicated to classification tasks, since event detection is usually done semi-automatically using standard data analysis algorithms such as STA/LTA (short-time-average, long-time-average ratio trigger) or the recently proposed VINEDA (*Bueno et al.*, 2019), with classification done manually (*Malfante et al.*, 2018) or with machine-learning based systems (*Carniel and Guzmán*, 2020). Some interesting results using ML are schemes that combine event detection and classification using Hidden Markov Models (HMM) (e.g., *Hammer et al.*, 2012; *Bhatti et al.*, 2016) as opposed to only classification.

Machine learning techniques have been used to improve debris flow warnings through the use of Random Forests (RF) (*Chmiel et al.*, 2021), and for snow avalanche detection using Support Vector Machines (SVM) in combination with acoustic array processing (*Thüring et al.*, 2015). In both cases, the methods reduced false positive detection considerably.

Supervised Learning - Signal Classification: In the volcanic event classification problem, a wide range of ML methods have been extensively tried for automatic classification tasks with varying success (e.g., *Malfante et al.*, 2018; *Titos et al.*, 2018; *Khan et al.*, 2019; *Martínez et al.*, 2021; *Manley et al.*, 2022). For example, Decision Trees (DT) have been applied to automatically detect precursors for real time-forecasting of sudden explosive eruptions from tremor signals (*Dempsey et al.*, 2020). Multi-Layer Perceptron (MLP) has been used to discriminate between volcanic hybrid events and explosion-generated earthquakes (e.g., *Esposito et al.*, 2018). A SVM classifier was tested against and outperformed Linear Discriminat Analysis (LDA), MLP, and Random Forests (RF) classifiers for the classification of volcano-tectonic (VT), long period (LP), Hybrid, Tremor and Explosive seismic events (*Lara et al.*, 2020). RF have also been tried for other subsets of events: VT, LP, Nested, Hybrid and Tornillo events, with good results (*Falcin et al.*, 2021).

Deep Neural Networks (DNN) have been applied with very good results. For instance, a Convolutional Autoencoder was used to find potential precursory signals to volcano eruption (*Zali et al.*, 2024). A Recurrent, Short-Long Term Memory (SLTM) and Gated Recurrent Unit (GRU) Artificial Neural Networks were tested with raw data and extracted features for LP, VT, Tremor, Hybrid and Silence event classification and showed that feature extraction outperforms analysis of raw data for this task (*Titos et al.*, 2019).

In the acoustic realm, the application of various advanced Neural Networks enabled classification of auroral, mountain associated, microbarom and volcanic signals from the Library of Typical Infrasonic Signals in the Comprehensive Nuclear-Test-Ban Treaty of the United Nations, and showed good performance (*Solomon et al.*, 2018). Automatic classification of volcano, tsunami and earthquake acoustic signals from the same data set was similarly explored with SVM (*Li et al.*, 2016). Another experiment using SVM showed excellent performance for subcrater explosion location indentification (*Toney et al.*, 2022).

Unsupervised Learning - Signal clustering: Supervised ML techniques usually learn from large amounts of data in order to better generalize predictions (*Géron*, 2019). When sufficient labelled data is unavailable or if in the search of specific families of events within a certain class is desired, unsupervised learning becomes useful. In the case of unsupervised learning, techniques can be applied to features or direct signals, or to a set of measurements of (dis)similarity between observations. In the latter case, this (dis)similarity can be calculated by statistical means, or by defining distances between signals or their features.

Hierarchical clustering of correlation values, as a measure of similarity, have been used in many settings. For instance, to identify families of icequakes (*Lamb et al.*, 2020) or to distinguish explosions between different volcanoes (*Ortiz et al.*, 2020). Other similarity measures such as Dynamic Time Warping have been applied to identify families of potential eruption precursory events (*Anzieta et al.*, 2019).

Self Organizing Maps (SOM) (*Kohonen*, 2001) have been used to study the relationship between volcanic degassing behaviour and seismo-acoustic signals in experimental settings (e.g., *Giudicepietro et al.*, 2021), to distinguish between seismic events and artificial explosions (*Kuyuk et al.*, 2011), and to identify signals related to seismic volcanic activity and noise (e.g., *Reyes and Jiménez Mosquera*, 2017). The K-means algorithm has also been used to study transitions in the style of activity of volcanoes associated to families of different acoustic signals (e.g., *Witsil and Johnson*, 2020; *Watson*, 2020).

Implementation

In this work, the Python programming language and several of its standard libraries (scipy, numpy, matplotlib, etc.), along with the specialized package "Obspy" (*Beyreuther et al.*, 2010) -a python framework for seismology- were used to explore and analyze the bulk of the data. Other specific packages/tools are mentioned in each Chapter when relevant. Additionally, to apply many different machine learning algorithms, I used the "Scikit-Learn" python package (*Pedregosa et al.*, 2011).

1.2 Thesis objective and specific goals

Data analysis, and Machine Learning in particular, continue to be fruitful areas of research since it is expected that with more widespread instrumentation continuously recording underground and atmospheric phenomena, automated systems will become so effective that more rapid and reliable forecasting of natural hazards will be possible. As such, it is worth trying to improve all steps towards the creation of robust systems dealing with different tasks related to various types of datasets, whether they be locked or unlocked, or whether there is plenty or little prior knowledge about them. At the same time, there are still many regions in the world without comprehensive monitoring networks, with perhaps only one sensor to monitor entire areas. Thus, the general objective of this research is to design or improve data analysis and/or machine learning tools and methods that can be applied to recordings from single stations, to deepen the knowledge of various types of datasets with increasing uncertainties as shown in Figure 1.1.

Specific Goals

Locked datasets

When locked datasets are fully determined in terms of the number of signals they possess, and their events are catalogued with more than one type or class, they are useful as the potential input for training automatic classification systems. As such, I aim to assess the quality of the information these datasets hold. More precisely, I want to answer the following questions:

1. Can the quality of a given catalogue (i.e., how well are the events within labelled) be assessed?
2. Can one determine which of the events within a catalogue may be mislabelled?
3. Can a degree of uncertainty be assigned to the assumption that each event's label may or may not be mislabelled?
4. Are the previous questions answerable in an objective and fast way?

For this study I used two publicly available seismic datasets with characteristics not typically found in similar open-access data sources –one from Cotopaxi volcano, Ecuador and another from Llaima volcano, Chile. The datasets possess selected waveforms, along with their respective class membership labels as determined by the observatories in charge of monitoring them, and I applied a machine learning dataset-cleaning procedure to answer the posed questions.

Unlocked dataset with a human created catalogue

Unlocked datasets imply that new information can be extracted from them by defining tasks according to previous knowledge. If a dataset possesses a human defined catalogue, a natural task is to build an expanded catalogue based upon the information found by the operators, and then to try to gain more knowledge from this new dataset. The precise sequence of tasks are:

1. Take advantage of traditional data-analysis tools for preliminary detection of events.
2. Use the human created catalogue to train supervised classifiers that help refine the preliminary detections.
3. Use unsupervised machine learning techniques to assert the capabilities of gaining deeper knowledge from the newly identified events.

4. Combine deep-learning with unsupervised learning to allow faster retrieval of the deeper knowledge.

For these tasks I used nearly 10 years of continuous acoustic recordings from Tungurahua volcano, Ecuador, and an explosion catalogue constructed by a human operator for this dataset. This fairly unique **unlocked** dataset served as a means to test the capabilities of various machine learning tools.

Unlocked dataset with no prior knowledge

When there is no previous knowledge for an unlocked catalogue, the first task is to establish a baseline of information, so that other applications can be built from there. In this case the objectives are:

1. Develop an efficient algorithm to extract useful signals from background noise to monitor river-flow discharge rates that can be applied in-situ.

2. Make use of this baseline information to identify/clean potentially anomalous data. This work is conducted by applying data-analysis tools to acoustic data collected at the confluence point of the Squamish River and Mud Creek at Mount Cayley, British Columbia, Canada during the spring-summer of 2021.

1.3 Thesis outine

The rest of the Chapters of this thesis are organized as follows: In Chapter 2, a machine learning procedure was implemented to assess the quality of two publicly-available seismic volcanic locked datasets, one under suspicion of having potentially misclassified events, and the other regarded as clean and used for reference. The results for this chapter have already been published. In Chapter 3, an unlocked acoustic dataset with a catalogue of volcanic explosions previously identified by a human analyst was explored and expanded by using a combination of classical data analysis techniques along with supervised, unsupervised and deep learning algorithms and these results are currently under review as of the writing of this thesis. In Chapter 4, a novel acoustic unlocked dataset with no previous analysis was surveyed, and a procedure for baseline river streamflow characterization was determined and further results are expected to be published in the future. Chapter 5 summarizes the results of the previous chapters and describes the most important lessons learned from the whole study.

References

Abraham, K., M. Abdelwahab, and M. Abo-Zahhad (2024), Classification and detection of natural disasters using machine learning and deep learning techniques: A review, *Earth Science Informatics*, *17*, 869–891, doi:10.1007/s12145-023-01205-2.

Allstadt, K. E., R. S. Matoza, A. B. Lockhart, S. C. Moran, J. Caplan-Auerbach, M. M. Haney, W. A. Thelen, and S. D. Malone (2018), Seismic and acoustic signatures of surficial mass movements at volcanoes, *Journal of Volcanology and Geothermal Research*, *364*, 76–106, doi:https://doi.org/10.1016/j.jvolgeores.2018.09.007.

Anderson, J. F., J. B. Johnson, D. C. Bowman, and T. J. Ronan (2018), The Gem Infrasound Logger and Custom-Built Instrumentation, *Seismological Research Letters*, *89*(1), 153–164, doi:10.1785/0220170067.

Anthony, R. E., A. T. Ringler, D. C. Wilson, and E. Wolin (2019), Do low-cost seismographs perform well enough for your network? An overview of laboratory tests and field observations of the OSOP raspberry shake 4D, *Seismological Research Letters*, *90*(1), 219–228, doi:10.1785/0220180251.

Anzieta, J. C., H. D. Ortiz, G. L. Arias, and M. C. Ruiz (2019), Finding Possible Precursors for the 2015 Cotopaxi Volcano Eruption Using Unsupervised Machine Learning Techniques, *International Journal of Geophysics*, *2019*, 1–8, doi:10.1155/2019/6526898.

Berk, R. A. (2016), *Statistical Learning from a Regression Perspective*, Springer Texts in Statistics, Springer International Publishing, Cham, doi:10.1007/978-3-319-44048-4.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010), ObsPy: A Python Toolbox for Seismology, *Seismological Research Letters*, *81*(3), 530–533, doi:10.1785/gssrl.81.3.530.

Bhatti, S. M., M. S. Khan, J. Wuth, F. Huenupan, M. Curilem, L. Franco, and N. B. Yoma (2016), Automatic detection of volcano-seismic events by modeling state and event duration in hidden Markov models, *Journal of Volcanology and Geothermal Research*, *324*, 134–143, doi:10.1016/j.jvolgeores.2016.05.015.

Bueno, A., A. Diaz-Moreno, I. Álvarez, A. De la Torre, O. D. Lamb, L. Zuccarello, and S. De Angelis (2019), Vineda—volcanic infrasound explosions detector algorithm, *Frontiers in Earth Science*, *7*, 1–8, doi:10.3389/feart.2019.00335.

Carniel, R., and S. R. Guzmán (2020), Machine learning in volcanology: A review, in *Updates in Volcanology*, edited by K. Németh, chap. 5, pp. 1–26, IntechOpen, Rijeka, doi:10.5772/intechopen.94217.

Chiarle, M., M. Geertsema, G. Mortara, and J. J. Clague (2021), Relations between climate change and mass movement: Perspectives from the canadian cordillera and the european alps, *Global and Planetary Change*, *202*, 103,499, doi:https://doi.org/10.1016/j.gloplacha.2021.103499.

Chmiel, M., F. Walter, M. Wenner, Z. Zhang, B. W. McArdell, and C. Hibert (2021), Machine Learning Improves Debris Flow Warning, *Geophysical Research Letters*, *48*(3), doi:10.1029/2020GL090874.

Chouet, B. A., and R. S. Matoza (2013), A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption, *Journal of Volcanology and Geothermal Research*, *252*, 108–175, doi:10.1016/j.jvolgeores.2012.11.013.

Connor, C., M. Bebbington, and W. Marzocchi (2015), Probabilistic Volcanic Hazard Assessment, in *The Encyclopedia of Volcanoes*, pp. 897–910, Elsevier, doi:10.1016/B978-0-12-385938-9.00051-1.

Cortes, G., M. C. Benitez, L. Garcia, I. Alvarez, and J. M. Ibanez (2016), A Comparative Study of Dimensionality Reduction Algorithms Applied to Volcano-Seismic Signals, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(1), 253–263, doi:10.1109/JSTARS.2015.2479300.

Dannemann, F., C. Koch, E. Berg, S. Arrowsmith, and S. Albert (2023), A New Decade in Seismoacoustics (2010–2022), *Bulletin of the Seismological Society of America*, *113*(4), 1390–1423, doi:10.1785/0120220157.

Dempsey, D. E., S. J. Cronin, S. Mei, and A. W. Kempa-Liehr (2020), Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand, *Nature Communications*, *11*(1), 3562, doi:10.1038/s41467-020-17375-2.

Esposito, A. M., F. Giudicepietro, S. Scarpetta, and S. Khilnani (2018), A Neural Approach for Hybrid Events Discrimination at Stromboli Volcano, pp. 11–21, doi:10.1007/978-3-319-56904-8_2.

Falcin, A., J.-P. Métaxian, J. Mars, É. Stutzmann, J.-C. Komorowski, R. Moretti, M. Malfante, F. Beauducel, J.-M. Saurel, C. Dessert, A. Burtin, G. Ucciani, J.-B. de Chabalier, and A. Lemarchand (2021), A machine-learning approach for automatic classification of volcanic seismicity at La Soufrière Volcano, Guadeloupe, *Journal of Volcanology and Geothermal Research*, *411*, 107,151, doi:10.1016/j.jvolgeores.2020.107151.

Freire, S., A. J. Florczyk, M. Pesaresi, and R. Sliuzas (2019), An improved global analysis of population distribution in proximity to active volcanoes, 1975–2015, *ISPRS International Journal of Geo-Information*, *8*(8), doi:10.3390/ijgi8080341.

Gariano, S. L., and F. Guzzetti (2016), Landslides in a changing climate, *Earth-Science Reviews*, *162*, 227–252, doi:10.1016/j.earscirev.2016.08.011.

Géron, A. (2019), *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, second ed., O'Reilly, CA 95472.

Giudicepietro, F., A. M. Esposito, L. Spina, A. Cannata, D. Morgavi, L. Layer, and G. Macedonio (2021), Clustering of Experimental Seismo-Acoustic Events Using Self-Organizing Map (SOM), *Frontiers in Earth Science*, *8*, doi:10.3389/feart.2020.581742.

Hammer, C., M. Beyreuther, and M. Ohrnberger (2012), A Seismic-Event Spotting System for Volcano Fast-Response Systems, *Bulletin of the Seismological Society of America*, *102*(3), 948–960, doi:10.1785/0120110167.

Jain, H., R. Dhupper, A. Shrivastava, D. Kumar, and M. Kumari (2023), Leveraging machine learning algorithms for improved disaster preparedness and response through accurate weather pattern and natural disaster prediction, *Frontiers in Environmental Science*, *11*, doi:10.3389/fenvs.2023.1194918.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning*, *Springer Texts in Statistics*, vol. 103, Springer New York, New York, NY, doi: 10.1007/978-1-4614-7138-7.

Khan, M. S., M. Curilem, F. Huenupan, M. F. Khan, and N. Becerra Yoma (2019), A Signal Processing Perspective of Monitoring Active Volcanoes [Applications Corner], *IEEE Signal Processing Magazine*, *36*(6), 125–163, doi:10.1109/MSP.2019.2930427.

Kohonen, T. (2001), *Self-Organizing Maps*, *Springer Series in Information Sciences*, vol. 30, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-56927-2.

Kuyuk, H. S., E. Yildirim, E. Dogan, and G. Horasan (2011), An unsupervised learning algorithm: application to the discrimination of seismic events and quarry blasts in the vicinity of Istanbul, *Natural Hazards and Earth System Sciences*, *11*(1), 93–100, doi: 10.5194/nhess-11-93-2011.

Lamb, O., J. Lees, L. F. Marin, J. Lazo, A. Rivera, M. Shore, and S. Lee (2020), Investigating potential icequakes at Llaima volcano, Chile, *Volcanica*, *3*(1), 29–42, doi:10.30909/vol.03.01.2942.

Lara, P. E. E., C. A. R. Fernandes, A. Inza, J. I. Mars, J.-P. Metaxian, M. Dalla Mura, and M. Malfante (2020), Automatic Multichannel Volcano-Seismic Classification Using Machine Learning and EMD, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 1322–1331, doi:10.1109/JSTARS.2020.2982714.

Li, M., X. Liu, and X. Liu (2016), Infrasound signal classification based on spectral entropy and support vector machine, *Applied Acoustics*, *113*, 116–120, doi:10.1016/j.apacoust.2016.06.019.

Liu, D., D. Tang, S. Zhang, X. Leng, K. Hu, and L. He (2021), Method for feature analysis and intelligent recognition of infrasound signals of soil landslides, *Bulletin of Engineering Geology and the Environment*, *80*(2), 917–932, doi:10.1007/s10064-020-01982-w.

Malfante, M., M. Dalla Mura, J.-P. Metaxian, J. I. Mars, O. Macedo, and A. Inza (2018), Machine Learning for Volcano-Seismic Signals: Challenges and Perspectives, *IEEE Signal Processing Magazine*, *35*(2), 20–30, doi:10.1109/MSP.2017.2779166.

Manley, G. F., T. A. Mather, D. M. Pyle, D. A. Clifton, M. Rodgers, G. Thompson, and J. M. Londoño (2022), A deep active learning approach to the automatic classification of volcano-seismic events, *Frontiers in Earth Science*, *10*, 1–13, doi:10.3389/feart.2022.807926.

Martínez, V. L., M. Titos, C. Benítez, G. Badi, J. A. Casas, V. H. O. Craig, and J. M. Ibáñez (2021), Advanced signal recognition methods applied to seismo-volcanic events from planchon peteroa volcanic complex: Deep neural network classifier, *Journal of South American Earth Sciences*, *107*, 103,115, doi:https://doi.org/10.1016/j.jsames.2020.103115.

Mousavi, S. M., and G. C. Beroza (2023), Machine learning in earthquake seismology, *Annual Review of Earth and Planetary Sciences*, *51*(1), 105–129, doi:10.1146/annurev-earth-071822-100323.

Ortiz, H. D., R. S. Matoza, C. Garapaty, K. Rose, P. Ramón, and M. C. Ruiz (2020), Multiyear regional infrasound detection of Tungurahua, El Reventador, and Sangay volcanoes in Ecuador from 2006 to 2013, p. 022003, doi:10.1121/2.0001362.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, *12*, 2825–2830.

Petley, D. (2012), Global patterns of loss of life from landslides, *Geology*, *40*(10), 927–930, doi:10.1130/G33217.1.

Ponziani, M., D. Ponziani, A. Giorgi, H. Stevenin, and S. M. Ratto (2023), The use of machine learning techniques for a predictive model of debris flows triggered by short intense rainfall, *Natural Hazards*, *117*, 143–162, doi:10.1007/s11069-023-05853-x.

Raspberry Shake (2024), Raspberry Shake, 2024, `https://raspberryshake.org/`.

Reyes, J. A., and C. J. Jiménez Mosquera (2017), Non-supervised classification of volcanic-seismic events for tungurahua-volcano ecuador, in *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–6, doi:10.1109/ETCM.2017.8247446.

Shearer, P. M. (2009), *Introduction to Seismology*, Cambridge University Press, Cambridge, doi:10.1017/CBO9780511841552.

Smithsonian Institution (2024), Global Volcanism Program, 2024, `https://volcano.si.edu/`.

Solomon, M., K. Smith, K. Bryan, A. O. Smith, D. A. Clauter, and A. M. Peter (2018), Infrasound threat classification: a statistical comparison of deep learning architectures, in *Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XIX*, edited by A. W. Fountain, J. A. Guicheteau and C. R. Howle, p. 42, SPIE, doi:10.1117/12.2304030.

Tehrani, F. S., M. Calvello, Z. Liu, L. Zhang, and S. Lacasse (2022), Machine learning and landslide studies: recent advances and applications, *Natural Hazards*, *114*, 1197–1245, doi:10.1007/s11069-022-05423-7.

Thüring, T., M. Schoch, A. van Herwijnen, and J. Schweizer (2015), Robust snow avalanche detection using supervised machine learning with infrasonic sensor arrays, *Cold Regions Science and Technology*, *111*, 60–66, doi:10.1016/j.coldregions.2014.12.014.

Titos, M., A. Bueno, L. García, and C. Benítez (2018), A deep neural networks approach to automatic recognition systems for volcano-seismic events, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(5), 1533–1544, doi:10.1109/JSTARS.2018.2803198.

Titos, M., A. Bueno, L. Garcia, M. C. Benitez, and J. Ibanez (2019), Detection and Classification of Continuous Volcano-Seismic Signals With Recurrent Neural Networks, *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 1936–1948, doi:10.1109/TGRS.2018.2870202.

Toney, L., D. Fee, A. Witsil, and R. S. Matoza (2022), Waveform Features Strongly Control Subcrater Classification Performance for a Large, Labeled Volcano Infrasound Dataset, *The Seismic Record*, *2*(3), 167–175, doi:10.1785/0320220019.

Wasserman, L. (2004), *All of Statistics*, Springer Texts in Statistics, Springer New York, New York, NY, doi:10.1007/978-0-387-21736-9.

Watson, L. M. (2020), Using unsupervised machine learning to identify changes in eruptive behavior at mount etna, italy, *Journal of Volcanology and Geothermal Research*, *405*, 107,042, doi:https://doi.org/10.1016/j.jvolgeores.2020.107042.

Witsil, A. J., and J. B. Johnson (2020), Analyzing continuous infrasound from Stromboli volcano, Italy using unsupervised machine learning, *Computers & Geosciences*, *140*, 104,494, doi:10.1016/j.cageo.2020.104494.

Zali, Z., M. Mousavi, M. Ohrnberger, E. P. Eibl, and F. Cotton (2024), Tremor clustering reveals pre-eruptive signals and evolution of the 2021 Geldingadalir eruption of the Fagradalsfjall Fires, Iceland, *Communications Earth & Environment*, *5*, 1–11, doi:10.1038/s43247-023-01166-w.

Chapter 2

Cleaning volcano-seismic event catalogues: a machine learning application for robust systems and potential crises in volcano observatories

Anzieta, J., D. Pacheco, G. Williams-Jones, and M.C. Ruiz (2023), Cleaning volcano-seismic event catalogues: a machine learning application for robust systems and potential crises in volcano observatories, *Bulletin of Volcanology, 85*, 1–16, doi:10.1007/s00445-023-01674-9.

2.1 Introduction

The detection and identification of different volcano-seismic events has been used for decades as a tool to characterize the state of volcanic activity, with early observatories dating back to the end of the nineteenth and beginning of the twentieth centuries (*McNutt and Roman*, 2015). Thus, extensive research has focused on distinguishing and describing different classes of events for a given volcano so that today a few types have become standard in volcano observatories around the world. Common types of events related to volcanic activity are low frequency (or long period—LP), high frequency (or volcano tectonic—VT), tremor (TR), hybrid (HB), very long period (VLP) and explosive (EX) events (*McNutt et al.*, 2015; *Chouet and Matoza*, 2013), pyroclastic density currents (PDC—which are surficial events that result from lateral blasts, boiling over, the collapse of lava domes, and eruptive columns;

Dufek et al., 2015), and more sub-classifications that can be volcano specific (e.g., Tornillos which are a particular kind of highly harmonic, long-duration LP; *Narváez et al.*, 1997). Other events can be detected by volcano-seismic networks that are not necessarily related to volcanic activity, including icequakes (IQ), rockfalls, lahars, regional quakes (RE), and noise (NS). Among all the classes of seismic events, LPs and VTs typically stand out given their relation to important physical processes (*Chouet and Matoza*, 2013) and as diagnostic signals for volcanic activity. LP origins are often associated with, but not limited to, conduit resonances due to fluid movement within the volcanic edifice (*Chouet*, 1996), oscillations due to pumping mechanisms in shallow fractures, brittle failure of flowing magma undergoing glass transition near the conduit wall, escape of gases through fractures (*Chouet and Matoza*, 2013), or fast and large input of gas into gas pockets below a permeable sealing cap (*Girona et al.*, 2019), while VTs are often explained as brittle fractures in the edifice driven by local stress changes or magma movement (*McNutt*, 2005). We acknowledge that in reality there is a continuum of events between VTs and LPs, and a hard classification is complex for many reasons, e.g., VTs may transition to LPs due to seismic attenuation or changes in the volcano as shown experimentally and in simulations (*Benson et al.*, 2008; *Clarke et al.*, 2021). Nevertheless, the distinction of such types of events is still central in volcano monitoring, and a thorough tracking of the occurrence of these events can be used as a tool for volcanic eruption forecasting (*Bell et al.*, 2018), or reawakening estimation (*Kilburn*, 2003). Volcano observatories review and evaluate the volcanic activity by creating and examining catalogues, and monitoring tasks are possible when catalogues are accurate and comprehensive enough (*Boué et al.*, 2016).

These short-lived events (LPs and VTs) are in principle distinguishable reasonably quickly through human inspection, by looking at the seismogram appearance (*McNutt and Roman*, 2015), or at other parameters such as frequency content. However, it can become exceedingly difficult for analysts to maintain a correct identification when short-handed or during unexpected circumstances, such as a sudden volcano reactivation or personnel changes. For these and other reasons, particularly the decreasing costs of powerful computers, the use of automated systems is becoming increasingly prevalent at observatories for monitoring and forecasting volcano activity (*Pallister et al.*, 2019). Technological and methodological advances have led to the exploration and use of an ever-increasing wealth of machine learning techniques that explicitly or implicitly extract and select several parameters from the data to achieve highly performing systems (*Malfante et al.*, 2018; *Khan et al.*, 2019). Well-trained automated systems have many advantages compared to human analysts including but not limited to unbiased (i.e., not influenced by analyst subjectivity; *Falcin et al.*, 2021), consistent over time (i.e., repeatable performance after implementation when volcanic conditions remain the same; *Bueno et al.*, 2020), and faster execution. Supervised machine learning classifiers, which are generally the core of automated recognition systems, rely on the labels provided in a catalogue to be trained, and thus, if a catalogue is biased, its derived models will also be biased. These systems usually require copious amounts of well-labelled data to be constructed as unbiased as possible, and a catalogue's integrity can be compromised due to unexpected situations like a sudden volcano reactivation, or when analysts or sensors are changed. Mislabelling is a significant source of model misclassification with seismic data (*Linville et al.*, 2020), and thus, clean, unbiased catalogues are necessary to train adequate models. However, cleaning an existing catalogue can be a costly procedure since it requires human and time resources for the revision of each event. It is noteworthy that Bayesian deep learning approaches, using transfer (i.e., using data from one volcano to train systems for another e.g., *Bueno et al.*, 2020) and active (i.e., systems that interactively query users to label new observations, e.g., *Bueno et al.*, 2022) learning, have been successfully applied (even on small datasets) to determine the potential uncertainty in event labels as volcanic activity evolves, simultaneously as the events are being catalogued. These procedures allow mapping uncertainties in the class memberships to later retrain the model. Another advancement in the area consists of augmentation of the dataset so that state-of-the-art classifiers can learn from large amounts of known good quality data (*Witsil et al.*, 2022). Nevertheless, they still require clean initial catalogues, posterior human validation for unknown events, or complex system fine-tuning.

To identify and overcome potential classification issues in volcanic catalogues, we propose an easy and rapid method for volcano observatories to assess the quality of a volcano catalogue in general, and of the specific events in it. Hence, events with "low confidence" can be revised by a human to improve the catalogue with less effort than going through the whole catalogue. Herein, we apply techniques developed in the machine learning and data mining domains to deal with potentially noisy labels in open catalogues with their respective waveforms from two different volcanoes: Cotopaxi volcano in Ecuador, and Llaima volcano in Chile (locations shown in Fig. 1). Both are ice-clad, subduction-driven stratovolcanoes that have been recently active, and thus, their catalogues are ideal for methodological performance comparisons.

The rest of this work is organized as follows: In the "Methodology" section, we present the procedural framework of the technique applied to clean databases in general along with proposed extensions to it. In the "Data processing and models" section, we detail the specific datasets used, enumerate the chosen models along with their parameters, and describe the data standardization steps for the application of the method. In the "Results" section, we present a subsample showing how events are evaluated to determine if they are potentially mislabelled, and we summarize the general results from both catalogues. In the "Discussion" section, we comment on and show the products and consequences of the applied method comparing both datasets. The "Conclusion" section discusses the potential extensions to this work.

2.2 Methodology

The accuracy of the classification models used for the creation of automated systems can be impacted by two types of noise in the training datasets: *label* or *class* noise, and *attribute* or *feature* noise (*García et al.*, 2015; *Frenay and Verleysen*, 2014). In the volcano-seismic event classification task, label refers to the specific family or class (type of event) an event has been designated as, and attributes or features to the different quantities that are extracted from the waveforms (e.g., peak frequency, duration, amplitude skewness, etc.). Both types of noise can have varying impact on the performance of the classifiers. If the noisy attributes are highly correlated with the labels, they can be more impactful than label noise (*García et al.*, 2015). In contrast, if the noise in the attributes is not widespread or strong enough, using a reasonable number of uncorrelated features could mitigate the effects of such noise, as opposed to noise in labels that have direct impact on the learning process (*Frenay and Verleysen*, 2014). Here, we implement a strategy that can simultaneously address the hypothetical existence of both types of noise in the context of volcano-seismic event catalogues. There are many strategies to deal with noisy datasets. One strategy consists of creating stable classifiers that are mildly affected by relatively low levels of noise in the data. These classifiers are trained with the data as is and are designed to perform well for future observations (i.e., robust learners; *García et al.*, 2015; *Frenay and Verleysen*, 2014). In the context of volcano-seismic event classification, however, it is not only desired to create systems that can identify future events properly, but also to have correct historical records of all events because they could have statistical value in hindsight (*Bell et al.*, 2018). Therefore, it is more relevant to apply a strategy to correct wrongly assigned labels by assessing the quality of a catalogue as a whole and of each event one by one. This task falls into the "data cleansing" or "noise filtering" domain. To do this, we apply a modified version of voting filtering (as first described by *Brodley and Friedl*, 1999) also known as ensemble filters (*García et al.*, 2015). The basic idea of the ensemble filters scheme is to create different predictive models in several disjoint partitions (non-overlapping subsets) of the data, train the models in one partition of the data, and then quantify how the different models "see" observations in the complementary left-out subsets in the dataset. Each event's quality or confidence is then evaluated based on the number of predictions from the models that agree with the original labels. It is expected that the majority of predictions from the different classifiers will agree with the original label for a given "unseen" instance if the original label is consistent with the information learned from the observations in the training partitions, assuming that the noise is not too large (*Brodley and Friedl*, 1999). Conversely, noisy observations will show a disagreement between the original label and some of the models' predictions if the information from the training partitions is strongly incompatible. Consequently, one can define an instance's quality based on the number of agreements and disagreements and define criteria for their eventual correction or definitive removal from the dataset.

In our approach, we modify the voting filtering method in four aspects by (1) using a LOO

(Leave-One-Out, or “one versus the rest”) partition scheme instead of partitions of arbitrary size; (2) extracting and using mostly uncorrelated "sufficient" sets of features instead of a single fixed set; (3) requiring a few "shallow learners" (*Deng and Yu*, 2014) for supervised learning; and (4) exploring unsupervised and/or transfer learning for the raw data if possible. Before describing the procedure in detail, we present an analogy illustrating the scheme.

In the supervised part of the scheme, we choose s shallow supervised classifiers that act as independent volcano analysts; they will learn from the both the catalogue's labels and information extracted from their corresponding waveforms and separately evaluate each event. The information is extracted in different features that are grouped in g complementary sets so that each set describes the n events sufficiently. This leads to several algorithms repeating while learning from different features at a time. To learn, one of the events is excluded, and each "analyst" will be trained with each of the feature sets and the labels of the other events and predict the label of the excluded event based on what they learned. This process is repeated for each of the events so that each is judged almost identically to the others. This means the full dataset minus one event acts as the training partition each time leading to n different training partitions and helping to mitigate issues when a class has too few events compared to others. Shallow learners are thus required for speed since $s \times g \times n$ supervised models are trained, but also because it makes the models easy to implement in popular programming and machine learning environments (e.g., Python, R, Weka, Orange, etc.). They may need to be mildly regularized to help avoid over-fitting into the noisy instances, even if label noise might lead internal validation procedures –such as bootstrapping– to the election of non-optimal hyperparameters (*Frenay and Verleysen*, 2014). The g sets of features are proposed to extract diverse information from the dataset to evaluate each instance while reducing potential attribute noise or bias. We argue that by creating many similar models with separate feature groups, we simultaneously achieve a desirable diversity (complementary classifiers as in *Sáez et al.*, 2013) while not overloading the procedure with excessively diverse ensembles that may lead to a decrease in noise detection (*Sluban and Lavrač*, 2015).

The u unsupervised classifiers can be thought of as analysts that are only given information about the relations between events, either similarities or differences –and not the labels– and are expected to group them "blindly". Since unsupervised models are inherently different to supervised models (for a detailed review see *Carniel and Guzmán*, 2020), they will find class memberships independently of the catalogue's original labels (with or without label noise). These classifications may occasionally disagree with human perception or definitions of classes and are in principle "nameless" in the sense that they lack direct one-to-one correspondence to the supervised classes. Thus, when exploring unsupervised classifiers, the "best" class assignations should first be identified by performing some kind of validation scheme such as examination of the confusion matrix between the original labels and the labels from the clustering. If the unsupervised algorithm has unfixed hyperparameters (e.g., epsilon in "Density-Based Spatial Clustering of Applications with Noise", DBSCAN), a search for the best hyperparameter(s) should be done, again by performing some validation of choice with the known labels. While unsupervised learning is generally less successful than supervised learning at matching human expertise, it still groups together instances that are very similar and leaves out from a coherent cluster instances that are not so similar (not necessarily but potentially noisy).

Finally, using t transfer learning models is analogous to asking analysts to label events in one catalogue using knowledge learned from another catalogue. One should consider that knowledge from other volcanoes may not be directly applicable to the catalogue of interest without proper unification of data characteristics from different sources (*Keramati et al.*, 2023). In this work, we did not use transfer learning models because some do not currently achieve state-of-the-art performance even in clean catalogues (*Cortés Moreno et al.*, 2019) or are not yet easy/fast enough to apply for the catalogue cleaning task (*Bueno et al.*, 2020; *Keramati et al.*, 2023). However, in contexts where the revision of a catalogue is not urgent, they could be applied, even given the potential lower performance, to check for highly consistent events that could be seen as "extremely archetypal" or "globally consistent".

Both unsupervised and transfer models have the advantage that they do not need to be retrained each time an event is labelled. In the end, each event will have m quasi-independent predictions where $m = g \times s + u + t$ coming from $g \times s$ supervised, u unsupervised, and t transfer classifiers. A comparison of all these predictions with the original label helps us define the following unique magnitude, c_i:

$$c_i = \frac{1}{m} \sum_{j}^{m} \mathbb{I}_{i-agree} j \tag{2.1}$$

as the "consistency" of the i-th event. Here, $\mathbb{I}$ is the indicator function based on the m predicted labels that takes value 1 if the j prediction agrees with the i-th original label, and 0 if it does not. This magnitude is the arithmetic mean of agreements per event, and while it is akin to the definition of accuracy, we use the term "consistency" as we do not actually know in principle if the models' predictions are accurate or not due to noise; we can thus only talk about consistency between the original label and the models' labels. The consistency value of each observation represents its quality given the rest of the data and the models, and all of them describe the quality of the catalogue in general. Thus, following Eq. 1, a "highly confident" instance is one where most or all the models agree with the original label ($c \sim 1$), while an instance with a few disagreements might require inspection. The specific steps to implement this approach are summarized in the following sequence:

0. Preprocess the data accordingly (e.g., trim data to contain only signals, filter/detrend when needed, scale amplitudes).

1. Extract g different groups of features from the n seismic waveforms.

2. Calculate a (dis)similarity measure matrix between all the events in the catalogue.

3. Select s low complexity supervised classifiers and explore hyperparameters to be used in the predictions of each of the n observations.

4. Select and apply u unsupervised classifiers, and when required, optimize hyperparameters by comparing the clustering labels with the original labels and selecting values that give the best accuracy or other performance measures.

5. Select t transfer learning models to the events making sure the transfer learning models are compatible with the dataset.

6. After choosing hyperparameters from step 3, for each event, train the different models with the rest of the data and predict its label.

7. For each observation, assign the appropriate label determined from the comparison and optimization of the clustering labels with the original labels.

8. Apply the t transfer learning models to each event and obtain the label predictions.

9. Calculate the consistency of each event for all the predictions as defined in Equation 2.1.

As numerous models may introduce too many uncertainties in a catalogue's quality assessment, one could opt to stay with the more classic voting filtering scheme and use only supervised models (therefore abandoning steps 2, 4, 5, 7 and 8 in the implementation sequence above). The inclusion or exclusion of models in the implementation depends on how much time researchers can devote to the catalogue's quality assessment. A definite exclusion of a model and its predictions can be avoided, if desired, if each model is given an importance value in Equation 2.1 in a weighted average fashion (weighted majority voting; *Sáez et al.*, 2013; *García et al.*, 2015) –this should depend on the trust in the model or comparable performance measures instead of discarding altogether. Here, we use Equation 2.1 as is, and assume all models are of equal importance for each consistency calculation. Finally, a catalogue's quality can be assessed from the consistencies of all the events and the different statistics that can be derived from them. As a first action, after performing all steps, lower quality events can be manually inspected (by one or more human analysts) for their removal or relabelling if possible. Additional insights obtained from full catalogue or class consistency statistics are described in the "Results" and "Discussions" sections.

2.3 Data processing and models

The open catalogues studied here for Cotopaxi and Llaima volcanoes (Figure 2.1) contain seismic waveforms with their respective labels provided in different formats. For access to these catalogues, see the Data availability section. The Cotopaxi dataset has waveforms recorded from two instruments, one short period and one broadband, while the Llaima

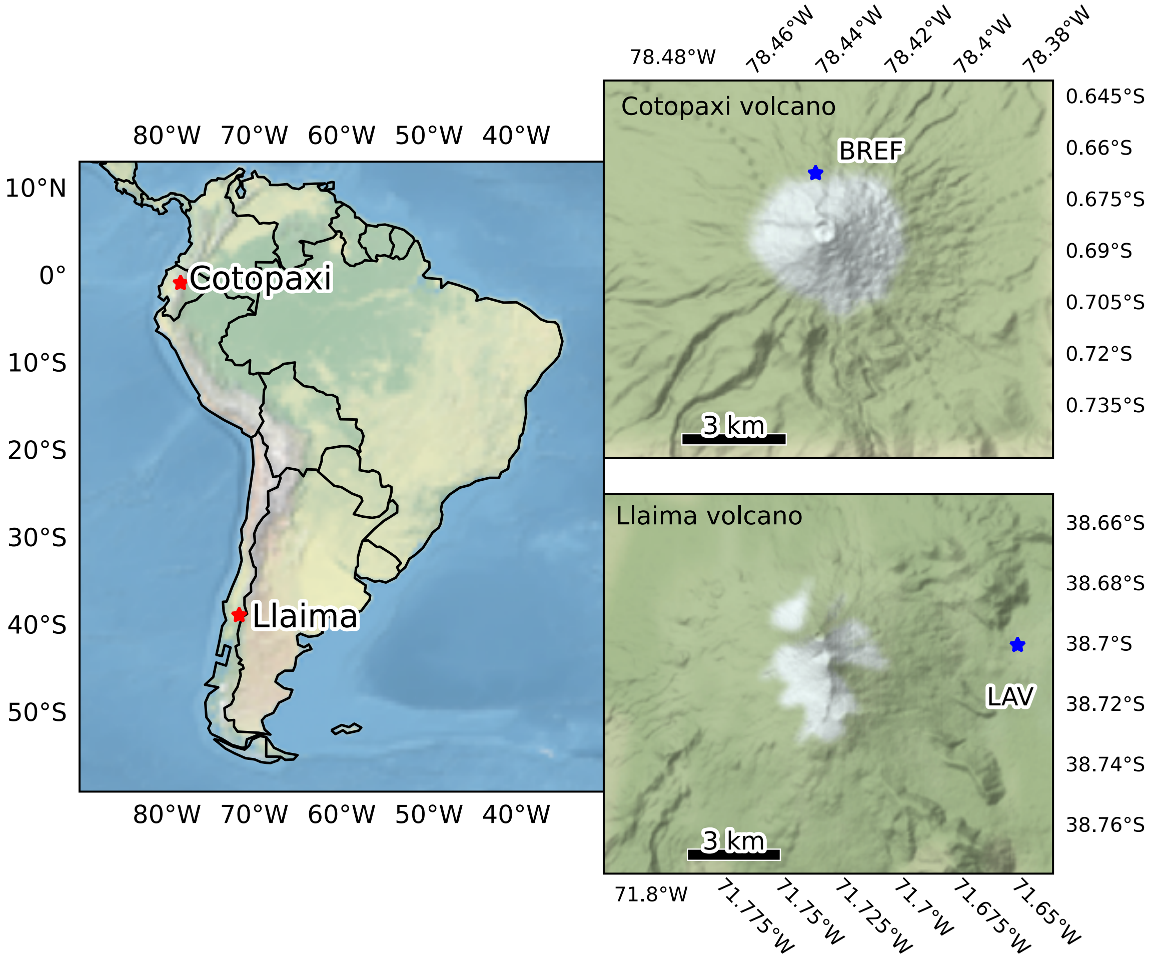

Figure 2.1: Locations for Cotopaxi and Llaima volcanoes in South America shown as red stars. Names and locations of the seismic stations used in this study are indicated by blue stars. Maps created using Cartopy, *Met Office* (2010 - 2015)

dataset has data from one broadband sensor only. Since the Cotopaxi events from both sensors do not overlap (from 2012 to 2014 for the short period sensor and from 2018 to 2019 for the broadband sensor), we restrict our analysis to the broadband seismometer because that instrument possesses more events and a wider frequency band for analysis. In both datasets, only the vertical component is provided since it contains all the necessary information to classify the event types (*Canário et al.*, 2020). In contexts where overlapping data is available from various seismic stations, the procedure can be implemented for each instrument's records independently by adding the different predictions to the consistency calculation in Equation 2.1 for each event and weighting them according to the confidence on each sensor if necessary. Cotopaxi events are provided along with 84 pre-calculated attributes for each event (*Pérez et al.*, 2020). Llaima events have been pre-processed to have fixed amplitudes and zero-padded to last 60 s (*Canário et al.*, 2020). While both catalogues possess various classes of events, we restrict this analysis to LPs and VTs because they are the most relevant for volcanic precursory activity, they are the most numerous in the Cotopaxi dataset (619 LPs and 46 VTs out of 706 total events), and one class vs one class noise detection systems yield better performances than multi-class approaches (*García et al.*, 2015). Out of 3592 total events in the Llaima dataset (*Canário et al.*, 2020), we used the 1310 LPs and 304 VTs available. To make the method comparable between both catalogues and reproducible for others, we preprocess the data so that events have similar properties before the feature extraction step. Similar steps are recommended for other datasets as they intend to optimize the performance of the models. Waveforms fed to the system should only contain signals from the event itself and not pre- and post-event noise since that could alter performance. Thus, for the Cotopaxi catalogue, we removed the first and last 10 s of noise that was included in the raw signals, and for the Llaima catalogue, we removed the zero-padded tail of the signals. In real-world applications, this requirement can be fulfilled by applying automatic event onset and end detection algorithms such as STA/LTA if waveforms contain signals other than the events. Additionally, the Llaima dataset was resampled from 100 to 50 Hz to ensure direct procedural comparisons between datasets. Waveforms from both datasets are provided pre-filtered to have frequencies higher than 1 Hz (*Pérez et al.*, 2020; *Canário et al.*, 2020), and in general, we recommend filtering the data to reduce the effect of the background noise that may alter analyses, especially in events with low signal to noise ratio. Finally, we normalize the traces with respect to the maximum absolute amplitude so that events are characterized based on the contrast of attributes, not magnitude/scale.

After preprocessing, we extracted three complementary sets of attributes from independent research groups that have been tested and found adequate for volcanic event classification. The three sets of features are selected and applied independently: the first following those proposed by *Watson* (2020), the second following *Titos et al.* (2018), and the third by combining features inspired by *Soto et al.* (2018) and *Reyes and Jiménez Mosquera* (2017), as described in Appendix B.1. While the features proposed by *Watson* (2020) were originally used for volcano-acoustic signals, they comprise a subset of the features described in *Malfante et al.* (2018) for volcano-seismic signals and are studied as a minimal working set. The attribute description of the three sets of features obtained in temporal and frequency domains is summarized in Table 2.1. At first, these sets of features may appear arbitrary, and some are "low level" and thus may not be ideal for some of the more classic machine learning algorithms. However, since there is not a definite "best" set of features to describe seismic waveforms, and there is no way to determine the "best" set for a given volcano under the assumption of imperfect labelling, we recommended choosing any desired set of features as long as they have already been successfully proven in clean datasets and are simple and fast to extract without any volcano specific preparation. Additionally, after preprocessing, we calculated two (dis)similarity matrices based on waveform cross-correlation (ccorr) and dynamic time warping (DTW) distances (*Anzieta et al.*, 2019; *Ida et al.*, 2022) for each event against the rest. It must be stressed that since distances are calculated for each event against the others into a (dis)similarity matrix, this leads to $n(n-1)/2$ calculations for each matrix that may need to be parallelized to be feasible in an admissible time (i.e., a couple of hours on a modern desktop PC) for datasets with a few thousand events, depending on the processing power of the computing system.

Following our condition of using easily applicable and low computational cost classifiers, we tested the methodology using supervised and unsupervised models present in the standard scikit-learn Python library (*Pedregosa et al.*, 2011) summarized along with their explored respective hyperparameters in Table 2.2, in combination with the Python seismological framework ObsPy (*Beyreuther et al.*, 2010).

For both datasets, we applied the algorithms to the three simple sets of features described earlier. In addition, since the features from *Pérez et al.* (2020) were already precomputed, we applied them to the Cotopaxi catalogue. We did this to test the robustness of the methodology, but in general, we discourage the use of feature sets with so many parameters especially because without data reduction they may lead to over-fitting (*Carniel and Guzmán*, 2020) and because they make calculations slower. We tuned only the most commonly explored hyperparameters shown in Table 2.2, and they were set to values that increased the ac-

Table 2.1: Description of the sets of features used to train the supervised classifiers. Set 1 features are derived following *Watson* (2020), set 2 feature extraction follows *Titos et al.* (2018), and set 3 features are obtained as proposed in Appendix B.1. The "Pérez" group refers to the pre-calculated features provided by *Pérez et al.* (2020). The distance metrics based on waveform cross-correlation and dynamic time warping (*Anzieta et al.*, 2019; *Ida et al.*, 2022) for unsupervised classifiers are also listed.

Group	**Description**	**Domain**
set 1, feature 1	standard deviation	time
set 1, feature 2	skewness	time
set 1, feature 3	kurtosis	time
set 1, feature 4	peak frequency	frequency
set 1, feature 5	quality factor	frequency
set 1, feature 6	skewness	frequency
set 1, feature 7	50th percentile of the cumulative sum of the signal amplitude in frequency domain	frequency
set 2, features 1-5	5th-order Linear Predictive Coding coefficients of the first segment of the waveform out of three	time
set 2, features 6-10	5th-order Linear Predictive Coding coefficients of the second segment of the waveform out of three	time
set 2, features 11-15	5th-order Linear Predictive Coding coefficients of the third segment of the waveform out of three	time
set 2, features 16-18	The 20th, 50th, and 80th percentiles of the cumulative sum of the signal amplitude in time domain	time
set 2, features 19-21	The 20th, 50th, and 80th percentiles of the cumulative sum of the signal amplitude in frequency domain	frequency
set3, feature 1	Duration	time
set3, features 2-8	7th-order Legendre Coefficients from transformed waveform	time
set3, features 9-16	Mean values of amplitude spectrum at seven intervals of same size from 0 to 12.5 Hz	time
Pérez	84 features provided by *Pérez et al.* (2020)	multi
Unsupervised 1	Waveform cross-correlation (ccorr)	time
Unsupervised 2	Dynamic time warping–based distance (DTW)	time

curacy of the models using a 10-fold cross-validation scheme, avoiding fine-tuning them to prevent over-fitting due to lack of knowledge of the number of mislabelled instances. While we explored several numbers of neighbours for the k-nearest neighbours (kNN) classifier, we defaulted to 1-NN as it generally yielded better consistency scores than other number of neighbours (albeit scores were lower than those of the other supervised models). The 1-NN classifier is also a local distance-dependent method sensitive to noise (*García et al.*, 2015) that can help to identify highly confident instances assuming mislabelled observations occur

Table 2.2: Classification models and their adjusted parameters and abbreviations used in this work.

	Adjusted hyperparameters
Supervised models	
K-nearest neighbours (KNN)	Number of neighbours
Logistic regression (LogisReg)	Inverse of regularization strength
Linear support vector machines (SVM-l)	Inverse of regularization strength
Polynomial support vector machines (SVM-p)	Inverse of regularization strength, polynomial order
Radial basis function support vector machines (SVM-rbf)	Inverse of regularization strength, gamma (sample influence radius)
Neural network (shallow) (sNN)	Alpha (strength of regularization), learning rate, hidden layers' sizes
Random forest (RF)	Number of trees in the forest, maximum features, minimum samples for split
Unsupervised models	
k-means (kMeans)	Number of clusters (fxed to 2)
Hierarchical clustering (HierClust)	Number of clusters (fixed to 2), linkage criterion
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	Maximum distance for neighbourhood (eps)
Spectral clustering (SpecClust)	Number of clusters (fixed to 2)

mostly in the "subjective decision boundaries" (where it is difficult to separate VTs and LPs due to the continuum of possible waveforms).

In the case of unsupervised algorithms, we discarded models that led to very low accuracies or delivered no predictions of one of the classes (generally VTs since they constitute the class with the fewest events in both datasets). It should be noted that depending on the comparison method, a similarity or dissimilarity (distances) matrix should be constructed as required by the unsupervised algorithms. Here, we did not use spectrogram-based features because they usually require tuning more parameters than total PSD or signal amplitude spectrum, e.g., spectrogram time window length and percentage of overlapping. Besides these basic preprocessing steps for spectrograms, one methodology requires cross-correlating the smoothed spectrograms (SPCC) of all of the events or designated events in a dataset and then voting between the labels of the top five correlated events (*Curilem et al.*, 2018). Another example requires calculating dissimilarity measurements of the events with an "Atlas" (a kind of spectrogram-based image stacking) of the families or classes of events (*Pérez et al.*, 2022). In our scheme, both methods could become computationally expensive for large datasets: in the first case because each event's smoothed spectrogram in the catalogue is contrasted with the rest to look for the top five cross-correlations (requiring $n(n-1)/2$ comparisons plus spectrogram smoothing), and in the second case because, for each partition, a new "Atlas" would be calculated without the event of interest. Additionally, mislabelled events may distort the spectrogram templates and stacks of spectrograms of events with noisy labels will potentially have blurry boundaries that make decisions less reliable.

After all the models are trained/defined, predictions are obtained for each event along with their consistencies. Then, general metrics about the catalogue and its classes can be obtained and events with low consistencies can be reviewed and relabelled or discarded from the dataset altogether. The general procedure is summarized in Figure 2.2.

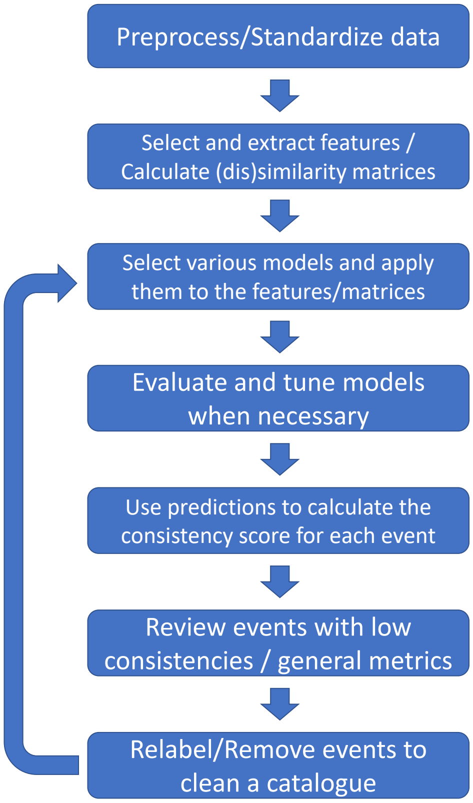
Preprocess/Standardize data
Select and extract features / Calculate (dis)similarity matrices
Select various models and apply them to the features/matrices
Evaluate and tune models when necessary
Use predictions to calculate the consistency score for each event
Review events with low consistencies / general metrics
Relabel/Remove events to clean a catalogue

Figure 2.2: Summary of the general procedure for the catalogue cleaning task.

2.4 Results

Based on the unique consistency scores of all events, we present several metrics tailored and examined by event, class, and globally, and compare them for both catalogues. We do not investigate metrics per classifier since the method's objective is focused on cleaning noisy labels, classifier performance is uncertain in the presence of noise, and they can be studied and reflected by weighting Equation 2.1 if needed.

First, we show a visual example of subsets of events from both catalogues displaying (dis)agreements between the original label and predictions from the different models (Figure 2.3). Consistency is asserted for each event on a dataset following Equation 2.1. In this figure, the first column of each diagram depicts an event's original label coded in blue for LPs and yellow for VTs, and successive columns show agreement with each model in green and disagreement in red. The events shown are ordered as provided in the datasets (unordered in the Cotopaxi dataset and ordered in the Llaima dataset), and the subsets were picked with the intention of depicting different degrees of consistency shown at the right of each subplot.

Events with low consistency have fewer agreements (i.e., fewer green squares after the first column from the left) or equivalently some disagreements (some red squares after the first column from the left). Note that the Cotopaxi catalogue has more models than the Llaima catalogue because we used the pre-calculated features exactly as provided by *Pérez et al.* (2020). We calculate per class consistency statistics for each catalogue by grouping according to the original labels (Table 2.3).

Table 2.3: Consistency statistics of classes in the Cotopaxi and Llaima catalogues.

Class consistencies	**Cotopaxi LPs**	**Cotopaxi VTs**	**Llaima LPs**	**Llaima VTs**
Average	0.9321	0.4832	0.9547	0.8595
Standard dev.	0.0754	0.2585	0.0756	0.2055
Maximum	1	0.8857	1	0.9643
Minimum	0.2857	0.0571	0.1428	0

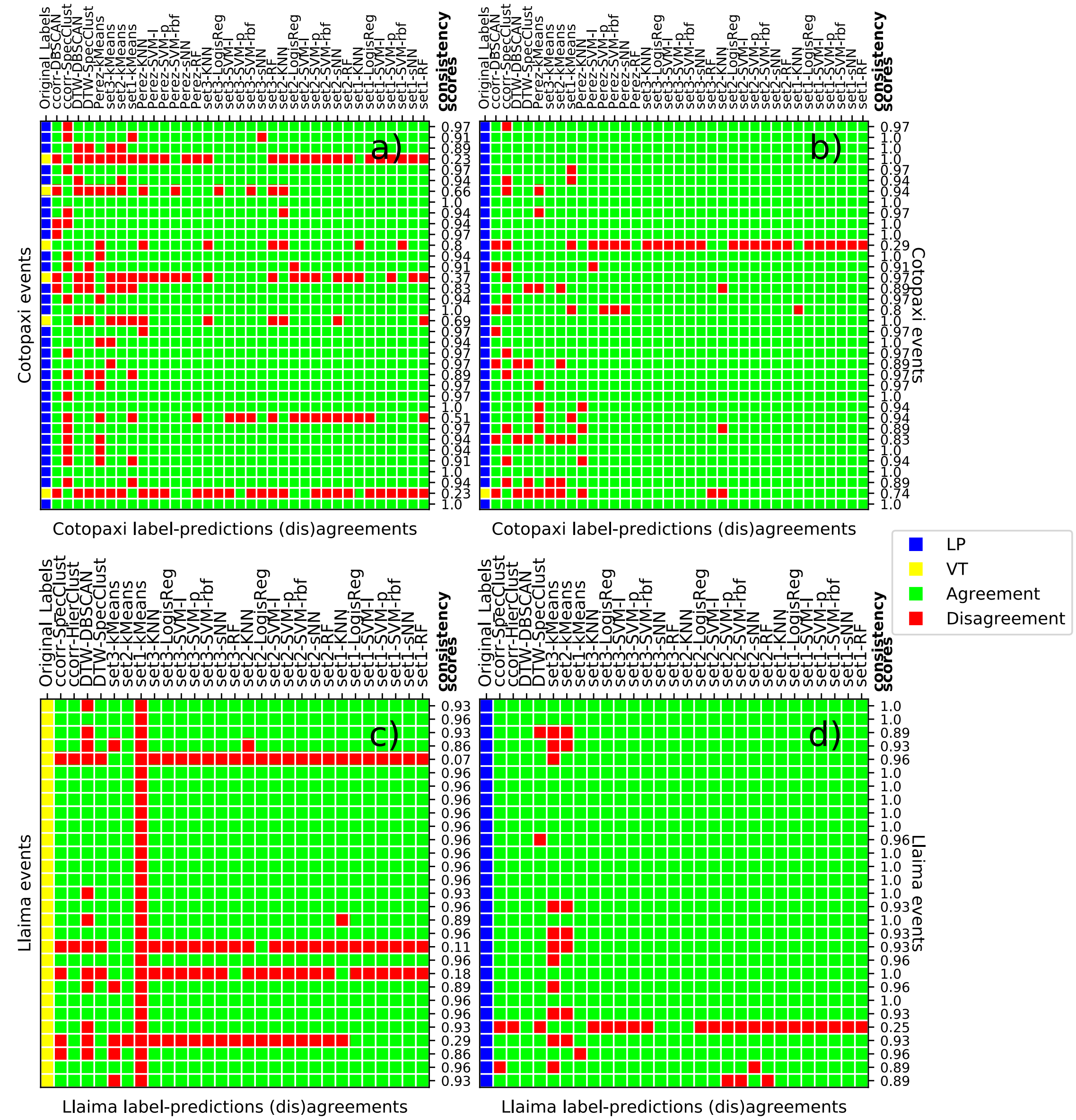

Figure 2.3: Examples of the varying consistency of seismic events from Cotopaxi volcano (a, b) and Llaima volcano (c, d). The labels on top of each plot refer to the feature set/distance metric used as described in Table 2.1 followed by the model applied named after abbreviations in Table 2.2. The original labels given in the first column of each plot and (dis)agreements are coded as shown in the legend, and rounded consistencies of each event are attached to the right of each plot.

These statistics show that different classes can have different overall consistency behaviors within a catalogue and between catalogues. Figure 2.4 shows the **Frequency Distribution** (histogram) **of Consistencies (*FDC*)** per class for Cotopaxi volcano (Figure 2.4a, b) and Llaima volcano (Figure 2.4c, d).

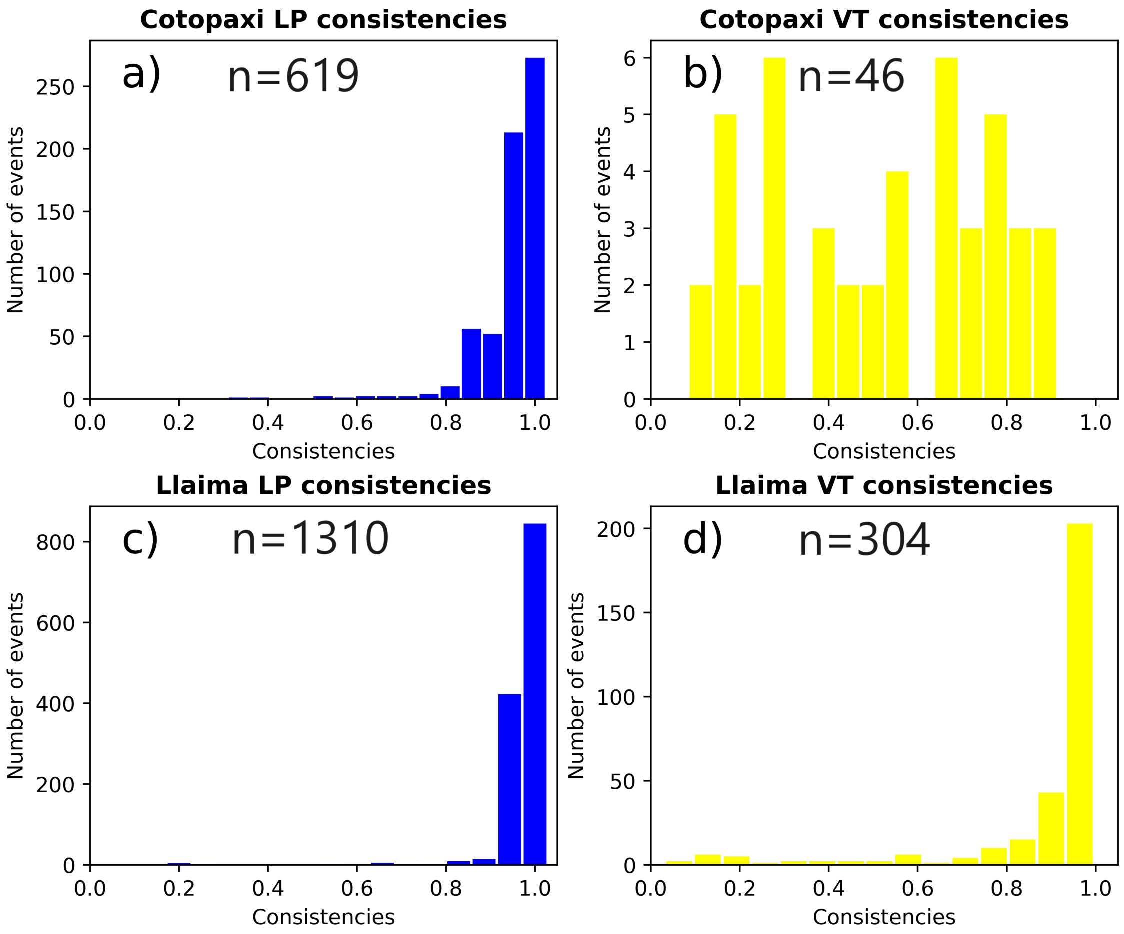

Figure 2.4: Frequency distributions of consistencies of LP (a, c) and VT (b, d) events in the Cotopaxi (a, b) and Llaima (c, d) catalogues, where n is the number of events in each group.

The consistency comprising all events without distinction for both catalogues is calculated, and the same general statistics are presented in Table 2.4.

Table 2.4: Consistency statistics of all events in the Cotopaxi and Llaima catalogues.

Global consistencies	**Cotopaxi**	**Llaima**
Average	0.9011	0.9368
Standard dev.	0.1510	0.1182
Maximum	1	1
Minimum	0.0571	0

Lastly, we show the FDC of both catalogues as a whole with no distinction between classes in Figure 2.5.

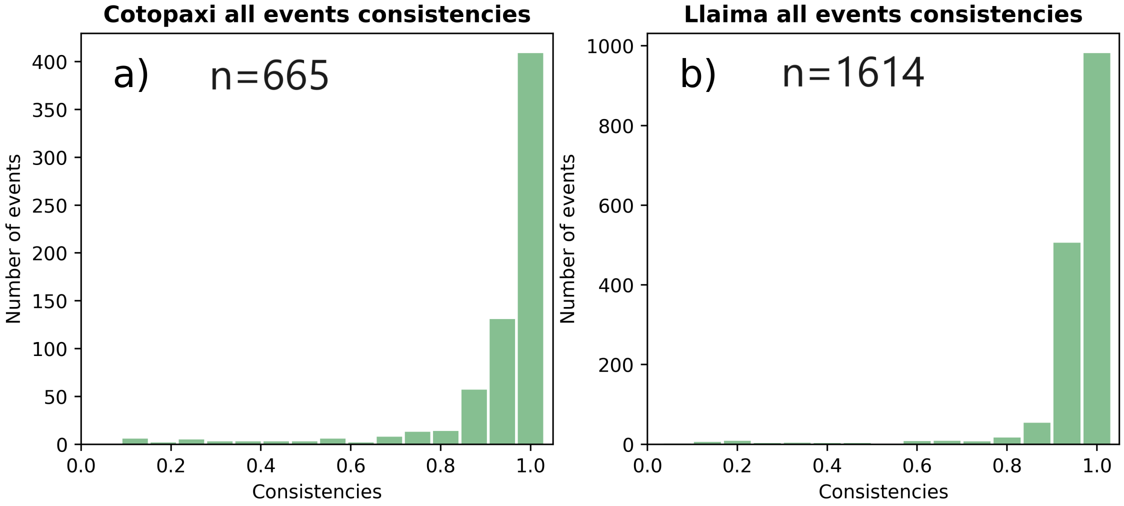

Figure 2.5: Global frequency distributions of consistencies for the Cotopaxi (a) and Llaima (b) catalogues, where n is the number of events in each group.

2.5 Discussions

Figure 2.3 shows that unsupervised classifiers (columns two-to-nine for the Cotopaxi events and two-to-eight for the Llaima events) present more disagreements-red squares-than supervised ones compared to original labels, but they still provide information for events that appear to belong together when there is agreement in events with high consistency values. This leads to the existence of "perfectly coherent" events, but these belong to the LP class. The existence of these events is most likely due to the fact that for both datasets, LPs overwhelmingly represent the majority of events (93.08% of events in the Cotopaxi catalogue and 81.16% of events in the Llaima catalogue) so classifiers will tend to favour LP predictions. Same models can sometimes agree in their predictions even when using different features (for example the kNN predictions for VTs in Figure 2.3a) but they do not need to always agree (for example classifiers from set1 vs the others in the VT with consistency 0.29 in Figure 2.3c). This suggests that many features/models are useful to explore ambiguities that arise when information learned from the dataset does not fully agree with an event's original label. Furthermore, with respect to data imbalance, LPs are on average highly consistent for both datasets, and the variability of the consistency is similar (Table 2.3).

By contrast, for VTs in both datasets, average consistencies are smaller, and variabilities are larger than for LPs. Furthermore, there are notable differences in consistency statistics for VTs. The average consistency for VTs in the Llaima catalogue is much higher than that in the Cotopaxi catalogue despite the variability being similar, and in general, the FDC for VTs from Llaima follows an expected left-tailed distribution (Figure 2.4d). The clean catalogues naturally ought to display consistencies distributed similar to the ones shown for the Llaima catalogue for all classes, since regardless of class, most events should confidently belong to their supposed class. Conversely, only a few should be outliers belonging to the same class or mislabels that belong to another category.

It is also important to stress that if one does not examine the classes in a separated manner, global statistics are not sufficiently different to easily state the comparative quality of the catalogues (Table 2.4, Figure 2.5). Figure 2.5 shows that if the data is too imbalanced (which is extreme for the Cotopaxi dataset), the frequency distribution of consistencies of the larger family of events will "absorb" the distribution of the underrepresented family if they are not separated. Using this comprehensive information from the calculated consistencies, we believe there is strong evidence to confirm the hypothesis that the Cotopaxi catalogue is, in comparison, at least noisier than the Llaima dataset (i.e., seen as a group, events in Cotopaxi's VT family tend to be confused and not clearly distinguishable, as opposed to Llaima's events –either LPs or VTs).

While the consistency distribution for VTs in the Cotopaxi catalogue appears less stable compared to the other more natural distributions (Fig. 2.4b versus Fig. 2.4a, c, d), it is not clear if the distribution's odd shape is due to mislabels in the VT class itself, or if VTs are poorly consistent because the models are biased towards LP predictions due to misclassifications within the LP class. To explicitly assert if the noise in the Cotopaxi catalogue comes from a poor evaluation of LP or VT events, or both, we performed a quick blind test of a few LPs and VTs designated highly and vaguely consistent based solely on the raw waveforms of the selected events. Three experienced volcano seismology analysts from IGEPN (Instituto Geofísico de la Escuela Politécnica Nacional, Ecuador) were asked to independently use any criteria/tools to indicate the class membership of each "anonymized" waveform as being LP or VT. We henceforth call events with high consistencies in their family "high quality" events or **HQ** for short, and use "low quality" or **LQ** for events with low consistency within their family. A total of 28 events from Cotopaxi were selected that included seven LQ LPs (consistencies < 0.7), eight LQ VTs (consistencies < 0.55), six HQ LPs (consistencies > 0.9), and seven HQ VTs (consistencies > 0.7). Of the seven LQ LPs, all were labelled as VTs by at least one analyst, and three were identified as VTs by at least two of three analysts. Of the eight LQ VTs, three were not relabelled by any analyst, four were relabelled by just one analyst, and just one was relabelled by two analysts. As for the HQ LPs, only one was not relabelled by any analyst. Finally, for the HQ VTs, no analyst relabelled any of the eight VTs. From this experiment, we confirm that most ambiguity in the Cotopaxi dataset comes from the LPs, since LQ and especially HQ VTs are consistent with their original labelling for the analysts, while for LPs, even HQ events present a level of ambiguity. Details of the experiment are provided in Appendix B.3.

Although both "low quality" LPs and VTs could be mislabeled, low consistency scores can help reduce the revision process to subsets of the full dataset (by looking at events in the left tails of the FDC). After visual (Figure 2.6) revision, relabelling or even removal from the dataset could help increase the quality of the catalogue. Additionally, while we cannot assert the cleanness of the Llaima dataset since criteria for labeling seismic volcanic events vary from observatory to observatory and volcano to volcano, and we are not familiar with the operational details at OVDAS (Observatorio Volcanológico de los Andes del Sur, Chile), we nevertheless identified a few low consistency LPs and VTs which may require revision for relabelling or filtering (Figure 2.7).

To further explore the results of analyzing the consistency scores in a simulated real-world scenario, we tested the influence of relabelling only the "bad" events that at least two out of three analysts switched in the Cotopaxi catalogue. We do this assuming that in a real application of this methodology, even if we showed that there could be "good" events that

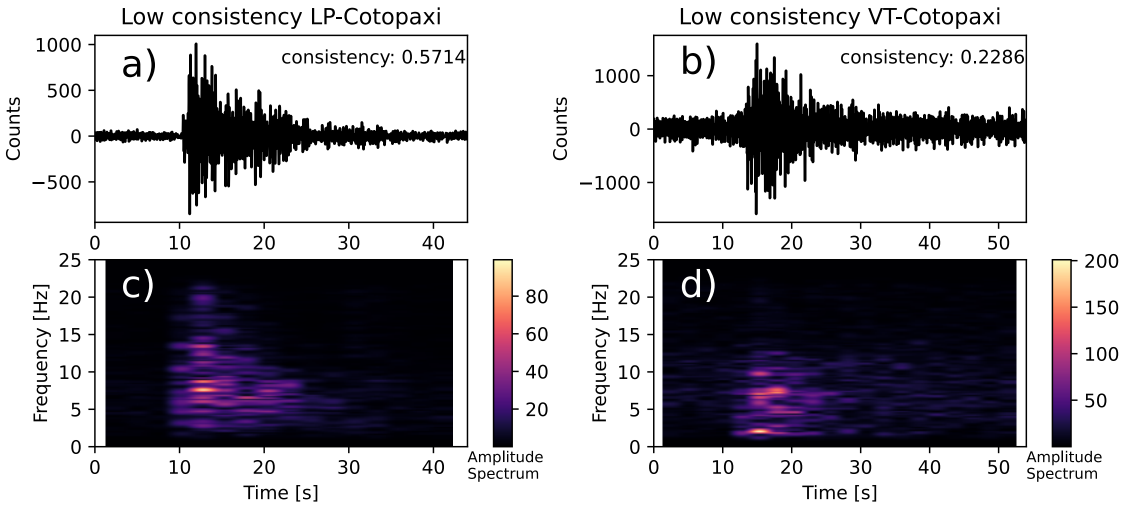

Figure 2.6: A “low quality” (LQ) LP (a) and “low quality” (LQ) VT (b) from the Cotopaxi catalogue with their spectrograms (c, d). The LQ LP may very likely be a mislabelled VT, whereas the LQ VT is more ambiguous.

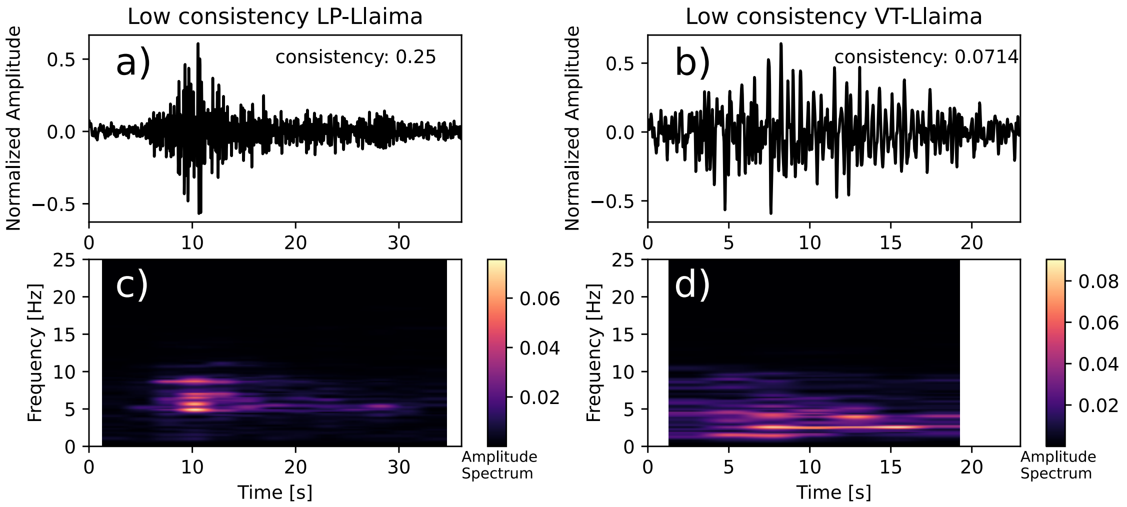

Figure 2.7: A “low quality” (LQ) LP (a) and “low quality” (LQ) VT (b) from the Llaima catalogue with their spectrograms (c, d).

analysts may relabel after revision, only bad events would be reasonably reviewed in large datasets -for this experiment, it amounts to three LPs and one VT. To assess the real effect on the catalogue’s quality overall, we thus re-calculated the statistics of the previous section based only on switching the classes of the "original labels". The resulting statistics are compared to the statistics from Tables 2.3 and 2.4. By doing so, we can measure the actual global changes due to correcting "bad labels" (Table 2.5).

Table 2.5: Consistency statistics for the Cotopaxi catalogue after changing the labels and reapplying the method.

Re-calculated statistics	**Cotopaxi Lps**	**Cotopaxi VTs**	**Cotopaxi all-events**
Average	0.9329	0.5060	0.9021
Standard dev.	0.0738	0.2535	0.1477
Maximum	1	0.8857	1
Minimum	0.371429	0.0571	0.0571

Following this, we show that if the labels had originally been as the analysts redefined, all statistics noticeably improve. Additionally, the effects of these seemingly small changes in the consistency score distributions become apparent (Figure 2.8).

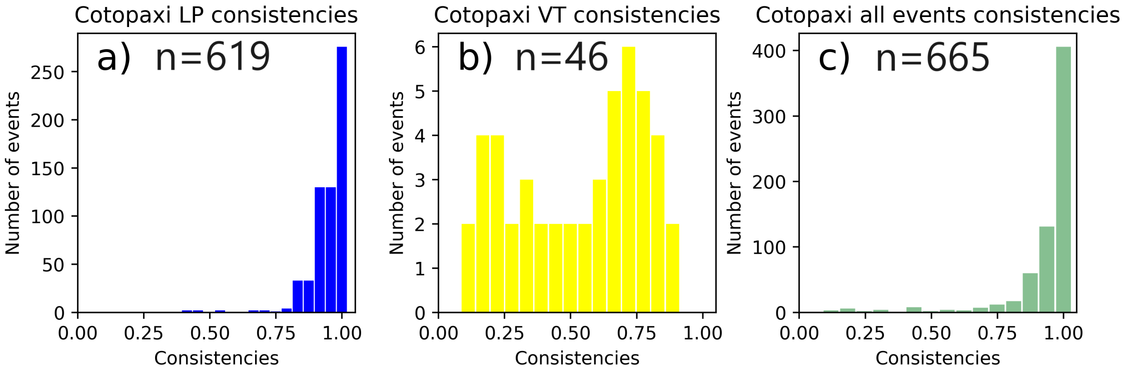

Figure 2.8: Consistency distributions for LPs, VTs, and all events from the Cotopaxi catalogue after four events have been relabeled, where n is the number of events in each group.

From these comparisons (Table 2.5 vs Tables 2.3 and 2.4, and Figure 2.8b vs Figure 2.4b), it is clear that both the VTs and LPs became on average more consistent and less variable. The average consistency increased from 0.9321 to 0.9329, 0.4832 to 0.5060, and 0.9011 to 0.9021 for LPs, VTs, and globally, respectively, while at the same time, all standard deviations decreased. Most notably, the FDC for VTs became closer to an expected left-tailed distribution. This indicates that the process is meaningful for cleaning datasets, even if only a few (here 4 out of 665) events are reviewed and relabelled; this process could be repeated until no further changes are needed.

To further verify what a clean catalogue's consistency should look like, we inverted the usual scenario where only "low quality" events are reviewed. We created an "ideal catalogue" from a dubious one by inspecting "distinctively good" events with notably high consistencies -

in this scenario, the agreement of unsupervised models with original labels becomes more meaningful. The suspicion that the Cotopaxi catalogue has considerable noise in the LP family was previously confirmed by this methodology and corroborated by experts at the IGEPN. Therefore, we created a sub-catalogue inspecting the high quality (events with consistencies > 0.9) LP events along with all the VTs to see how this would change the results in the VT consistencies. From 486 "distinctively good" LPs, we picked 125 that were very visually compelling (i.e., events with emergent onsets and dominant low frequencies) as LP events based on waveform, spectrogram, and spectrum observations. We applied the methodology to the 125 LP events along with the original 46 VTs, initially with their original labels, to assess the effect of using HQ LPs in the calculations before cleaning the catalogue. We then reran the procedure by changing the label of the single VT event that was identified as a LP in the previous test.

By using the original labels employing a subset of LP events that fall within the usual characteristics of LPs, we show that the FDC of VTs follows the expected left-tailed distribution closer than in the full dataset (Figure 2.9b vs Figure 2.4b) as the removed noisy LPs no longer affect the classifiers' performances. More importantly, by changing the label of the single VT that the analysts identified earlier as a potential LP, the consistency distribution of the VTs became even closer to the expected left-tailed distribution of a clean dataset (Figure 2.9d). Interestingly, the event that was changed by analysts in the original dataset has a higher consistency score within the "cleaner" dataset (from 0.5429 to 0.8333). While this may seem contradictory, it supports the idea that since high consistencies are more abundant in cleaner datasets (or within robust families), a higher threshold should be considered for the definition of potential noise boundary. However, this does not mean that large amounts of data should be examined since in cleaner catalogues, the great majority of the events' consistencies will be high as seen in the Llaima dataset (see Figure 2.4c, d). Consequently, the frequency distribution of consistencies is a tool that can help diagnose the potential existence and even level of noise in a catalogue. If distributions show close to expected ideal behaviour (left-tailed distributions), the catalogue is more likely to have a low level of noise, while the more noise in a dataset, the more it will depart from the normal

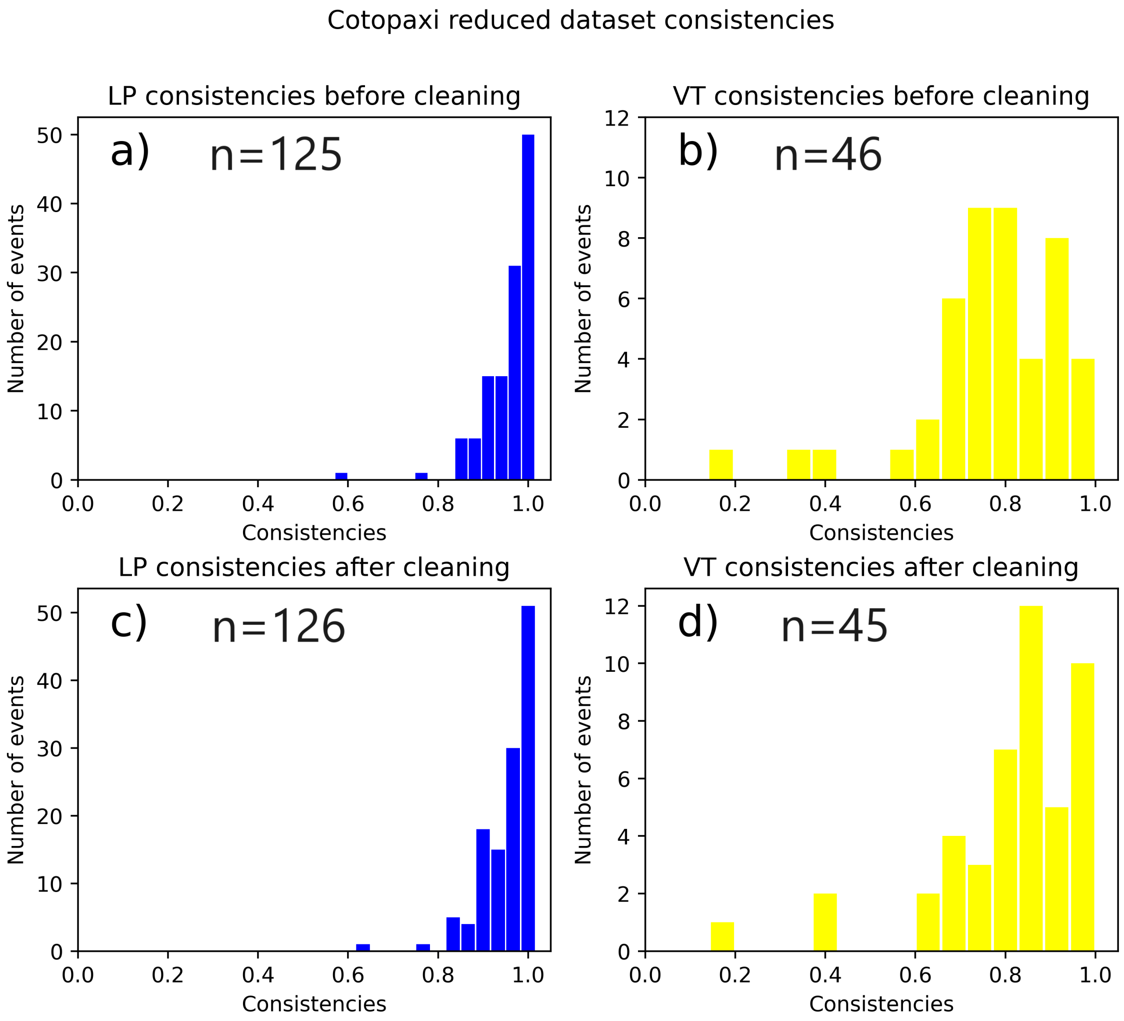

Figure 2.9: Consistency frequency distributions of the reduced Cotopaxi dataset that contains the same VT events but with visually picked "good quality" LP events before (a, b) and after (c, d) correcting the label of one VT event, where n is the number of events in each group.

behaviour. As seen from this experiment, the effects of using more balanced datasets will likely enhance the results from the proposed methodology since classifiers will not favor a family or class based on the number of instances, but on the label noise.

Some important final remarks about implementation of this method are that by using independent classifiers, one can introduce or discard classifiers with varying complexity and thus this method works efficiently for small and large datasets. For very large datasets (several thousand seismic signals or more), the method can be easily adapted to save computing time. For instance, since the classifiers are independent, their training can be done in a parallelized manner not only using multiple cores of a single system simultaneously, but also using many computing devices with different capabilities (e.g., using independent and dedicated systems with powerful CPUs or GPUs to train neural network models, which are currently the bottleneck in our experiments). Other strategies could be to partition the data in any way (randomly or by year/activity period) so that smaller and more tractable subsets can be studied, thus greatly reducing the computational burden of classifiers that scale sharply with the number of samples. Another advantage of the method is that once models have been obtained, and events have been relabelled or discarded, features and (dis)similarity matrices have already been calculated so new iterations due to new labels only involve hyperparameter and model re-calculations (Figure 2.2). Another characteristic of the proposed methodology is that the explored sets of features are sufficiently complementary that they present high agreement in general but allow enough variability so that events which might be slightly odd within a class could still be considered for examination (e.g., the relabelled VT in the "cleaner dataset" scenario still had too low a consistency to be considered of "exceptional quality" and thus required revision). Furthermore, in our experiment, the set described by *Watson* (2020) led to higher consistencies than the other two (the proposed set of features in Appendix B.1 shows slightly lower consistencies, followed by the features of *Titos et al.*, 2018) with the Cotopaxi data. For the Llaima catalogue, the features proposed in this article yielded higher consistencies than the features by *Watson* (2020) and than those by *Titos et al.* (2018) in that order. Furthermore, the correlation matrix visualizations between all the features for the Cotopaxi and the Llaima catalogues (Appendix B.2) show some features that are inevitably slightly correlated in the case of similar feature definitions (e.g., percentiles in FD, of *Titos et al.*, 2018, compared to skewness in FD, used in *Watson*, 2020), but are not strongly correlated as groups between the proposed sets of features. In the case of longer available exploration times (e.g., outside of volcanic crises), more comprehensive sets of features (*Toney et al.*, 2022) and models could be tested without altering the general procedure, adding to the flexibility of the method. We also investigated what would happen in extreme situations when the catalogue cleaning needs to be done rapidly using default values for the supervised classifiers and skipping the hyperparameter tuning. While this leads to slightly less consistent models, events with very low/high consistencies are still sufficiently well identified. A clear disadvantage of the method is that it performs best in one-to-one class comparisons, and thus becomes increasingly cumbersome to apply when many classes are present within a catalogue. The method would need to be applied $k(k-1)/2$ times if k classes are investigated, although when catalogues have specific classes that are sufficiently different to the rest, those comparisons can be bypassed.

2.6 Conclusions

We explored the results of applying a machine learning approach for the quality assessment of two volcano-seismic event catalogues. The proposed methodology enables the identification of potentially noisy labels in catalogues, provides steps to find out where (which class) the source of noise could be in the catalogues, and yields a rough estimation of the significance of this noise. The methodology is easier to apply than those using complex machine learning algorithms since few hyperparameters need to be mildly tuned for the proposed models, and preprocessing steps are general enough that they do not require volcano-specific knowledge. During emergencies, the hyperparameter determination steps can be skipped and default models could be trained with similar results. The use of complementary but uncomplicated features in the creation of different simple models helps achieve the right amount of model diversity in a manner that is straightforward to apply. The advantage of the fast implementation is that one can quickly and iteratively apply the process. The statistics obtained from this process allow for visual as well as quantitative verification of the quality of the catalogues, along with assessment of the effect of cleaning these catalogues. In this study, we showed that the methodology helped to objectively verify the existence of presumed noise in the Cotopaxi catalogue and showed what a good catalogue should look like in the case of the Llaima dataset. Noisy events were unsurprisingly identified in both catalogues, and thus, we believe that similar procedures should be considered in other settings where noisy labels could be present affecting classification tasks. In particular, this same experiment could be repeated to examine the compatibility of events within sub-families derived from unsupervised learning (clustering), or to clean dubious but useful historical catalogues. While the results of this study were derived from catalogues of volcanoes with similar characteristics, the procedure is sufficiently general that it can be applied to other types of volcanoes.

References

Anzieta, J., H. Ortiz, G. Arias, and M. C. Ruiz (2019), Finding Possible Precursors for the 2015 Cotopaxi Volcano Eruption Using Unsupervised Machine Learning Techniques, *International Journal of Geophysics*, *2019*, 1–9, doi:10.1155/2019/6526898.

Bell, A. F., M. Naylor, S. Hernandez, I. G. Main, H. E. Gaunt, P. Mothes, and M. Ruiz (2018), Volcanic eruption forecasts from accelerating rates of drumbeat long-period earthquakes, *Geophysical Research Letters*, *45*(3), 1339–1348, doi:https://doi.org/10.1002/2017GL076429.

Benson, P. M., S. Vinciguerra, P. G. Meredith, and R. P. Young (2008), Laboratory simulation of volcano seismicity, *Science*, *322*(5899), 249–252, doi:10.1126/science.1161927.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010), ObsPy: A Python Toolbox for Seismology, *Seismological Research Letters*, *81*(3), 530–533, doi:10.1785/gssrl.81.3.530.

Boué, A., P. Lesage, G. Cortés, B. Valette, G. Reyes-Dávila, R. Arámbula-Mendoza, and A. Budi-Santoso (2016), Performance of the 'material failure forecast method' in real-time situations: A bayesian approach applied on effusive and explosive eruptions, *Journal of Volcanology and Geothermal Research*, *327*, 622–633, doi:https://doi.org/10.1016/j.jvolgeores.2016.10.002.

Brodley, C. E., and M. A. Friedl (1999), Identifying mislabeled training data, *J. Artif. Intell. Res.*, *11*, 131–167, doi:10.1613/jair.606.

Bueno, A., C. Benítez, S. De Angelis, A. Díaz Moreno, and J. M. Ibáñez (2020), Volcano-seismic transfer learning and uncertainty quantification with bayesian neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, *58*(2), 892–902, doi:10.1109/TGRS.2019.2941494.

Bueno, A., M. Titos, C. Benítez, and J. M. Ibáñez (2022), Continuous active learning for seismo-volcanic monitoring, *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5, doi:10.1109/LGRS.2021.3121611.

Canário, J. P., R. F. de Mello, M. Curilem, F. Huenupan, and R. A. Rios (2020), Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification, *Data in Brief*, *30*(105627), 1–6, doi:https://doi.org/10.1016/j.dib.2020.105627.

Carniel, R., and S. R. Guzmán (2020), Machine learning in volcanology: A review, in *Updates in Volcanology*, edited by K. Németh, chap. 5, pp. 1–26, IntechOpen, Rijeka, doi:10.5772/intechopen.94217.

Chouet, B. A. (1996), Long-period volcano seismicity: its source and use in eruption forecasting, *Nature*, pp. 309–3016, doi:10.1038/380309a0.

Chouet, B. A., and R. S. Matoza (2013), A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption, *Journal of Volcanology and Geothermal Research*, *252*, 108–175, doi:10.1016/j.jvolgeores.2012.11.013.

Clarke, J., L. Adam, and K. van Wijk (2021), Lp or vt signals? how intrinsic attenuation influences volcano seismic signatures constrained by whakaari volcano parameters, *Journal of Volcanology and Geothermal Research*, *418*, 107,337, doi:https://doi.org/10.1016/j.jvolgeores.2021.107337.

Cortés Moreno, G., R. Carniel, P. Lesage, M. A. Mendoza Perez, and I. Della Lucia (2019), Volcano-independent seismic recognition: detecting and classifying events of a given volcano using data from others, *ESS Open Archive*, doi:10.1002/essoar.10500900.1.

Curilem, M., R. F. de Mello, F. Huenupan, C. San Martin, L. Franco, E. Hernández, and R. A. Rios (2018), Discriminating seismic events of the llaima volcano (chile) based on spectrogram cross-correlations, *Journal of Volcanology and Geothermal Research*, *367*, 63–78, doi:https://doi.org/10.1016/j.jvolgeores.2018.10.023.

Deng, L., and D. Yu (2014), Deep learning: Methods and applications, *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387, doi:10.1561/2000000039.

Dufek, J., T. Esposti Ongaro, and O. Roche (2015), Chapter 35 - pyroclastic density currents: Processes and models, in *The Encyclopedia of Volcanoes (Second Edition)*, edited by H. Sigurdsson, second edition ed., pp. 617–629, Academic Press, Amsterdam, doi: https://doi.org/10.1016/B978-0-12-385938-9.00035-3.

Falcin, A., J.-P. Métaxian, J. Mars, Éléonore Stutzmann, J.-C. Komorowski, R. Moretti, M. Malfante, F. Beauducel, J.-M. Saurel, C. Dessert, A. Burtin, G. Ucciani, J.-B. de Chabalier, and A. Lemarchand (2021), A machine-learning approach for automatic classification of volcanic seismicity at la soufrière volcano, guadeloupe, *Journal of Volcanology and Geothermal Research*, *411*, 107,151, doi:https://doi.org/10.1016/j.jvolgeores.2020.107151.

Frenay, B., and M. Verleysen (2014), Classification in the presence of label noise: A survey, *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869, doi:10.1109/TNNLS.2013.2292894.

García, S., J. Luengo, and F. Herrera (2015), *Dealing with Noisy Data*, pp. 107–145, Springer International Publishing, Cham, doi:10.1007/978-3-319-10247-4_5.

Girona, T., C. Caudron, and C. Huber (2019), Origin of shallow volcanic tremor: The dynamics of gas pockets trapped beneath thin permeable media, *Journal of Geophysical Research: Solid Earth*, *124*(5), 4831–4861, doi:https://doi.org/10.1029/2019JB017482.

Ida, Y., E. Fujita, and T. Hirose (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*, 107,616, doi:https://doi.org/10.1016/j.jvolgeores.2022.107616.

Keramati, M., M. A. Tayebi, Z. Zohrevand, U. Glässer, J. Anzieta, and G. Williams-Jones (2023), Cubism: Co-balanced mixup for unsupervised volcano-seismic knowledge transfer, in *Machine Learning and Knowledge Discovery in Databases*, edited by M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak and G. Tsoumakas, pp. 581–597, Springer Nature Switzerland, Cham.

Khan, M. S., M. Curilem, F. Huenupan, M. F. Khan, and N. Becerra Yoma (2019), A Signal Processing Perspective of Monitoring Active Volcanoes [Applications Corner], *IEEE Signal Processing Magazine*, *36*(6), 125–163, doi:10.1109/MSP.2019.2930427.

Kilburn, C. R. (2003), Multiscale fracturing as a key to forecasting volcanic eruptions, *Journal of Volcanology and Geothermal Research*, *125*(3), 271–289, doi:https://doi.org/10.1016/S0377-0273(03)00117-3.

Linville, L., D. Anderson, J. Michalenko, J. Galasso, and T. Draelos (2020), Semisupervised Learning for Seismic Monitoring Applications, *Seismological Research Letters*, *92*(1), 388–395, doi:10.1785/0220200195.

Malfante, M., M. Dalla Mura, J. I. Mars, J.-P. Métaxian, O. Macedo, and A. Inza (2018), Automatic classification of volcano seismic signatures, *Journal of Geophysical Research: Solid Earth*, *123*(12), 10,645–10,658, doi:https://doi.org/10.1029/2018JB015470.

McNutt, S. R. (2005), VOLCANIC SEISMOLOGY, *Annual Review of Earth and Planetary Sciences*, *33*(1), 461–491, doi:10.1146/annurev.earth.33.092203.122459.

McNutt, S. R., and D. C. Roman (2015), Volcanic Seismicity, in *The Encyclopedia of Volcanoes*, pp. 1011–1034, Elsevier, doi:10.1016/B978-0-12-385938-9.00059-6.

McNutt, S. R., G. Thompson, J. Johnson, S. D. Angelis, and D. Fee (2015), Chapter 63 - seismic and infrasonic monitoring, in *The Encyclopedia of Volcanoes (Second Edition)*, edited by H. Sigurdsson, second edition ed., pp. 1071–1099, Academic Press, Amsterdam, doi:https://doi.org/10.1016/B978-0-12-385938-9.00063-8.

Met Office (2010 - 2015), *Cartopy: a cartographic python library with a Matplotlib interface*, Exeter, Devon.

Narváez, L., R. A. Torres, D. M. Gómez, G. P. Cortés, H. Cepeda, and J. Stix (1997), 'tornillo'-type seismic signals at galeras volcano, colombia, 1992–1993, *Journal of Volcanology and Geothermal Research*, *77*(1), 159–171, doi:https://doi.org/10.1016/S0377-0273(96)00092-3, galeras Volcano, Colombia: Interdisciplinary Study of a Decade Volcano.

Pallister, J., P. Papale, J. Eichelberger, C. Newhall, C. Mandeville, S. Nakada, W. Marzocchi, S. Loughlin, G. Jolly, J. Ewert, and J. Selva (2019), Volcano observatory best practices

(VOBP) workshops - a summary of findings and best-practice recommendations, *Journal of Applied Volcanology*, *8*(1), doi:10.1186/s13617-019-0082-8.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, *12*, 2825–2830.

Pérez, N., D. Benítez, F. Grijalva, R. Lara-Cueva, M. Ruiz, and J. Aguilar (2020), Eseismic: Towards an ecuadorian volcano seismic repository, *Journal of Volcanology and Geothermal Research*, *396*, 106,855, doi:https://doi.org/10.1016/j.jvolgeores.2020.106855.

Pérez, N., F. S. Granda, D. Benítez, F. Grijalva, and R. Lara (2022), 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, doi:10.1109/TGRS.2020.3048107.

Reyes, J. A., and C. J. Jiménez Mosquera (2017), Non-supervised classification of volcanic-seismic events for tungurahua-volcano ecuador, in *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–6, doi:10.1109/ETCM.2017.8247446.

Sluban, B., and N. Lavrač (2015), Relating ensemble diversity and performance: A study in class noise detection, *Neurocomputing*, *160*, 120–131, doi:https://doi.org/10.1016/j.neucom.2014.10.086.

Soto, R., F. Huenupan, P. Meza, M. Curilem, and L. Franco (2018), Spectro-temporal features applied to the automatic classification of volcanic seismic events, *Journal of Volcanology and Geothermal Research*, *358*, 194–206, doi:https://doi.org/10.1016/j.jvolgeores.2018.04.025.

Sáez, J. A., M. Galar, J. Luengo, and F. Herrera (2013), Tackling the problem of classification with noisy data using multiple classifier systems: Analysis of the performance and robustness, *Information Sciences*, *247*, 1–20, doi:https://doi.org/10.1016/j.ins.2013.06.002.

Titos, M., A. Bueno, L. García, and C. Benítez (2018), A deep neural networks approach to automatic recognition systems for volcano-seismic events, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(5), 1533–1544, doi:10.1109/JSTARS.2018.2803198.

Toney, L., D. Fee, A. Witsil, and R. S. Matoza (2022), Waveform Features Strongly Control Subcrater Classification Performance for a Large, Labeled Volcano Infrasound Dataset, *The Seismic Record*, *2*(3), 167–175, doi:10.1785/0320220019.

Watson, L. M. (2020), Using unsupervised machine learning to identify changes in eruptive behavior at mount etna, italy, *Journal of Volcanology and Geothermal Research*, *405*, 107,042, doi:https://doi.org/10.1016/j.jvolgeores.2020.107042.

Witsil, A., D. Fee, J. Dickey, R. Peña, R. Waxler, and P. Blom (2022), Detecting large explosions with machine learning models trained on synthetic infrasound data, *Geophysical Research Letters*, *49*(11), e2022GL097,785, doi:https://doi.org/10.1029/2022GL097785, e2022GL097785 2022GL097785.

Chapter 3

Investigating 10 years of volcano acoustic activity at Tungurahua volcano, Ecuador aided by Machine Learning

Anzieta, J., G. Williams-Jones, H.D. Ortiz and M.C. Ruiz (2024), Investigating 10 years of volcano acoustic activity at Tungurahua volcano, Ecuador aided by Machine Learning, *Seismological Research Letters*, In Review.

3.1 Introduction

Seismology in general, and volcano seismology in particular, heavily relies on the recording and analysis of Earth's ground motion to try to understand geophysical processes such as earthquakes and volcanic eruptions. When the recordings of atmospheric vibrations are also included, seismo-acoustics (*Arrowsmith et al.*, 2010) allows for a better understanding and discrimination of subsurface and atmospheric signals related to volcanic activity, or to other phenomena of interest such as earthquakes and chemical or nuclear explosions (*Matoza and Roman*, 2022; *Dannemann et al.*, 2023).

Progress in the sensing, acquisition, and storage of seismo-acoustic data has led to the development of ever more sophisticated tools for volcano monitoring (*Matoza and Roman*, 2022), with the increasing amount and complexity of data requiring the creation and application of efficient and nearly automatic procedures.

The nature of such tools is varied, and among them, traditional multi-station methods stand out because the breadth of applicable techniques surpasses the capabilities of single-station approaches for the quantitative description of different phenomena sources (*Ortiz et al.*, 2018; *Matoza and Roman*, 2022; *Tan et al.*, 2023) (e.g., by determination of parameters related to source location, geometry, time-evolution, etc.). On the other hand, single-station routines still offer valuable information ranging from statistical approaches analyzing the time evolution of single event occurrences (*Varley et al.*, 2006), to complex tracking of seismic velocity changes related to subsurface changes in volcanoes (*De Plaen et al.*, 2016), or phenomena in their vicinity (*Ortiz et al.*, 2021a); these are often the sole option when multiple sensors are not available in a specific area.

Complementary to the more traditional tools, several machine learning methods have been adopted relatively early in volcano seismology (*Matoza and Roman*, 2022; *Carniel and Guzmán*, 2020), and led to a plethora of studies applying supervised (*Masotti et al.*, 2006; *Malfante et al.*, 2018a; *Titos et al.*, 2018, 2019; *Martínez et al.*, 2021; *Manley et al.*, 2022; *Toney et al.*, 2022; *Trani et al.*, 2022), unsupervised (*Esposito et al.*, 2008; *Reyes and Jiménez Mosquera*, 2017; *Ortiz et al.*, 2021b; *Watson*, 2020; *Witsil and Johnson*, 2020; *Duque et al.*, 2020), transfer learning techniques (*Bueno et al.*, 2020; *Titos et al.*, 2020; *Keramati et al.*, 2023), or their combinations (*Ren et al.*, 2020).

Here we explored the results of applying a combination of several traditional tools and contemporary machine learning techniques currently employed by the community to increase the information from a catalogue of explosions obtained manually by only using data from a single acoustic station (simulating situations when only one station is available or stable for a long period of time). We applied a succession of tools to examine a vast dataset in order to save time and human resources while ensuring good data quality. This study aims to showcase how much information can be extracted from a single station, to compare some modern tools with traditional ones, and to enhance an existing database, which is desirable even in hindsight for a deeper understanding of eruptive processes (*Yukutake et al.*, 2023).

3.2 Background and Data

3.2.1 Background

Tungurahua is an andesitic stratovolcano located in the Andean region of Ecuador (Figure 3.1a) with numerous historical episodes of major eruptive activity (*Mothes et al.*, 2015), the most recent of which occurred from October 1999 to March 2016. This recent activity was characterized by periods with several eruptive phases and styles interspersed with periods of relative quiescence. The active phases generated a broad range of phenomena that resulted from often continuous and sometimes intense degassing (*Arellano et al.*, 2008; *Hidalgo et al.*, 2015) processes. Ash emissions (*Eychenne et al.*, 2012; *Wright et al.*, 2012) of different sizes were often correlated with explosions of varying intensities from strombolian to sub-plinian (*Fee et al.*, 2010; *Kim et al.*, 2014; *Hall et al.*, 2015; *Anderson et al.*, 2018; *Battaglia et al.*, 2019). These explosions often included pyroclastic fountaining (*Hidalgo et al.*, 2015), but most notably pyroclastic density currents of great intensity and destructive power (*Douillet et al.*, 2013; *Hall et al.*, 2015).

The monitoring of Tungurahua volcano has been responsibility of Ecuador's Instituto Geofisico - Escuela Politecnica Nacional (IG-EPN), since 1983 with monitoring capabilities steadily increasing since the institutions' inception due to various international collaborations (*Hidalgo et al.*, 2023). The first real time transmission, permanent, collocated broadband seismic and infrasonic sensors were installed on the flanks of Tungurahua (Figure 3.1) gradually from July 2006 to 2008 in collaboration with the Japan International Cooperation Agency (JICA) (*Kumagai et al.*, 2007). This network of 5 permanent stations (Figure 3.1b) provided near continuous seismic and acoustic recordings during the last 10 years of activity (Figure 3.2a), and continues to operate presently.

3.2.2 Data

From these recordings, an unpublished explosion catalogue was created manually by IG-EPN personnel as eruptions unfolded. Discrete acoustic events that exceeded a threshold (∼40 Pascals [Pa] peak-to-peak pressure at Station BMAS) were hand-picked and generated a catalogue of 6523 relatively discrete events spanning from July 14, 2006, to August 25,

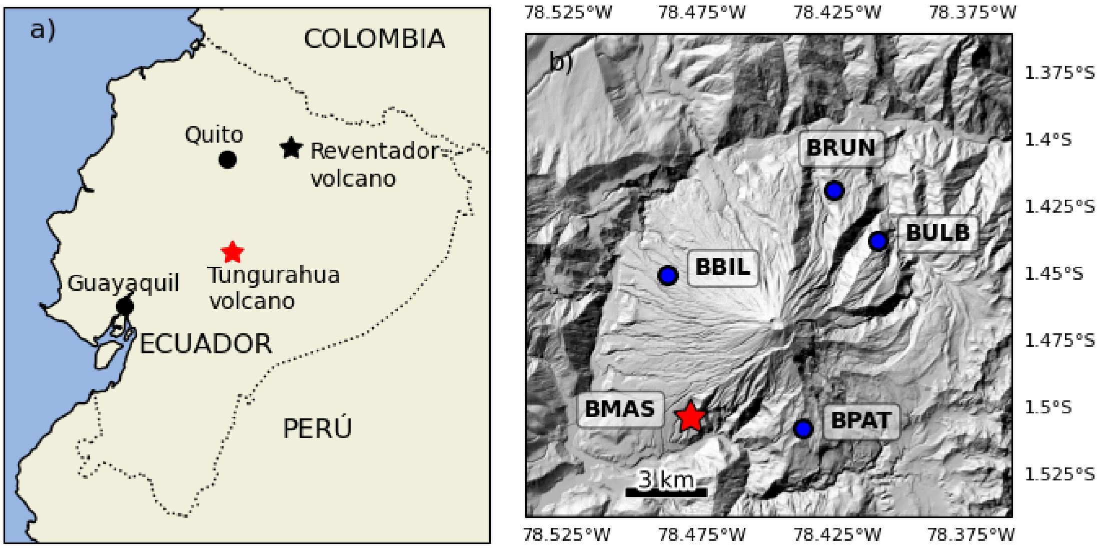

Figure 3.1: Location of Tungurahua volcano (red star) and Reventador volcano (black star), as well as of the two major cities in Ecuador (black dots). b) Seismo-acoustic network from the IG-EPN-JICA collaboration. Blue circles are unused stations, the red star indicates the location of BMAS used for this study.

2015 (Figure 3.2b). Given the large volume of data that needed to be manually processed, acoustic signals were scrutinized from daily drumplots with a relatively low time resolution of approximately one minute. Consequently, when examined in detail at smaller scales, some of the degassing signals displayed various shapes that deviated from the typical N-shaped explosions, bearing some resemblance to short "jetting" (rightmost column in Figure 3.3) events that were also present during explosive periods at Tungurahua (*Ruiz et al.*, 2006; *Matoza et al.*, 2009; *Fee et al.*, 2010; *Anderson et al.*, 2018).

Although some of the events lie in the continuum between discrete explosions and tremor-like behaviour, we used all the timestamps in this manual catalogue and extracted 20 seconds waveforms from the recordings at station BMAS (red star in Figure 3.1b) as input for this study. Here, like in the original manual catalogue, the focus is on signals with transient explosion-like resemblance and thus, we only tried window sizes of 10 and 20 seconds, sufficient to encompass the primary content of these signals. While chugging or jetting signals can persist for significantly longer periods (up to several hours), they were not central to the analysis and a small subset was included only as they were present in the original

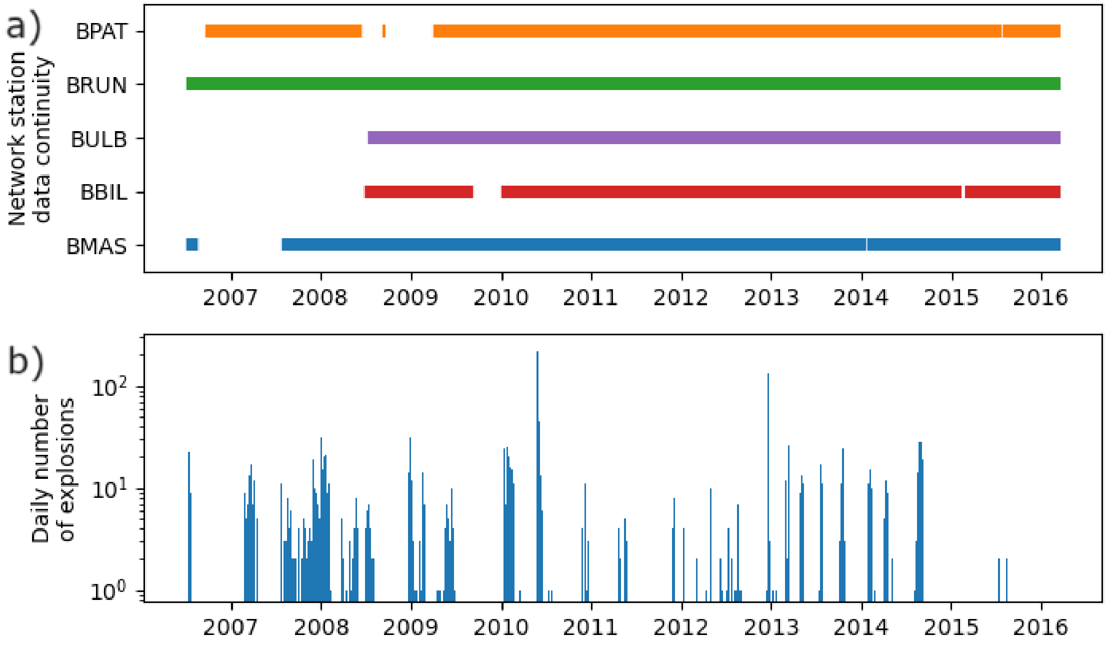

Figure 3.2: a) Data time availability per station in the Tungurahua IG-EPN-JICA network. b) Daily number of explosions from a manually constructed catalogue

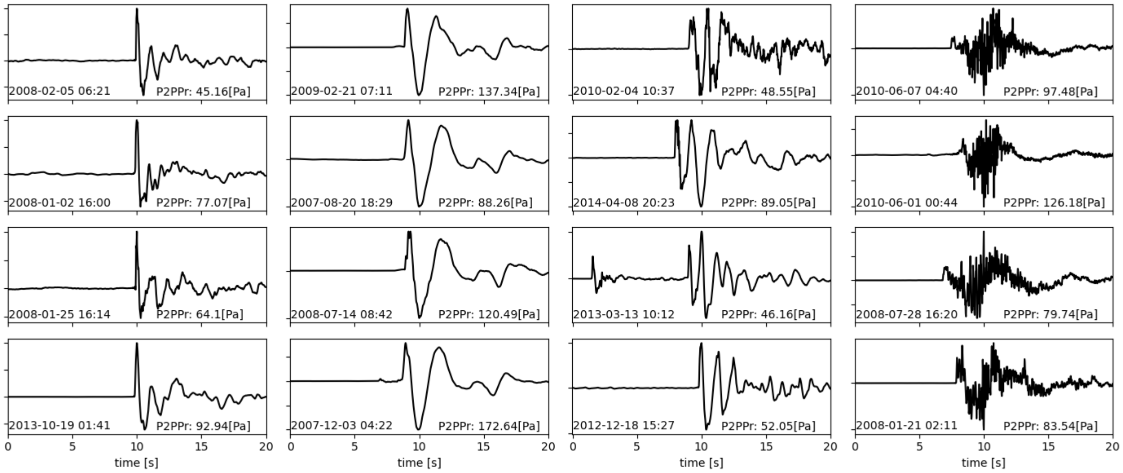

Figure 3.3: 20 seconds samples of some repeating acoustic events indexed in the manual catalogue as recorded in BMAS (unfiltered). P2PPr stands for peak-to-peak pressure.

catalogue. The 20 second windows allowed for better classification and cleaning results to be described later compared to the 10 second window. This approach enables consistent assessment of potential repeating patterns (Figure 3.3), while ensuring that variations in waveforms are not influenced by differences in station locations.

Station BMAS was selected as it was used as the reference for the manual catalogue, was one of the first stations to be installed, and had more data than all other stations except BRUN. Importantly, the signal-to-noise ratio of the BMAS station was second only to BBIL because Tungurahua's geometry favours acoustic propagation towards the north-west (*Kim et al.*, 2012; *Anderson et al.*, 2018; *Ortiz et al.*, 2018). Since station BMAS was inactive during some time periods (Figure 3.2a), the explosion events manually detected at the station were 5740. Additionally, we manually extracted 5715 non-explosion waveforms from the same station for training purposes described below.

3.3 Methods and Procedures

The general analysis scheme for the ten years of acoustic recordings based exclusively on station BMAS was as follows: first we attempted to contrast and optimize two automatic detection algorithms. Because we ignored other stations, we applied two successive procedures to discard potential false detections. The first consisted of testing a combination of several supervised classifiers and features using the manual explosions and handpicked noise, and then applying the best classifier to keep the detections that were classified as explosions. The second consisted of manually reviewing a subset of the explosions obtained in the previous step, and using the false positives to further discard wrong detections. This was done using the classifiers that were not previously chosen and retaining events labelled as explosions by most of them. These two steps were performed to reduce the number of events that must be reviewed manually to create a robust catalog from a single station, and can be ignored for example, by using a multi-station coincidence approach. Finally, these distilled events were used to apply classical clustering with two measures of (dis)similarity, and deep clustering using a Convolutional Autoencoder. All analyses involving waveforms were performed on a personal computer using Python, its popular libraries (e.g., Numpy, Scipy, Matplotlib, etc.), the seismological package Obspy (*Beyreuther et al.*, 2010), and several other software packages (referenced below when appropriate). As a first step before cutting the waveforms, daily acoustic traces were high-pass filtered above 0.05 Hz to preserve most of the information while eliminating intense low frequency drifts. This leads to the potential inclusion of microbaroms (low-frequency atmospheric pressure waves created by the ocean that propagate over long distances) or other low frequency phenomena into the study, but they will be taken out during the several sequential dataset cleaning procedures. The cut signals had fixed durations of 20 seconds and were normalized against maximum absolute amplitude so that they could be compared solely based on their shape and not their magnitudes (e.g., *Canário et al.*, 2020; *Anzieta et al.*, 2023). Since all the traces were extracted from the same station (BMAS), no further preprocessing was performed.

3.3.1 Part 1: Dataset increment

The first part of this study consisted of attempting to reduce the gap between manually detected explosions and missing events due to the ∼40 Pa peak-to-peak manual detection threshold.

While machine learning has achieved state-of-the-art performance in earthquake detection and phase picking nearly equalling (*Zhu and Beroza*, 2018; *Mousavi et al.*, 2020; *Mai et al.*, 2023) and sometimes even surpassing human capabilities (*Wickham-Piotrowski et al.*, 2023; *Okamoto et al.*, 2024), in volcano seismology machine learning is mostly devoted to event classification because event detection is still routinely applied semi-automatically using classical techniques (*Malfante et al.*, 2018b; *Mousavi and Beroza*, 2023). Many current machine learning applications still rely upon the classic STA/LTA algorithm for detection and then apply more sophisticated tools for clustering or classification (e.g., *Jenkins II et al.*, 2021; *Machacca et al.*, 2023; *Tan et al.*, 2023), although it is possible to deal with both tasks simultaneously (e.g., *Titos et al.*, 2019; *Cortés et al.*, 2021).

Here, we also aim to improve the original catalogue by dividing the problem in two parts: first by detecting signals with 'classical' tools, and then by improving the detections using machine learning classification between true signals and noise. We then apply a process to clean the dataset further.

Detection

For detection, we tested a version of the classical STA/LTA algorithm (*Allen*, 1978) against the "VINEDA" algorithm designed specifically for explosion detection (*Bueno et al.*, 2019).

Both, the STA/LTA and the VINEDA algorithms require adjusting or tuning many parameters that depend on the data and influence their detection performance. Thus, we explored a range of suitable parameters for both by examining data previously inspected visually that consisted of two days with low explosive activity (20 and 12 explosions) and one of intense activity (more than 800 explosions). For the recursive STA/LTA (used for explosion-like signals; e.g., *Matoza et al.*, 2014, 2019), we tried short-time windows of 0.5, 1, 2, and 3 seconds, long-time windows of 10 to 60 s in steps of 10 seconds, onset thresholds of 3, 5, 9, 12, 17 and 20, and off thresholds of 1, 3 and 5. For the VINEDA detector, we fixed the high frequency to 8 Hz, low frequency to 0.3 Hz and the average minimum and maximum event durations to 2 and 20 seconds, respectively. We then explored combinations of the number of frequency bands and duration bands to 3, 5, 7 separately, the beta value to 0.13, 0.2, 0.3, 0.5 and 1, and the detection threshold to values of 25, 50, 75 and 100. After comparing results from both detectors' combinations of parameters, we picked from the ones we deemed adequate (as discussed in a later section): a short-time window of 1 second, long time window of 40 seconds and trigger-on-off of 9, 3 for the recursive STA/TLA detector, and for VINEDA –besides the mentioned fixed parameters– we used both the number of frequency bands and number of duration bands of 3, and beta equal to 0.2.

Discrimination

To discriminate actual acoustic events from noisy signals with explosion-like shapes that slip through the detection stage (e.g., instrumental glitches), we trained a classifier using the 5740 events from the manual catalogue and the 5715 non-explosion waveforms described in the previous section. For ease and speed of implementation, we trained and validated several simple learning models ("shallow learners", *Deng and Yu*, 2014) implemented in the *scikit-learn* Python package (*Pedregosa et al.*, 2011) using three independent groups of features with demonstrated success in volcano seismo-acoustic datasets (*Titos et al.*, 2018; *Watson*, 2020; *Anzieta et al.*, 2023), and picked the one that performed the best. Among the different "shallow learners" (Support Vector Machine –SVM, Random Forest –RF, and shallow Neural Network –NN classifiers) and features we tried, the SVM classifier performed the best after executing a 10-fold cross-validation scheme using the manual explosions and non-explosive signals by comparing classification accuracies. When trained with features following those proposed by *Anzieta et al.* (2023), it replicated other successful SVM implementations (*Li et al.*, 2016; *Tang et al.*, 2020; *Toney et al.*, 2022).

Cleaning

As a final step to further clean the dataset, we created subsets of the newly detected events classified as explosions in the previous step, reused the manual non-explosions into new sets of events, and performed a cleaning procedure similar to the one described in *Anzieta et al.* (2023). This procedure consists of creating several mostly independent shallow classifiers to measure the "consistency" (or quality) of each event given its preassigned label, that is, how the many different classifiers label each event based on information learned from all the other events compared to the original label. If most classifiers agree with the original label then the event is highly consistent (thus most likely well labelled) and vice-versa. This way the quality of each event's label can be high or low depending on its consistency and a visual inspection or straight filtering of the low quality observations can then be applied in a "voting-filtering" scheme (*Brodley and Friedl*, 1999). In summary, for each event, many shallow classifiers are created with all the other events, and the quality of the event is defined by the number of agreements of the predictions with its original label. Poor quality events are those in which many classifiers disagreed with the original label and can be discarded while high-quality events can be kept.

3.3.2 Part 2: Classical Clustering

Once a clean catalogue was created, we applied, compared and combined two different techniques to identify families of repeating events. The first method was to cluster events based on waveform cross-correlation (CC) which is still one of the preferred techniques to establish template waveforms for matched filtering (e.g., *Machacca et al.*, 2023; *Tan et al.*, 2023). The second technique is clustering based on pseudo-distances between waveforms calculated using the Dynamic Time Warping (DTW) method (*Anzieta et al.*, 2019; *Ida et al.*, 2022). In the case of CC, the matrix is converted from correlation (similarity) to dissimilarity, and we refer to these operation or values as CC*.

These two processes are very computationally intensive (especially DTW) although they can be greatly sped up even in consumer grade computers by using Graphics Processing Unit (GPU) techniques when available (*Beaucé et al.*, 2017). For instance, DTW distance calculations between a few thousand observations can be performed in a few seconds using GPU approaches (*Hundt et al.*, 2014), as long as the number of observations can fit in the GPU memory ($\lesssim$ 5000 events). If a GPU is not available, several Python packages (*Tavenard et al.*, 2020; *Wannesm et al.*, 2022) have been optimized and can calculate DTW distances between hundreds to several thousands of observations from a few seconds to several hours in the CPU virtually without memory limitations. Nevertheless, since each event is compared to the rest, at some point increasing the number of events leads to a number of calculations ($\frac{n(n-1)}{2}$ calculations for a set of size n) that can rapidly exceed available resources (e.g., processing time or device memory). For such situations, it is possible to split datasets into subgroups to identify and isolate representative clusters within them, allowing for comparative analyses at a later stage (e.g., by clustering events in a day and then using those clusters to prove other days as *Allstadt and Malone*, 2014; *Lamb et al.*, 2020, 2022). Here, we divided the entire dataset into the distinct eruptive phases and grouped events within each phase to achieve processing times of seconds to a few hours for all phases, depending on the number of events (hundreds to a few thousands).

After the distance-matrices were found for each period, we visualize them as a proxy for evolution of explosivity, and then applied the Hierarchical Clustering algorithm using Ward's criterion (*Anzieta et al.*, 2019; *Ortiz et al.*, 2021b) to find representative families. There are several ways to cluster events: one is by defining a dissimilarity (or similarity) threshold and then considering the families of clustered events given this threshold, and another is by predefining a set number of families. Here we investigated fixing the number of families for several reasons: 1) In DTW there is no apparent value for strong association between events before inspecting the entire dataset so both (dis)similarity measures would not be comparable in a straightforward manner. 2) Thresholding is also arbitrary and different studies have used several different thresholds (e.g., 0.6 in *Allstadt and Malone*, 2014, 0.7 in *Lamb et al.*, 2020, and 0.8 in *Lamb et al.*, 2022), and even after a threshold is defined the minimum number of events for a family to be considered important is also arbitrary and may or may not leave out meaningful clusters. 3) Thresholding is not sufficient to guarantee high resemblance between events and can lead to associations that are not necessarily good (Figure C.1) or may leave events unassociated that may belong together (Figure C.2). 4) Fixing the number of families to a low number diminishes the clustering precision but allows for more potentially noisy events to be included in a cluster (*Green and Neuberg*, 2006). High thresholds imply smaller clusters that may appear disconnected or disappear altogether, while a coarse aggregation allows the study of general explosive behaviour during a long period, i.e., as long as 10 years of activity. 5) Despite coarser resolution, combining the cross-correlations and DTW can help alleviate the poor clustering by merging similar events using two factually different similarity approaches. Thus, for both (dis)similarity measures we clustered events into families of size 4 and 7 and explored coarse clustering. Since the labels coming from each distance clustering are arbitrary, to identify coincidental groupings between DTW and CC distances (i.e., events that were grouped simultaneously), we relabelled these such that the total amount of coincidences was maximal to produce families as large as possible –this is equivalent to permuting the rows and columns of the confusion matrix such that the trace of this matrix between the arbitrary labels is maximized as shown next:

$$\begin{array}{c} \\ CC\,f_1 \\ CC\,f_2 \\ CC\,f_3 \\ CC\,f_4 \end{array} \begin{array}{c} \begin{array}{cccc} DTW\,f_1 & DTW\,f_2 & DTW\,f_3 & DTW\,f_4 \end{array} \\ \begin{pmatrix} 857 & 117 & 476 & 76 \\ 938 & 1106 & 1312 & 20 \\ 519 & 23 & 57 & 417 \\ 1078 & 112 & 233 & 648 \end{pmatrix} \end{array} \rightarrow$$

$$\rightarrow \begin{array}{c} \\ CC\,f_1 \\ CC\,f_2 \\ CC\,f_3 \\ CC\,f_4 \end{array} \begin{array}{c} \begin{array}{cccc} DTW\,f_{1-3} & DTW\,f_2 & DTW\,f_{3-4} & DTW\,f_{4-1} \end{array} \\ \begin{pmatrix} 476 & 117 & 76 & 857 \\ 1312 & 1106 & 20 & 938 \\ 57 & 23 & 417 & 519 \\ 233 & 112 & 648 & 1078 \end{pmatrix} \end{array}$$

This optimal label swap can be found automatically in a fast manner using the Hungarian algorithm (a combinatorial optimization algorithm that minimizes or maximizes –the total cost– on a one-to-one assignment problem). By swapping labels, only events that are clustered together using both independent difference matrices are considered to belong to a given family, and by maximizing the trace, the most amount of consistent/coincidental clustered events are preserved. Finally, the time evolution of the events with respect to their families is studied for different periods.

3.3.3 Part 3: Deep Learning Clustering

The advent of deep learning methods –in particular Convolutional Neural Networks (CNN)– has been a major advance in pattern recognition in volcano seismology as these methods can learn feature representations from the raw data, and thus do not require "hand-crafted" features (*Titos et al.*, 2020; *Manley et al.*, 2022; *Trani et al.*, 2022). Nevertheless, as in more classic machine learning algorithms, there is no single "flavor" of Neural Network that fits all problems (*Titos et al.*, 2018, 2019; *Manley et al.*, 2022). Among the many available deep learning architectures one of the most interesting ones is the Convolutional Autoencoder (CA) because it can find efficient feature representations from images explicitly (*Géron*, 2019), making it well suited for spectrograms, scalograms or any other two-dimensional waveform characterization. Convolutional Autoencoders have had success in supervised (*Zheng et al.*, 2019; *Kong et al.*, 2021) and unsupervised learning tasks (*Jenkins II et al.*, 2021; *Zali et al.*, 2024), and are reasonably quick to train. Simply put, a CA will receive a signal, encode (compress) it, and attempt to reconstruct (decode) it from the compressed version. The interesting part is using this compressed representation to perform other tasks in an efficient manner. Here we also trained a CA focusing on compressing the waveform representations to then apply clustering, and additionally compared its performance with those of the classical methods described earlier.

To do so, as inputs we converted the 20 second time-domain acoustic traces to scalograms using the Continuous Wavelet Transform (CWT) (*Lee et al.*, 2019) as suggested by *Jenkins II et al.* (2021). To improve resolution on the most important portions of the signals, we created the 2D representations by focusing on frequencies between 0.17 and 8.0 Hz as most of the energy was contained in that frequency band (e.g., *Ruiz et al.*, 2006; *Ortiz et al.*, 2018, 2021b for Tungurahua volcano) and applied a complex Morlet wavelet with central frequency 1.5 and bandwidth 0.85 after exploration. We then reshaped the representations to images of resolution 80 × 144 pixels.

Then we constructed the CA architecture similar to those suggested by *Jenkins II et al.* (2021) and *Zali et al.* (2024) compatible with the 2D representations described in Table 3.1. We chose maximum amplitude normalization to normalize the 2D representations of the waveforms to be consistent with the previous techniques.

Table 3.1: Convolutional Autoencoder Architecture. The name, type, shape and activation function of the different layers forming the Convolutional Autoencoder network are described along with the number of trainable parameters for each layer.

Layer(depth)	**Type**	**Output Shape**	**Activation**	**Num. of params.**
Encoding				
Input(-)	InputLayer	(80,144,1)	-	-
Conv2D-1(8)	Conv2D	(40,72,8)	elu	288
Conv2D-2(16)	Conv2D	(20,36,16)	elu	1936
Conv2D-3(32)	Conv2D	(10,18,32)	elu	7712
Conv2D-4(64)	Conv2D	(5,9,64)	elu	30784
Flatten(-)	Flatten	(2880)	-	-
Dense1(1)	Dense	(26)	elu	74906
Decoding				
Dense2(1)	Dense	(2880)	elu	77760
Reshape(-)	Reshape	(5,9,64)	-	-
Conv2DT-4(64)	Conv2DTranspose	(10,18,32)	elu	30752
Conv2DT-3(32)	Conv2DTranspose	(20,36,16)	elu	7696
Conv2DT-2(16)	Conv2DTranspose	(40,72,8)	elu	1928
Conv2DT-1(8)	Conv2DTranspose	(80,144,1)	sigmoid	281
TOTAL				234043

After training and evaluating the Autoencoder, we used the encoder portion of the CA network to find the representation of each observation in a 26 dimensional vector. For the grouping of these latent representations, we used two classical schemes: k-means clustering as in *Zali et al.* (2024), and Gaussian mixtures modelling (GMM) clustering as in *Jenkins II et al.* (2021), and combined them in a similar fashion as for the clusterings from DTW and CC*, that is, by constructing a confusion matrix between both labellings and picking events that are grouped together by both schemes. Additionally, we retained clustered events that possessed a high degree of membership to their own class for a given clustering scheme. In the case of k-means clustering, the degree of membership was assessed by using the distance to the mean of each family as a measure of quality (less distant events to the same mean are considered more similar between them). For the GMM clustering, we used the probability of membership to each Gaussian mixture as the measure of quality (events with high probability of membership to the same mixture are considered more similar). In summary, we sought to ensure that events grouped into a family were assigned in a consistent way from two different methods' "perspectives" and with a high degree of membership for each method. Although we dealt with the whole dataset at once, the clustering process required a fraction of the elapsed time (a few minutes) compared to what would have been required to label all events as described in the previous subsection (due to the difference matrices calculations plus clustering). Finally, the total number of events belonging to both k-means and GMM clustering for 4 and 7 families was approximately two thirds of the explosions, but was reduced to 15,124 and 14,505 events respectively, after also applying quality restrictions (probability of membership higher than 0.99 for GMM clustering and distance lower than the 75th percentile of distances for k-means).

A workflow diagram summarizing the entire procedure is shown in Figure 3.4.

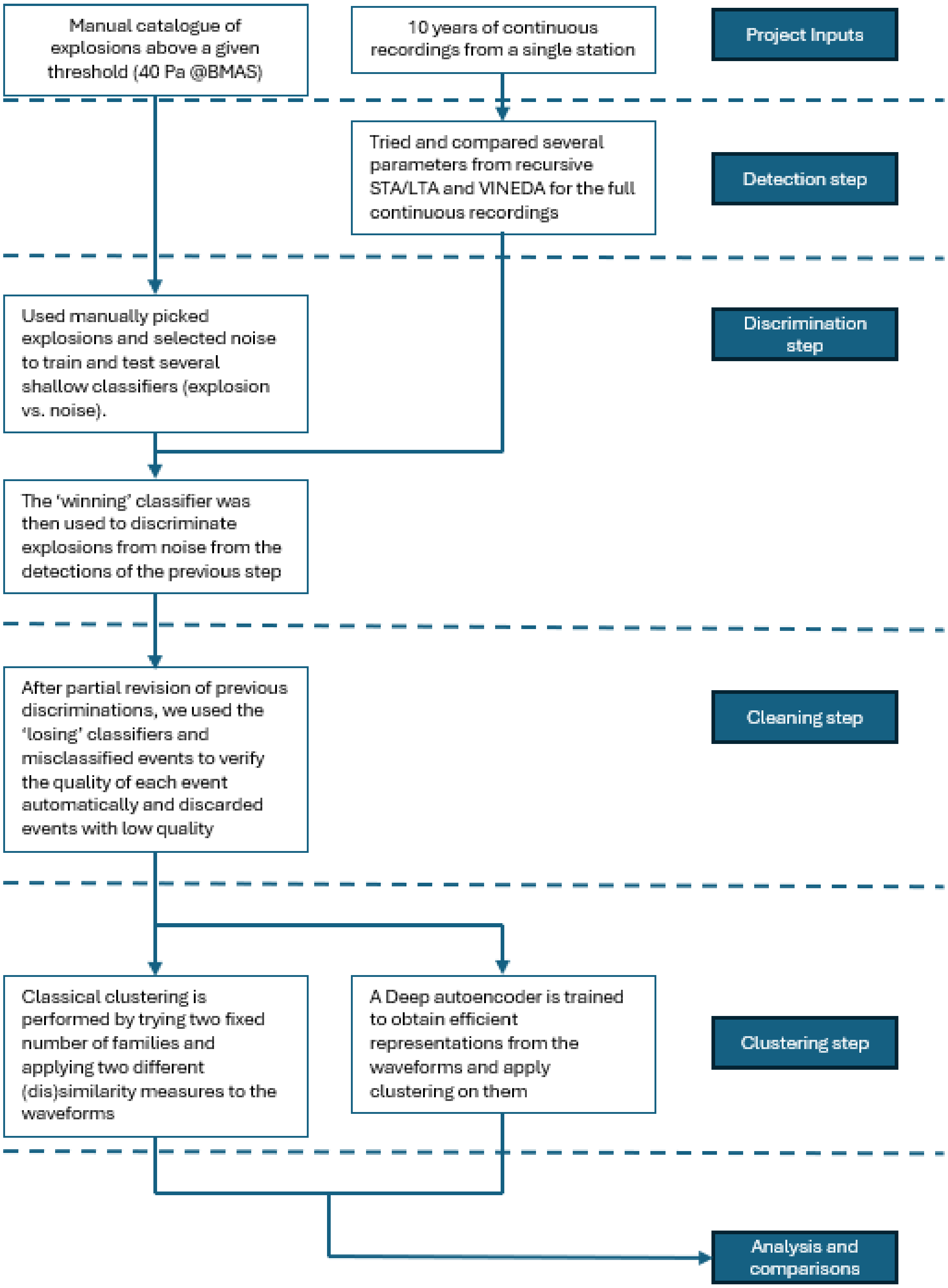
Manual catalogue of explosions above a given threshold (40 Pa @BMAS)
10 years of continuous recordings from a single station
Project Inputs
Tried and compared several parameters from recursive STA/LTA and VINEDA for the full continuous recordings
Detection step
Used manually picked explosions and selected noise to train and test several shallow classifiers (explosion vs. noise).
Discrimination step
The 'winning' classifier was then used to discriminate explosions from noise from the detections of the previous step
After partial revision of previous discriminations, we used the 'losing' classifiers and misclassified events to verify the quality of each event automatically and discarded events with low quality
Cleaning step
Classical clustering is performed by trying two fixed number of families and applying two different (dis)similarity measures to the waveforms
A Deep autoencoder is trained to obtain efficient representations from the waveforms and apply clustering on them
Clustering step
Analysis and comparisons

Figure 3.4: Workflow diagram of the processes applied in this study.

3.4 Results and Comments

3.4.1 Part 1: Data Increment

Detection

After inspection of the performance of both detection algorithms during different explosive rates, and attempts to identify a fixed set of parameters suitable for the diverse activity, we decided to risk incurring more false detections in favour of capturing potential explosion-like signals. Nevertheless, we opted to use VINEDA since even wrong detections had more impulsive behaviour than those that could be produced by the recursive STA/LTA when lowering the threshold too much (Figure 3.5).

From the application of VINEDA, we obtained 118,516 detections in a process that took a few minutes for the entire dataset. Because the algorithm still detected signals that did not necessarily belong to volcanic activity (e.g., instrumental/natural noise or other transient waveforms), we proceeded to use the manual explosions and manually picked non-explosions waveforms to discriminate between signals and noise.

Discrimination

From the 118,516 events obtained by VINEDA, 75,483 were classified as events related to volcanic activity and the rest were classified as noise using the SVM classifier trained with the events from the manual catalogue. However, while the internal cross-validation of the winning classifier led to an average accuracy greater than 93%, simple classifier systems tend to perform worse when faced with new data because they possess less generalization power than deep classifiers (*Géron*, 2019). To verify this, we manually reviewed nearly 30,000 out of the 75,483 events by inspecting traces from other stations (when available) and found that $\sim$ 21% of the events were dubiously classified as explosions and close to 7% of the events were definitely not explosions (i.e., noise with shapes resembling explosions). It is not surprising that there were so many events of dubious nature since we started from a manual catalogue created with a large peak-to-peak amplitude threshold ($\sim$ 40 Pa) many times larger than background noise (Figure 3.5a), leading to detection and classification of events with amplitudes of $\sim$ 1 Pascal (Figure 3.5c). As such, in order to preserve events

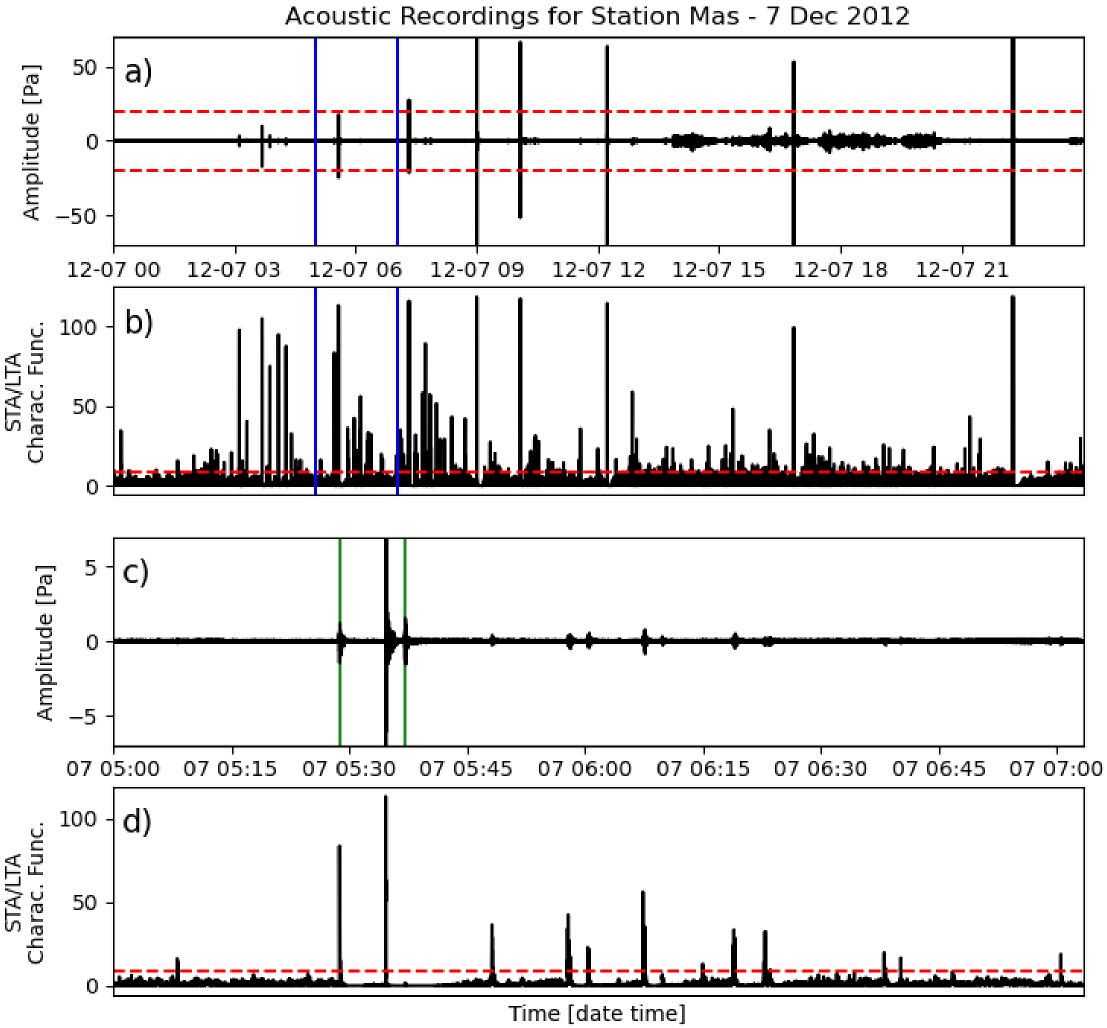

Figure 3.5: Example of detection and discrimination stages. a) Daily trace with visual ∼ 40 Pa threshold as the dashed red lines. Zoomed region delimited by vertical blue lines. b) STA/LTA Characteristic function for the whole day with detection threshold as the dashed red line. Blue lines define the zoomed regions in c-d. c) Zoomed trace. Small explosions detected by a "shallow learner" are highlighted in green. d) STA/LTA Characteristic Function for the zoomed trace. Lowering the detection threshold would increase false detections.

that are most likely related to volcanic activity, we applied the cleaning procedure with the following details.

Cleaning

In the original version of the catalogue-cleaning procedure proposed by *Anzieta et al.* (2023), the focus was on cleaning already reviewed/published catalogues to identify potentially mislabelled events by assessing events with low consistency compared to information from other events in the catalogue. However, a good quality catalogue can be created quickly by just preserving events with high consistency. Here, we opted to apply that procedure to the 75,483 remaining events from the previous step by partitioning the whole dataset into 12 non-overlapping subgroups of ∼ 6000 events labelled as "explosions" and adding the original manually picked non-explosions plus the truly noisy events we identified through visual inspection. To create the shallow learners, we used the classifiers and features not used in the previous discrimination step. This means that we used all the combinations between the features by *Titos et al.* (2018); *Watson* (2020); *Anzieta et al.* (2023) and SVM, RF and NN classifiers except the winner used in the discrimination step (SVM with features from *Anzieta et al.*, 2023). After applying the cleaning procedure to each of these subgroups and retaining only explosion events with consistencies higher than 0.8 (equivalent to ≥7/8 coincidences between original label and predictions), we identified 36,359 events with high confidence; a more than 6 fold increase compared to the original catalogue (Figure 3.6).

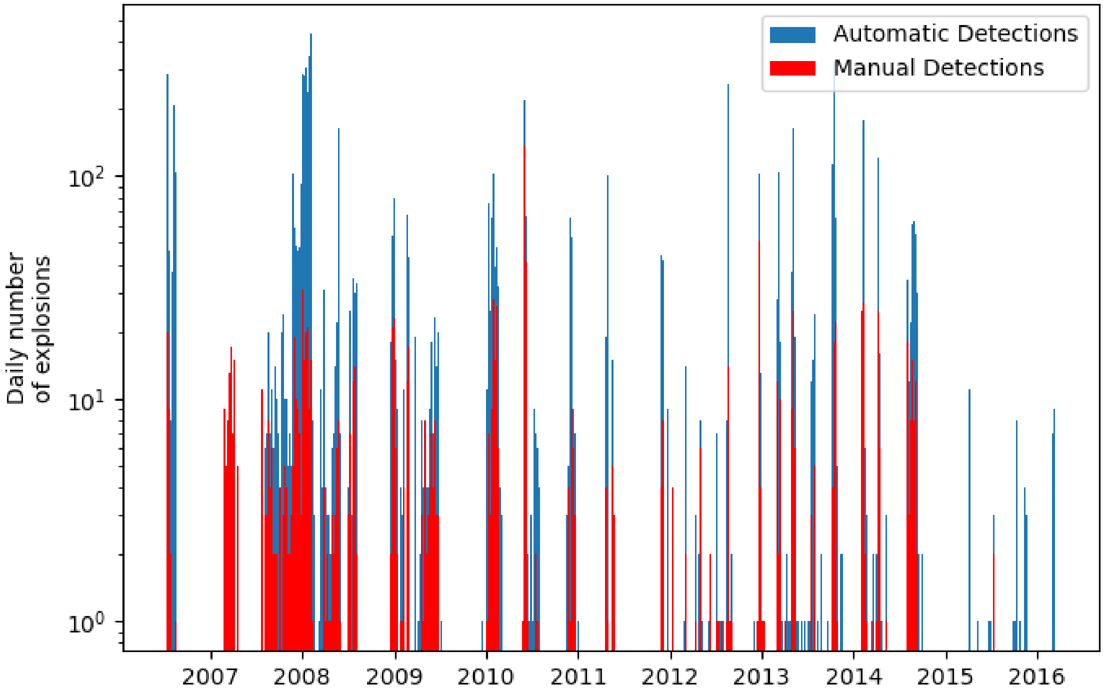

Figure 3.6: Superimposed daily explosions from the original manually determined catalogue (red) and new automatic detections (blue)

3.4.2 Part 2: Classical Clustering

After obtaining the 36,359 events, we divided the whole dataset into subsets of eruptive phases separated by quiescence of several days (*Hidalgo et al.*, 2015) to cluster events within periods. Before clustering, distance/dissimilarity (or similarity) matrices must be calculated and the matrices themselves can be very informative about the state of activity/evolution of a volcano. It is important to note that DTW pseudo-distances are inherently different than those from CC* not only because they are obtained from a non-linear operation, but also because they do not possess a fixed upper-bound for maximum dissimilarity (for CC* maximum dissimilarity is 1). Nevertheless, both measures should lead to similar results. To show that this behaviour also transferred between several eruptive phases, we calculated DTW and CC* dissimilarity matrices for the events of five consecutive eruptive periods studied by *Ortiz et al.* (2018) and described in Table 3.2; qualitatively analogous results were obtained for other periods.

Table 3.2: Focused Periods of Eruptive Activity.

Period	**Date**	**Number of Events**
1	Jan 2010 to March 2010	2046
2	May 2010 to Aug 2010	2490
3	Nov 2010 to Jan 2011	1305
4	Apr 2011 to May 2011	445
5	Nov 2011 to Sep 2012	1701
Total		**7987**

Once the difference matrices were generated and examined (Figure 3.7), the next step was to cluster and identify highly similar events combining them. To perform the clustering, we applied the hierarchical clustering method using Ward's criterion as in *Anzieta et al.* (2019) and *Ortiz et al.* (2021b). We predicted labels for families of size 4 and 7 to show a coarse and fine event discrimination/grouping. This process delivered 4 families with 3077 total events and 7 families with 2478 total events which correspond to 38% and 31% of the total detections during this time frame. The time evolution of events that were clustered together combining both matrices is shown in Figure 3.8.

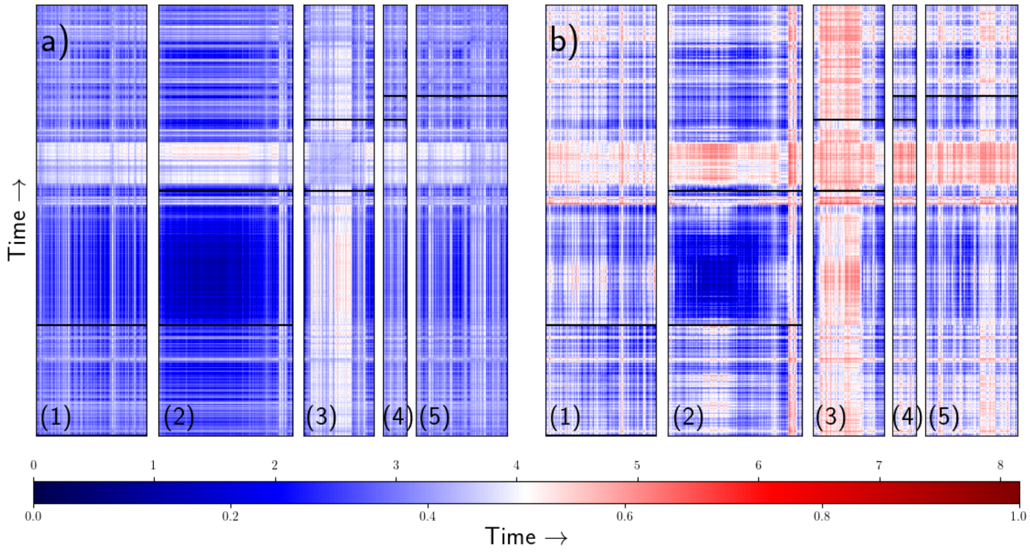

Figure 3.7: Dissimilarity matrices obtained using DTW and CC*. a) DTW dissimilarity matrix. b) CC* dissimilarity matrix. Periods were divided visually and numbered following Table 3.2. Blue depicts similarity and white/red dissimilarity. Black lines delimit same-period dissimilarity portions. Blue colors depict similarity and red dissimilarity according to the colorbar values.

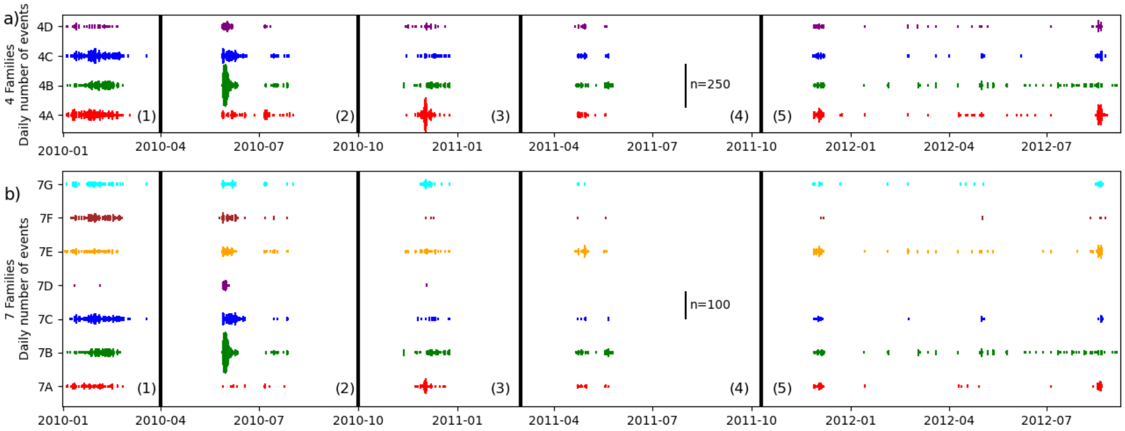

Figure 3.8: Daily occurrences of events clustered using DTW joint with CC* grouped into 4 (a) and 7 (b) families for the examined period. Larger marker sizes indicate more event occurrences on a given date as given by the solid black marker. Vertical black lines separate the periods and they are numbered according to Table 3.2. Gaps in detections are due to surficial inactivity.

3.4.3 Part 3: Deep Learning Clustering

We used 30% of the 36,359 acoustic events for training and 70% for validation and obtained good reconstruction results from the scalograms probably because the CWT is very well suited for non-stationary processes (*Kumar and Foufoula-Georgiou*, 1997).

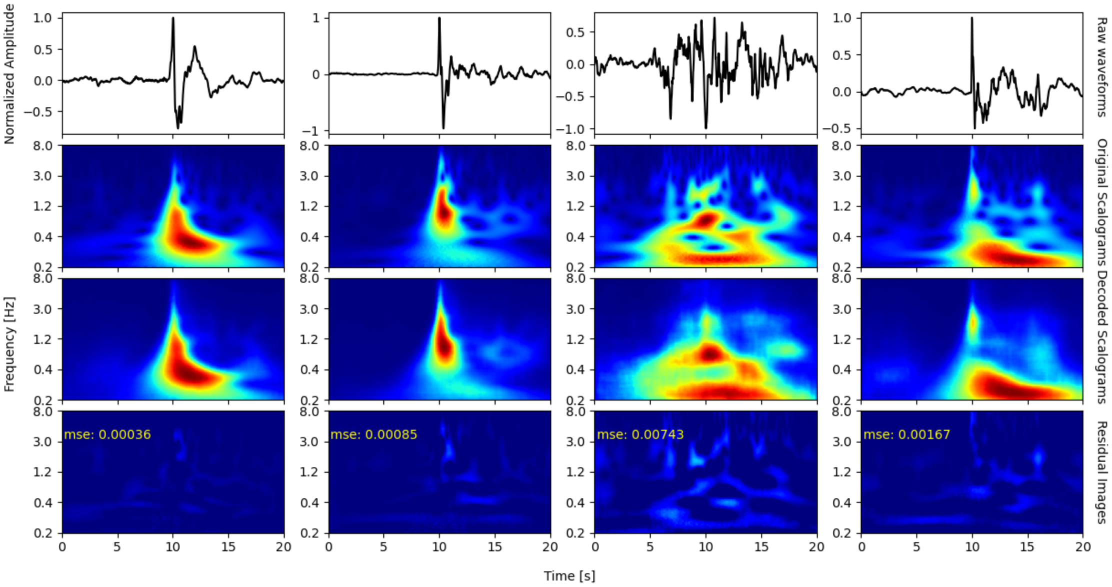

Figure 3.9: Examples of raw acoustic event waveforms, their normalized scalograms, their decoded images and residual images from the trained autoencoder network. The residual image shows the mean-squared error (MSE) between the reconstructed image and the original one.

As a reminder, to obtain good latent representations, CA must learn to reconstruct their inputs as faithfully as possible. Once an input is reconstructed, one can compare the original 2D-representation with the reconstructed one and characterize qualitatively as well as quantitatively differences between them. In Figure 3.9 we show 4 examples of acoustic events with different shapes, along with their original scalograms, their reconstructed scalograms, and their residual images. The residual image also shows the mean-squared error (MSE) between the reconstructed image and the original one; we only used that value as a measure of error to have an intuitive idea of the mean difference in "pixel value" between images, and not as the loss-function for training the network. This is because we weighed vector normalization as in *Jenkins II et al.* (2021) and *Zali et al.* (2024) against maximum amplitude normalization (with sigmoid activation functions and binary cross-entropy as the loss function) as suggested by *Géron* (2019), and opted for the later combination because it returned smaller MSE between original and reconstructed images even though the preferred loss-function for a sigmoid activation is the binary cross-entropy and we used that one here. This standard approach is better for image-like inputs because images have an inherent lower bound, i.e., there are no negative values for pixels, so mappings to reconstruct images (either spectrograms or scalograms) are naturally bounded (by zero). The reconstructions are mostly faithful to the original images, although some of the small amplitude fine details are lost in the noisier signals (Figure 3.9). This is another property of CA that, thanks to their power to reconstruct the main features of the signals, have been proposed as denoisers (*Zhu et al.*, 2019). It also suggests that the latent 26-dimensional representations are adequate to attempt clustering.

Figure 3.10 shows the distribution of events for clusterings of size 4 and 7 combining k-means and GMM, plus the quality restrictions (probability of membership higher than 0.99 and distance lower than the 75th percentile of distances).

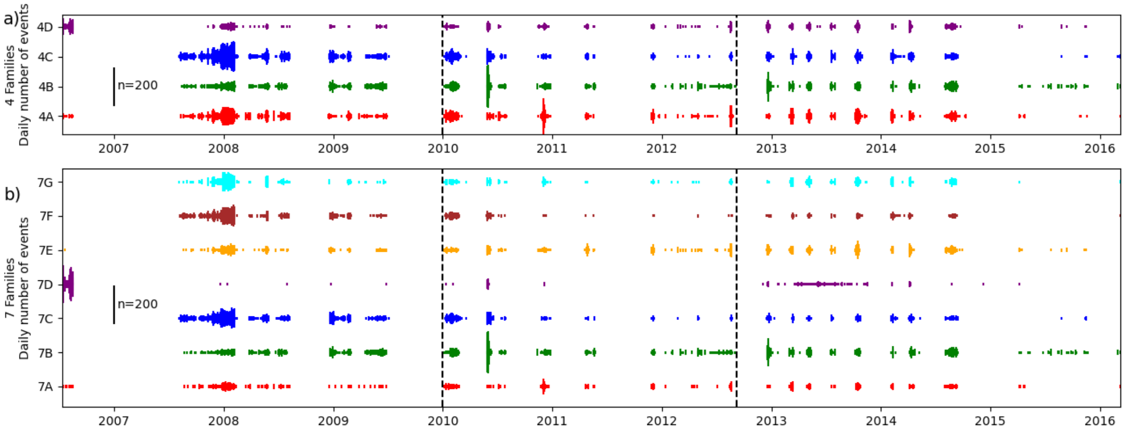

Figure 3.10: Daily occurrences of events clustered using DTW joint with CC* grouped into 4 (top) and 7 (bottom) families for the whole dataset. Larger marker sizes indicate more event occurrences on a given date as given by the solid black marker. Vertical dashed black lines enclose the periods studied in Figure 3.8. Gaps in detections are due to surficial inactivity.

3.5 Discussion

The inspection of various parameters of both the recursive STA/LTA and VINEDA led us to verify that there is no single set of parameters that maximize true detections and minimize false detections during days with different explosive rates. Although in cases of low explosive activity it was easier to tune the recursive STA/LTA, in cases of high activity or noise it had less sensitivity to impulsive signals than VINEDA. This could be especially problematic because explosions could sometimes occur in the middle of the transition between rapid explosions to emissions tremor or vice versa (e.g., *Fee et al.*, 2010; Figure C.3). Also, while VINEDA has many parameters that make it more complex to tune than STA/LTA, after extensive exploration, we found that given some prior knowledge of the events (expected durations and frequency content) that allow fixing the duration and frequency ranges, one can focus only on a couple of parameters to optimize performance. For instance, the most sensitive parameters for detections are beta and the number of frequency bands which cause noticeable changes in number of true and false detections in a consistent manner. Lower beta led to more detections overall, and a higher number of frequency bands reduced detections but favoured reduced false detections slightly, while the number of duration bands led to minor changes in detection numbers. Thus, after setting parameters that can be easily surveyed in a given dataset, emphasis can be placed on adjusting beta and the number of frequency bands, making it comparable to STA/LTA schemes (i.e., testing of short and long windows durations and on and off thresholds). Nevertheless, depending on environmental conditions or instrumental defects/problems, atypical signals unrelated to volcanic activity can still bypass classical detectors. For this reason it is important to obtain catalogues that are as clean as possible before applying sophisticated tools since the noise greatly affects the predictive capability of any system, independently of its sophistication. Cleaning efforts are usually time intensive and different schemes can be tried to alleviate the time devoted to this task. At every step in obtaining a good quality dataset, one must consider that the more restrictions one implements into an automatic or semi-automatic tool, the greater amount of potentially good data may be missed.

Here, we used traditional tools (not Machine Learning related) for detection because they are fast to apply, but then used different forms of supervised learning with classifiers that are easy to implement and require little tuning to further clean detections from classical algorithms. This created a robust dataset over 6 times larger than the original manually created catalogue in a timely manner, and we believe this procedure can be applied to similar datasets so that they become more complete for further knowledge extraction. While SVM was a good shallow classifier, it is still important to explore other classifiers such as Random Forests that have also been tried with good success (e.g., *Malfante et al.*, 2018b; *Falcin et al.*, 2021). When time permits, a deep study for the best features for each classifier and dataset following a comprehensive approach such as that proposed by *Toney et al.* (2022) is also recommended.

The increase of $\sim$ 6.33 times (from 5740 to 36,359) the number of events obtained through human inspection (Figure 3.6) through automated methods was also observed in *Ortiz et al.* (2018). While celebrating that there was an increase in the number of explosions just from the fact that the 40 Pa threshold was high to begin with may be too optimistic, the key point is that this procedure is semi-automatic and can reduce the inspection time substantially. The determination of the $\sim$ 6000 explosions detected manually took a few weeks of work with clear explosions (high signal to noise ratio), so we can argue that producing the same work for the current catalogue might take considerable efforts and time. Still, we note that just like other machine learning tools, human revision is advised when possible. While changing propagation conditions will affect detections and thus different periods could show artificial explosion rate variations, we argue that in the bad scenarios explosions will generally be obfuscated, but in times of high explosivity (in terms of number and intensity of explosions) the amount of detections will increase independently of families. Since we observed different types of explosions alternate in time arbitrarily, at least relative changes between clusters in a given day or period should be credible. However, it is a fact that many more explosions were not included because of the use of a single station and probable bias in the cleaning stages, and explosion sequences may not alternate between families randomly. Thus, we acknowledge that since this catalogue is but a first approximation, a more complete one can be created considering information from other stations and the seismic components, and correlations between number of detections and weather parameters may help resolve this question.

The clustering task is of particular interest and importance because the existence of diverse events repeating and interspersing in time tells us about recurrent dynamical processes related to the state of activity of a volcano. Even dissimilarity matrices on their own possess important analytical value to identify temporal patterns before clustering is applied, but clustering makes the picture clearer. At the same time, clustering is not a straightforward task and there is no single algorithm that suitably solves all problems.

The striking similarity between the two matrices (Figure 3.7) indicates that CC* as well as DTW methods provided similar insights into the temporal relationships between events. At the same time there are small differences that enable identification and discrimination of "super similar events", i.e., pairs or groups of events that are deemed highly similar using both criteria. Visual observation of period 2 suggests that a large number of events with high similarity in both matrices occurred consistently throughout the entire time frame, and it appears that these events may have been present in all periods to some extent. In contrast, the following period (period 3 in Figure 3.7) is comprised of events that are different to almost every other event during all times. This may be due to notable changes in the source of the events or related to a major vulcanian eruption on November 22, 2010 (*Ortiz et al.*, 2018). Nevertheless, several events repeat in time in spite of crater location changes detected during these periods by *Ortiz et al.* (2018). This suggests that source parameters controlling the appearance of explosive waveforms are likely influenced not only by the relative vent location or possible vent geometries, but also by the interplay of deeper factors over time (plus path effects). For instance, given a family of repeating events, it is assumed that their source should essentially be the same, yet amplitudes can vary. This suggests that for different repeating events, the generation process is similar except in intensity. This could imply then that even if the surficial location stays seemingly the same, it is on a deeper level below the crater that a transient phenomenon must repeat for a given family of events yet be different between families. Further, this transient process is non destructive at such levels so that waveforms ‘repeat’ even after large eruptions or after many years of activity.

Perhaps, a complex network of fractures (like those proposed by *Butcher et al.*, 2020) that connect roughly to the same surficial location become alternatingly permeable and shape the waveforms, or are part of the generating process for the explosions themselves. However, elucidating this question, is a step that will require a similar yet extended approach using the seismo-acoustic data from all available stations to better probe the sources (e.g., *Ruiz et al.*, 2006).

Notably, most events reoccur throughout all eruptive phases, although some families tend to be more prominent than others at certain times before they all reappear in a balanced manner towards the end of the studied periods. For both groupings in periods 2 and 3 on Figure 3.8, two different families of explosions (A and B) show a large number of events with respect to others but none completely disappear. This could be related to a change in eruptive behaviour due to the change in the location (and other parameters) of the vent/-source, but also suggests simultaneous competing source mechanisms that become more or less prominent over time. It is also clear that partitioning the events into more families can dilute the knowledge gained because some families become too under-represented. However, we cannot rule out that the existence of rare but repetitive events may be linked to distinct processes associated with eruptions as shown in *Anzieta et al.* (2019).

While this exercise provides interesting insights by clustering events between different periods, using larger number of explosions and applying these techniques as described eventually becomes infeasible as mentioned earlier. For this reason, as the objective of this exercise was to revise the full extent of the dataset ($\sim$ 10 years of events), to obtain similar detailed results for the whole dataset we applied and show the results of the more modern deep learning techniques below.

Here we contrasted the results of using two (dis)similarity measurements to look for clustering and patterns in the data distilled from the detection and cleaning steps, and showed that different criteria should still deliver similar knowledge (e.g., similar behaviour for the same periods of activity as shown in Figure 3.7), and that their combination helps solidify it. The main limitation of classical clustering through waveform direct comparisons is computational burden (if distances between all pairs of events are considered), so in cases of limited time they are mostly well suited for moderate datasets. Here, we consider moderate datasets those that have only a few thousand events and can fit in the memory of a consumer-grade GPU that allows all-against-all calculations in a few seconds as opposed to minutes/hours in a consumer-grade CPU.

While there is no single algorithm or procedure to solve all tasks, deep learning techniques are becoming more prevalent because of the strong generalization capabilities they possess. As noted in the *methods* section, two of the strongest points of Convolutional Autoencoders are the generalization ability (inherited from deep learning) and efficient representation (by optimized encoding-decoding) that make it a suitable tool to engage with the full dataset. The other important feature of the CA network architectures proposed by *Jenkins II et al.* (2021) and *Zali et al.* (2024) is that they are fairly quick to train. While tuning of training parameters (hyper-parameters) can be a lengthy process (and "an art" for some), using a consumer grade GPU can greatly decrease the time required to test different hyper-parameters, to later train the whole dataset in a CPU if memory is insufficient for the whole dataset. Although they generally require a large amount of good quality data (preferably tens of thousands of events or more for a process such as the one described here) and parameter tuning to be trained properly, after they "learn" a task, they perform satisfactorily (as seen in Figure 3.10 vs Figure 3.8, Figure 3.7 vs. Figure C.4, Figures C.5 to C.15) and generally faster than almost any other procedure given observations/datasets of the same nature. For instance, producing the inputs for Figure C.4 took 30 min with the same computer that required >7 hours to produce Figure 3.7b. Additionally, continuous improvements in hardware and software capabilities make deploying these deep learning systems easier with time. Figures C.5 to C.15 show samples of the families of events based on the CA clustering and they generally show similar semblance, which makes it in agreement with traditional results. Nevertheless, it must be mentioned that separating events into more families tends to blur the major differences between them (e.g., waveforms from families 7A vs 7G, or 7F vs 7C in Appendix C.4) and thus interpretations on what they precisely mean in time related to the behaviour of the volcano is harder.

For the entire dataset (Figure 3.10), there are obvious periods with more acoustic activity and there is a similar behaviour with the analyzed periods (Table 3.2, Figure 3.8), i.e., the corresponding periods A and B have families that surpass the others in terms of events, but no event family is totally absent.

In general, behaviour between clusterings is qualitatively similar (Figure 3.8 and Figure 3.10), with some families taking precedence over others at times, or bursts of events increasing in number for all families (e.g., Figure 3.10 from late 2007 to early 2008). However, due to the large number of events, in this case there is no family that is almost completely "diluted" by having too few events when using 7 families (family 7F in Figure 3.8b vs. families in Figure 3.10b bottom); although one of the families shows an almost complete and distinctively long "quiescence" since 2006 to 2013 (family 7D in Figure 3.10b). A few events from each family are shown in Appendix C.4.

To our knowledge, this is the first time classical clusterings of acoustic signals have been directly compared with clusterings from encoded signals from a CA. This opens the avenue to consider improving deep learning approaches that may lead to ultimately equal performances to traditional tools and could allow fine exploration of more large datasets faster (for example, after using all stations and discriminating even lower signal to noise events). Nevertheless, the traditional techniques remain a good, if not the best standard, when there are no resource restrictions and can be combined with the more modern tools (e.g., classical tools such as k-means and GMM clustering with "deep" latent representations). Comparatively, k-means clustering led to more consistent groupings than GMM but no algorithm was perfect at clustering events when analyzing results on their own. For instance, only a few of the examples of repeating events in Figure 3.3 met all the criteria for clustering, although those that did, belonged to families in concordance with their apparent shapes (similar to the samples in Figures C.5 to C.15). On the other hand, when events met all the criteria, the resemblance between same family observations increased noticeably but reduced the number of events dramatically, to less than half of the dataset. To overcome this problem, a few high-quality events from each family could be inspected quickly and they can be used as labelled events to implement a semi-supervised learning (learning from a few labelled observations) scheme (e.g., **?**), in a time critical situation, the same pre-trained CA network could be used to create a classifying network (*Géron*, 2019).

On the other hand, the application of improved deep learning networks or perhaps the widespread adoption of Neural Operators to learn and rapidly evaluate functions (*Kovachki et al.*, 2023), may provide the volcano-seismological community the tools to more effectively and efficiently analyze the ever expanding datasets. Neural Operators are a recent paradigm in deep learning that map between infinite dimensional function spaces as opposed to finite dimensional vectors (of regular neural networks). They are currently being used in the seismological community for multi-station simultaneous picking of seismic arrivals irrespective of the network geometry with excellent results (*Sun et al.*, 2023), and thus their potential to improve tasks is still unknown but promising.

Here we showed that a Convolutional Autoencoder, already tested successfully in seismic datasets, translated very well into a purely acoustic dataset with minor modifications (scalogram representations and maximum amplitude normalization with binary cross-entropy loss function) and thus can also translate well into similar datasets. This procedure greatly reduced the time to explore the whole dataset: after the Autoencoder was trained (which effectively took a couple of hours because of hyper-parameter optimization, but not training time), encoding the wavelets and clustering them takes from a few minutes (few thousand events) to a couple of hours (full dataset) depending on the size of the eruptive period used. This is in contrast to the several hours to several days required when following the classical procedures in our processing system and thus reinforces the significant advantages of deep learning approaches.

As a final comment, all these results can be replicated from a multi-station standpoint so that lower uncertainties can be achieved in all tasks: detection, cleaning, classification, clustering, etc., although independent subsystems are recommended in case of station loss or other data-loss problems. In particular, the same tools can be applied to combine acoustic with co-located seismic recordings, and obtain even deeper pattern recognition and understanding of volcanic processes. The salient advantage of using multi-station approaches is immediate as the cleaning stage is omitted because network coincidences can be used to discard false detections.

3.6 Conclusions

This work encompasses the exploration of many classical and novel data processing techniques applied to a very large dataset originating from almost continuous acoustic recordings from 2006 to 2016 at Tungurahua volcano while it was still erupting. The task consisted of trying to increase and improve the quality of an existing catalogue of manually detected explosions. We simulated the common situation of possessing a single acoustic sensor at a volcano and analyzed the data by dividing the problem into consecutive detection, cleaning and clustering steps. We compared and combined many available tools while trying to find a compromise between accuracy of detections and resource investment. This confirmed that there is no universal, perfect algorithm to complete a given task and it is best to actually use them in tandem to compensate for inherent weaknesses. As means to rapidly analyze the whole dataset simultaneously, we verified that a Convolutional Autoencoder network architecture already proven for seismic datasets handles the task successfully for acoustic signals with minor modifications. For this particular dataset, we found that the scalogram obtained from Continuous Wavelet Transform is a better two-dimensional descriptor than the more common spectrogram, and argue that it should be tried for other discrete events. In a similar fashion we also found that using maximum amplitude normalization for scalograms, sigmoid function activation for the output layer, and binary-cross entropy as the loss function for the Convolutional Autoencoder network delivers better results than vector normalization, linear activation function for the output layer, and mean squared error as the loss function, respectively. By employing a comprehensive approach, we extended the manually curated catalogue from around 6000 events to over 36,000 events. This allowed us to identify recurring groups of events and their interplay throughout the ten years of recordings at Tungurahua volcano. The relative simplicity of the implementation of this procedure should motivate its adoption for other extensive datasets.

References

Allen, R. V. (1978), Automatic earthquake recognition and timing from single traces, *Bulletin of the Seismological Society of America*, *68*(5), 1521–1532, doi:10.1785/BSSA0680051521.

Allstadt, K., and S. D. Malone (2014), Swarms of repeating stick-slip icequakes triggered by snow loading at mount rainier volcano, *Journal of Geophysical Research: Earth Surface*, *119*(5), 1180–1203, doi:https://doi.org/10.1002/2014JF003086.

Anderson, J. F., J. B. Johnson, A. L. Steele, M. C. Ruiz, and B. D. Brand (2018), Diverse eruptive activity revealed by acoustic and electromagnetic observations of the 14 july 2013 intense vulcanian eruption of tungurahua volcano, ecuador, *Geophysical Research Letters*, *45*(7), 2976–2985, doi:https://doi.org/10.1002/2017GL076419.

Anzieta, J., H. Ortiz, G. Arias, and M. C. Ruiz (2019), Finding Possible Precursors for the 2015 Cotopaxi Volcano Eruption Using Unsupervised Machine Learning Techniques, *International Journal of Geophysics*, *2019*, 1–9, doi:10.1155/2019/6526898.

Anzieta, J., D. Pacheco, G. Williams-Jones, and M. C. Ruiz (2023), Cleaning volcano-seismic event catalogues: a machine learning application for robust systems and potential crises in volcano observatories, *Bulletin of Volcanology*, *85*, 1–16, doi:10.1007/s00445-023-01674-9.

Arellano, S., M. Hall, P. Samaniego, J.-L. Le Pennec, A. Ruiz, I. Molina, and H. Yepes (2008), Degassing patterns of Tungurahua volcano (Ecuador) during the 1999–2006 eruptive period, inferred from remote spectroscopic measurements of SO2 emissions, *Journal of Volcanology and Geothermal Research*, *176*(1), 151–162, doi:10.1016/j.jvolgeores.2008.07.007.

Arrowsmith, S. J., J. B. Johnson, D. P. Drob, and M. A. H. Hedlin (2010), The seismoacoustic wavefield: A new paradigm in studying geophysical phenomena, *Reviews of Geophysics*, *48*(4), 1–23, doi:https://doi.org/10.1029/2010RG000335.

Battaglia, J., S. Hidalgo, B. Bernard, A. Steele, S. Arellano, and K. Acuña (2019), Autopsy of an eruptive phase of Tungurahua volcano (Ecuador) through coupling of seismo-acoustic and SO 2 recordings with ash characteristics, *Earth and Planetary Science Letters*, *511*, 223–232, doi:10.1016/j.epsl.2019.01.042.

Beaucé, E., W. B. Frank, and A. Romanenko (2017), Fast Matched Filter (FMF): An Efficient Seismic Matched-Filter Search for Both CPU and GPU Architectures, *Seismological Research Letters*, *89*(1), 165–172, doi:10.1785/0220170181.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010), ObsPy: A Python Toolbox for Seismology, *Seismological Research Letters*, *81*(3), 530–533, doi:10.1785/gssrl.81.3.530.

Brodley, C., and M. Friedl (1999), Identifying Mislabeled Training Data, *Journal of Artificial Intelligence Research*, *11*, 131–167, doi:10.1613/jair.606.

Bueno, A., A. Diaz-Moreno, I. Álvarez, A. De la Torre, O. D. Lamb, L. Zuccarello, and S. De Angelis (2019), Vineda—volcanic infrasound explosions detector algorithm, *Frontiers in Earth Science*, *7*, 1–8, doi:10.3389/feart.2019.00335.

Bueno, A., C. Benítez, S. De Angelis, A. Díaz Moreno, and J. M. Ibáñez (2020), Volcano-seismic transfer learning and uncertainty quantification with bayesian neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, *58*(2), 892–902, doi:10.1109/TGRS.2019.2941494.

Butcher, S., A. F. Bell, S. Hernandez, E. Calder, M. Ruiz, and P. Mothes (2020), Drumbeat lp “aftershocks” to a failed explosive eruption at tungurahua volcano, ecuador, *Geophysical Research Letters*, *47*(16), e2020GL088,301, doi:https://doi.org/10.1029/2020GL088301, e2020GL088301 10.1029/2020GL088301.

Canário, J. P., R. F. de Mello, M. Curilem, F. Huenupan, and R. A. Rios (2020), Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification, *Data in Brief*, *30*, 105,627, doi:https://doi.org/10.1016/j.dib.2020.105627.

Carniel, R., and S. R. Guzmán (2020), Machine learning in volcanology: A review, in *Updates in Volcanology*, edited by K. Németh, chap. 5, pp. 1–26, IntechOpen, Rijeka, doi:10.5772/intechopen.94217.

Cortés, G., R. Carniel, P. Lesage, M. A. Mendoza, and I. Della Lucia (2021), Practical volcano-independent recognition of seismic events: Vulcan.ears project, *Frontiers in Earth Science*, *8*, 1–11, doi:10.3389/feart.2020.616676.

Dannemann, F., C. Koch, E. Berg, S. Arrowsmith, and S. Albert (2023), A New Decade in Seismoacoustics (2010–2022), *Bulletin of the Seismological Society of America*, *113*(4), 1390–1423, doi:10.1785/0120220157.

De Plaen, R. S. M., T. Lecocq, C. Caudron, V. Ferrazzini, and O. Francis (2016), Single-station monitoring of volcanoes using seismic ambient noise, *Geophysical Research Letters*, *43*(16), 8511–8518, doi:https://doi.org/10.1002/2016GL070078.

Deng, L., and D. Yu (2014), *Deep Learning: Methods and Applications*, vol. 7, 197-387 pp., Now Foundations and Trends, doi:10.1561/2000000039.

Douillet, G. A., D. A. Pacheco, U. Kueppers, J. Letort, È. Tsang-Hin-Sun, J. Bustillos, M. Hall, P. Ramón, and D. B. Dingwell (2013), Dune bedforms produced by dilute pyroclastic density currents from the August 2006 eruption of Tungurahua volcano, Ecuador, *Bulletin of Volcanology*, *75*(11), 762, doi:10.1007/s00445-013-0762-x.

Duque, A., K. González, N. Pérez, D. Benítez, F. Grijalva, R. Lara-Cueva, and M. Ruiz (2020), Exploring the unsupervised classification of seismic events of cotopaxi volcano, *Journal of Volcanology and Geothermal Research*, *403*, 107,009, doi:https://doi.org/10.1016/j.jvolgeores.2020.107009.

Esposito, A. M., F. Giudicepietro, L. D'Auria, S. Scarpetta, M. G. Martini, M. Coltelli, and M. Marinaro (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, doi:10.1785/0120070110.

Eychenne, J., J.-L. Le Pennec, L. Troncoso, M. Gouhier, and J.-M. Nedelec (2012), Causes and consequences of bimodal grain-size distribution of tephra fall deposited during the August 2006 Tungurahua eruption (Ecuador), *Bulletin of Volcanology*, *74*(1), 187–205, doi:10.1007/s00445-011-0517-5.

Falcin, A., J.-P. Métaxian, J. Mars, Éléonore Stutzmann, J.-C. Komorowski, R. Moretti, M. Malfante, F. Beauducel, J.-M. Saurel, C. Dessert, A. Burtin, G. Ucciani, J.-B. de Chabalier, and A. Lemarchand (2021), A machine-learning approach for automatic classification of volcanic seismicity at la soufrière volcano, guadeloupe, *Journal of Volcanology and Geothermal Research*, *411*, 107,151, doi:https://doi.org/10.1016/j.jvolgeores.2020.107151.

Fee, D., M. Garces, and A. Steffke (2010), Infrasound from Tungurahua Volcano 2006–2008: Strombolian to Plinian eruptive activity, *Journal of Volcanology and Geothermal Research*, *193*(1-2), 67–81, doi:10.1016/j.jvolgeores.2010.03.006.

Géron, A. (2019), *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, second ed., O'Reilly, CA 95472.

Green, D. N., and J. Neuberg (2006), Waveform classification of volcanic low-frequency earthquake swarms and its implication at soufrière hills volcano, montserrat, *Journal of Volcanology and Geothermal Research*, *153*(1), 51–63, doi:https://doi.org/10.1016/j.jvolgeores.2005.08.003, mULTIMO: Multi-Parameter Monitoring, Modelling and Forecasting of Volcanic Hazard.

Hall, M. L., A. L. Steele, B. Bernard, P. A. Mothes, S. X. Vallejo, G. A. Douillet, P. A. Ramón, S. X. Aguaiza, and M. C. Ruiz (2015), Sequential plug formation, disintegration by Vulcanian explosions, and the generation of granular Pyroclastic Density Currents at Tungurahua volcano (2013–2014), Ecuador, *Journal of Volcanology and Geothermal Research*, *306*, 90–103, doi:10.1016/j.jvolgeores.2015.09.009.

Hidalgo, S., J. Battaglia, S. Arellano, A. Steele, B. Bernard, J. Bourquin, B. Galle, S. Arrais, and F. Vásconez (2015), SO2 degassing at Tungurahua volcano (Ecuador) between 2007 and 2013: Transition from continuous to episodic activity, *Journal of Volcanology and Geothermal Research*, *298*, 1–14, doi:10.1016/j.jvolgeores.2015.03.022.

Hidalgo, S., B. Bernard, P. Mothes, C. Ramos, J. Aguilar, D. Andrade, P. Samaniego, Y. Hugo, M. Hall, A. Alvarado, M. Segovia, M. Ruiz, R. Patricio, M. Vaca, and I.-E. Staff (2023), Hazard assessment and monitoring of Ecuadorian volcanoes: challenges and progresses during four decades since IG-EPN foundation, *Bulletin of Volcanology*, *86*, 1–20, doi:10.1007/s00445-023-01685-6.

Hundt, C., B. Schmidt, and E. Schömer (2014), Cuda-accelerated alignment of subsequences in streamed time series data, in *2014 43rd International Conference on Parallel Processing*, pp. 10–19, doi:10.1109/ICPP.2014.10.

Ida, Y., E. Fujita, and T. Hirose (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*, 107,616, doi:https://doi.org/10.1016/j.jvolgeores.2022.107616.

Jenkins II, W. F., P. Gerstoft, M. J. Bianco, and P. D. Bromirski (2021), Unsupervised deep clustering of seismic data: Monitoring the ross ice shelf, antarctica, *Journal of Geophysical Research: Solid Earth*, *126*(9), e2021JB021,716, doi:https://doi.org/10.1029/2021JB021716, e2021JB021716 2021JB021716.

Keramati, M., M. A. Tayebi, Z. Zohrevand, U. Glässer, J. Anzieta, and G. Williams-Jones (2023), Cubism: Co-balanced mixup for unsupervised volcano-seismic knowledge transfer, in *Machine Learning and Knowledge Discovery in Databases*, edited by M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak and G. Tsoumakas, pp. 581–597, Springer Nature Switzerland, Cham.

Kim, K., J. M. Lees, and M. Ruiz (2012), Acoustic multipole source model for volcanic explosions and inversion for source parameters, *Geophysical Journal International*, *191*(3), 1192–1204, doi:10.1111/j.1365-246X.2012.05696.x.

Kim, K., J. M. Lees, and M. C. Ruiz (2014), Source mechanism of vulcanian eruption at tungurahua volcano, ecuador, derived from seismic moment tensor inversions, *Journal of Geophysical Research: Solid Earth*, *119*(2), 1145–1164, doi:https://doi.org/10.1002/2013JB010590.

Kong, Q., A. Chiang, A. C. Aguiar, M. G. Fernández-Godino, S. C. Myers, and D. D. Lucas (2021), Deep convolutional autoencoders as generic feature extractors in seismological applications, *Artificial Intelligence in Geosciences*, *2*, 96–106, doi:https://doi.org/10.1016/j.aiig.2021.12.002.

Kovachki, N., Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar (2023), Neural operator: Learning maps between function spaces with applications to pdes, *Journal of Machine Learning Research*, *24*(89), 1–97.

Kumagai, H., H. Yepes, M. Vaca, V. Caceres, T. Naga, K. Yokoe, T. Imai, K. Miyakawa, T. Yamashina, S. Arrais, F. Vasconez, E. Pinajota, C. Cisneros, C. Ramos, M. Paredes, L. Gomezjurado, A. Garcia-Aristizabal, I. Molina, P. Ramon, M. Segovia, P. Palacios, L. Troncoso, A. Alvarado, J. Aguilar, J. Pozo, W. Enriquez, P. Mothes, M. Hall, I. Inoue, M. Nakano, and H. Inoue (2007), Enhancing volcano-monitoring capabilities in Ecuador, *Eos, Transactions American Geophysical Union*, *88*(23), 245–246, doi:10.1029/2007EO230001.

Kumar, P., and E. Foufoula-Georgiou (1997), Wavelet analysis for geophysical applications, *Reviews of Geophysics*, *35*(4), 385–412, doi:https://doi.org/10.1029/97RG00427.

Lamb, O., J. Lees, L. Franco Marin, J. Lazo, A. Rivera, M. Shore, and S. Lee (2020), Investigating potential icequakes at llaima volcano, chile, *Volcanica*, *3*(1), 29–42, doi:10.30909/vol.03.01.2942.

Lamb, O. D., J. M. Lees, L. Franco-Marin, J. Lazo, A. Rivera, M. J. Shore, and S. J. Lee (2022), Persistent shallow micro-seismicity at llaima volcano, chile, with implications for long-term monitoring, *Journal of Volcanology and Geothermal Research*, *426*, 107,528, doi:https://doi.org/10.1016/j.jvolgeores.2022.107528.

Lee, G. R., R. Gommers, F. Waselewski, K. Wohlfahrt, and O. Aaron (2019), Pywavelets: A python package for wavelet analysis, *Journal of Open Source Software*, *4*(36), 1237, doi:10.21105/joss.01237.

Li, M., X. Liu, and X. Liu (2016), Infrasound signal classification based on spectral entropy and support vector machine, *Applied Acoustics*, *113*, 116–120, doi:https://doi.org/10.1016/j.apacoust.2016.06.019.

Machacca, R., P. Lesage, H. Tavera, J. D. Pesicek, C. Caudron, J. L. Torres, N. Puma, K. Vargas, I. Lazarte, M. Rivera, and A. Burgisser (2023), The 2013–2020 seismic activity at sabancaya volcano (peru): Long lasting unrest and eruption, *Journal of Volcanology and Geothermal Research*, *435*, 107,767, doi:https://doi.org/10.1016/j.jvolgeores.2023.107767.

Mai, H., P. Audet, H. C. Perry, S. M. Mousavi, and Q. Zhang (2023), Blockly earthquake transformer: A deep learning platform for custom phase picking, *Artificial Intelligence in Geosciences*, *4*, 84–94, doi:https://doi.org/10.1016/j.aiig.2023.05.003.

Malfante, M., M. Dalla Mura, J. I. Mars, J.-P. Métaxian, O. Macedo, and A. Inza (2018a), Automatic classification of volcano seismic signatures, *Journal of Geophysical Research: Solid Earth*, *123*(12), 10,645–10,658, doi:https://doi.org/10.1029/2018JB015470.

Malfante, M., M. Dalla Mura, J.-P. Metaxian, J. I. Mars, O. Macedo, and A. Inza (2018b), Machine learning for volcano-seismic signals: Challenges and perspectives, *IEEE Signal Processing Magazine*, *35*(2), 20–30, doi:10.1109/MSP.2017.2779166.

Manley, G. F., T. A. Mather, D. M. Pyle, D. A. Clifton, M. Rodgers, G. Thompson, and J. M. Londoño (2022), A deep active learning approach to the automatic classification of volcano-seismic events, *Frontiers in Earth Science*, *10*, 1–13, doi:10.3389/feart.2022.807926.

Martínez, V. L., M. Titos, C. Benítez, G. Badi, J. A. Casas, V. H. O. Craig, and J. M. Ibáñez (2021), Advanced signal recognition methods applied to seismo-volcanic events from planchon peteroa volcanic complex: Deep neural network classifier, *Journal of South American Earth Sciences*, *107*, 103,115, doi:https://doi.org/10.1016/j.jsames.2020.103115.

Masotti, M., S. Falsaperla, H. Langer, S. Spampinato, and R. Campanini (2006), Application of support vector machine to the classification of volcanic tremor at etna, italy, *Geophysical Research Letters*, *33*(20), 1–5, doi:https://doi.org/10.1029/2006GL027441.

Matoza, R., and D. Roman (2022), One hundred years of advances in volcano seismology and acoustics, *Bulletin of Volcanology*, *84*, 1–49, doi:10.1007/s00445-022-01586-0.

Matoza, R. S., D. Fee, M. A. Garcés, J. M. Seiner, P. A. Ramón, and M. A. H. Hedlin (2009), Infrasonic jet noise from volcanic eruptions, *Geophysical Research Letters*, *36*(8), 1–5, doi:https://doi.org/10.1029/2008GL036486.

Matoza, R. S., D. Fee, and T. M. López (2014), Acoustic Characterization of Explosion Complexity at Sakurajima, Karymsky, and Tungurahua Volcanoes, *Seismological Research Letters*, *85*(6), 1187–1199, doi:10.1785/0220140110.

Matoza, R. S., A. Arciniega-Ceballos, R. W. Sanderson, G. Mendo-Pérez, A. Rosado-Fuentes, and B. A. Chouet (2019), High-broadband seismoacoustic signature of vulcanian explosions at popocatépetl volcano, mexico, *Geophysical Research Letters*, *46*(1), 148–157, doi:https://doi.org/10.1029/2018GL080802.

Mothes, P. A., H. A. Yepes, M. L. Hall, P. A. Ramón, A. L. Steele, and M. C. Ruiz (2015), The scientific-community interface over the fifteen-year eruptive episode of Tungurahua Volcano, Ecuador, *Journal of Applied Volcanology*, *4*(1), 1–15, doi:10.1186/s13617-015-0025-y.

Mousavi, M., W. Ellsworth, W. Zhu, L. Chuang, and G. Beroza (2020), Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nature Communications*, *11*, 1–12, doi:10.1038/s41467-020-17591-w.

Mousavi, S. M., and G. C. Beroza (2023), Machine learning in earthquake seismology, *Annual Review of Earth and Planetary Sciences*, *51*(1), 105–129, doi:10.1146/annurev-earth-071822-100323.

Okamoto, K., Y. Mukuhira, D. Darisma, H. Asanuma, and H. Moriya (2024), Machine learning automatic picker for geothermal microseismicity analysis for practical procedure to reveal fine reservoir structures, *Geothermics*, *116*, 102,832, doi:https://doi.org/10.1016/j.geothermics.2023.102832.

Ortiz, H. D., J. B. Johnson, P. G. Ramón, and M. C. Ruiz (2018), Using infrasound waves to monitor tropospheric weather and crater morphology changes at volcán tungurahua, ecuador, *Journal of Volcanology and Geothermal Research*, *349*, 205–216, doi:https://doi.org/10.1016/j.jvolgeores.2017.11.001.

Ortiz, H. D., R. S. Matoza, J. B. Johnson, S. Hernandez, J. C. Anzieta, and M. C. Ruiz (2021a), Autocorrelation infrasound interferometry, *Journal of Geophysical Research: Solid Earth*, *126*(4), e2020JB020,513, doi:https://doi.org/10.1029/2020JB020513, e2020JB020513 2020JB020513.

Ortiz, H. D., R. S. Matoza, C. Garapaty, K. Rose, P. Ramón, and M. C. Ruiz (2021b), Multi-year regional infrasound detection of Tungurahua, El Reventador, and Sangay volcanoes in Ecuador from 2006 to 2013, *Proceedings of Meetings on Acoustics*, *41*(1), 022,003, doi:10.1121/2.0001362.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, *12*, 2825–2830.

Ren, C. X., A. Peltier, V. Ferrazzini, B. Rouet-Leduc, P. A. Johnson, and F. Brenguier (2020), Machine learning reveals the seismic signature of eruptive behavior at piton de la fournaise volcano, *Geophysical Research Letters*, *47*(3), e2019GL085,523, doi:https://doi.org/10.1029/2019GL085523, e2019GL085523 2019GL085523.

Reyes, J. A., and C. J. Jiménez Mosquera (2017), Non-supervised classification of volcanic-seismic events for tungurahua-volcano ecuador, in *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–6, doi:10.1109/ETCM.2017.8247446.

Ruiz, M. C., J. Lees, and J. Johnson (2006), Source constraints of Tungurahua volcano explosion events, *Bulletin of Volcanology*, *68*, 480–490, doi:10.1007/s00445-005-0023-8.

Sun, H., Z. E. Ross, W. Zhu, and K. Azizzadenesheli (2023), Phase neural operator for multi-station picking of seismic arrivals, *Geophysical Research Letters*, *50*(24), e2023GL106,434, doi:https://doi.org/10.1029/2023GL106434, e2023GL106434 2023GL106434.

Tan, D., D. Fee, A. J. Hotovec-Ellis, J. D. Pesicek, M. M. Haney, J. A. Power, and T. Girona (2023), Volcanic earthquake catalog enhancement using integrated detection, matched-filtering, and relocation tools, *Frontiers in Earth Science*, *11*, 1–16, doi:10.3389/feart.2023.1158442.

Tang, L., M. Zhang, and L. Wen (2020), Support vector machine classification of seismic events in the tianshan orogenic belt, *Journal of Geophysical Research: Solid Earth*, *125*(1), e2019JB018,132, doi:https://doi.org/10.1029/2019JB018132, e2019JB018132 2019JB018132.

Tavenard, R., J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods (2020), Tslearn, a machine learning toolkit for time series data, *Journal of Machine Learning Research*, *21*(118), 1–6.

Titos, M., A. Bueno, L. García, and C. Benítez (2018), A deep neural networks approach to automatic recognition systems for volcano-seismic events, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(5), 1533–1544, doi:10.1109/JSTARS.2018.2803198.

Titos, M., A. Bueno, L. García, M. C. Benítez, and J. Ibañez (2019), Detection and classification of continuous volcano-seismic signals with recurrent neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 1936–1948, doi:10.1109/TGRS.2018.2870202.

Titos, M., A. Bueno, L. García, C. Benítez, and J. C. Segura (2020), Classification of isolated volcano-seismic events based on inductive transfer learning, *IEEE Geoscience and Remote Sensing Letters*, *17*(5), 869–873, doi:10.1109/LGRS.2019.2931063.

Toney, L., D. Fee, A. Witsil, and R. S. Matoza (2022), Waveform Features Strongly Control Subcrater Classification Performance for a Large, Labeled Volcano Infrasound Dataset, *The Seismic Record*, *2*(3), 167–175, doi:10.1785/0320220019.

Trani, L., G. A. Pagani, J. P. P. Zanetti, C. Chapeland, and L. Evers (2022), Deepquake — an application of cnn for seismo-acoustic event classification in the netherlands, *Computers & Geosciences*, *159*, 104,980, doi:https://doi.org/10.1016/j.cageo.2021.104980.

Varley, N., J. Johnson, M. Ruiz, G. Reyes, and K. Martin (2006), Applying statistical analysis to understanding the dynamics of volcanic explosions, in *Statistics in Volcanology*, Geological Society of London, doi:10.1144/IAVCEI001.6.

Wannesm, Khendrickx, A. Yurtman, P. Robberechts, D. Vohl, E. Ma, G. Verbruggen, M. Rossi, M. Shaikh, M. Yasirroni, Todd, W. Zieliński, T. V. Craenendonck, and S. Wu (2022), wannesm/dtaidistance: v2.3.5, doi:10.5281/zenodo.5901139.

Watson, L. M. (2020), Using unsupervised machine learning to identify changes in eruptive behavior at mount etna, italy, *Journal of Volcanology and Geothermal Research*, *405*, 107,042, doi:https://doi.org/10.1016/j.jvolgeores.2020.107042.

Wickham-Piotrowski, A., Y. Font, M. Regnier, B. Delouis, O. Lengliné, M. Segovia, and Q. Bletery (2023), Achieving a Comprehensive Microseismicity Catalog through a Deep-Learning-Based Workflow: Applications in the Central Ecuadorian Subduction Zone, *Bulletin of the Seismological Society of America*, *XX*(XX), 1–19, doi:10.1785/0120230128.

Witsil, A. J., and J. B. Johnson (2020), Analyzing continuous infrasound from stromboli volcano, italy using unsupervised machine learning, *Computers & Geosciences*, *140*, 104,494, doi:https://doi.org/10.1016/j.cageo.2020.104494.

Wright, H. M. N., K. V. Cashman, P. A. Mothes, M. L. Hall, A. G. Ruiz, and J.-L. Le Pennec (2012), Estimating rates of decompression from textures of erupted ash particles produced by 1999-2006 eruptions of Tungurahua volcano, Ecuador, *Geology*, *40*(7), 619–622, doi:10.1130/G32948.1.

Yukutake, Y., A. Kim, and T. Ohminato (2023), Reappraisal of volcanic seismicity at the Kirishima volcano using machine learning, *Earth, Planets and Space*, *75*, 1–17, doi:10.1186/s40623-023-01939-9.

Zali, Z., M. Mousavi, M. Ohrnberger, E. P. Eibl, and F. Cotton (2024), Tremor clustering reveals pre-eruptive signals and evolution of the 2021 Geldingadalir eruption of the Fagradalsfjall Fires, Iceland, *Communications Earth & Environment*, *5*, 1–11, doi:10.1038/s43247-023-01166-w.

Zheng, J., S. Shen, T. Jiang, and W. Zhu (2019), Deep neural networks design and analysis for automatic phase pickers from three-component microseismic recordings, *Geophysical Journal International*, *220*(1), 323–334, doi:10.1093/gji/ggz441.

Zhu, W., and G. C. Beroza (2018), PhaseNet: a deep-neural-network-based seismic arrival-time picking method, *Geophysical Journal International*, *216*(1), 261–273, doi:10.1093/gji/ggy423.

Zhu, W., S. M. Mousavi, and G. C. Beroza (2019), Seismic signal denoising and decomposition using deep neural networks, *IEEE Transactions on Geoscience and Remote Sensing*, *57*(11), 9476–9488, doi:10.1109/TGRS.2019.2926772.

Chapter 4

Tracking changes in natural streams using acoustic data: The 2021 Western North America heat wave at the Squamish River, BC, Canada

4.1 Introduction

The number of scientific efforts focusing on the effects of climate change are increasing with urgency, especially since escalating extreme weather events result in large environmental, societal and economic impacts (*Naveau et al.*, 2020; *Krichen et al.*, 2024). This is reflected in the steady increase of literature containing the phrases "climate change" and "extreme weather events" since the mid-1980s (Figure 4.1). Most of these efforts are devoted to trying to mitigate the effects of natural hazards by improving the understanding or modelling the predictability of these phenomena (*Finkel et al.*, 2023), or by improving monitoring capabilities with continuously evolving classical and unconventional technological tools (*Krichen et al.*, 2024). In spite of all the combined efforts, natural hazards monitoring remains challenging for many reasons including economic and logistical challenges of installing and maintaining instruments in areas of interest, and technical complexities of analyzing and interpreting the derived data in a timely manner (*Krichen et al.*, 2024).

Among the many hazards, there is great interest in monitoring those related to water streams such as rivers and their flows. These hazards can occur in two contrasting scenarios:

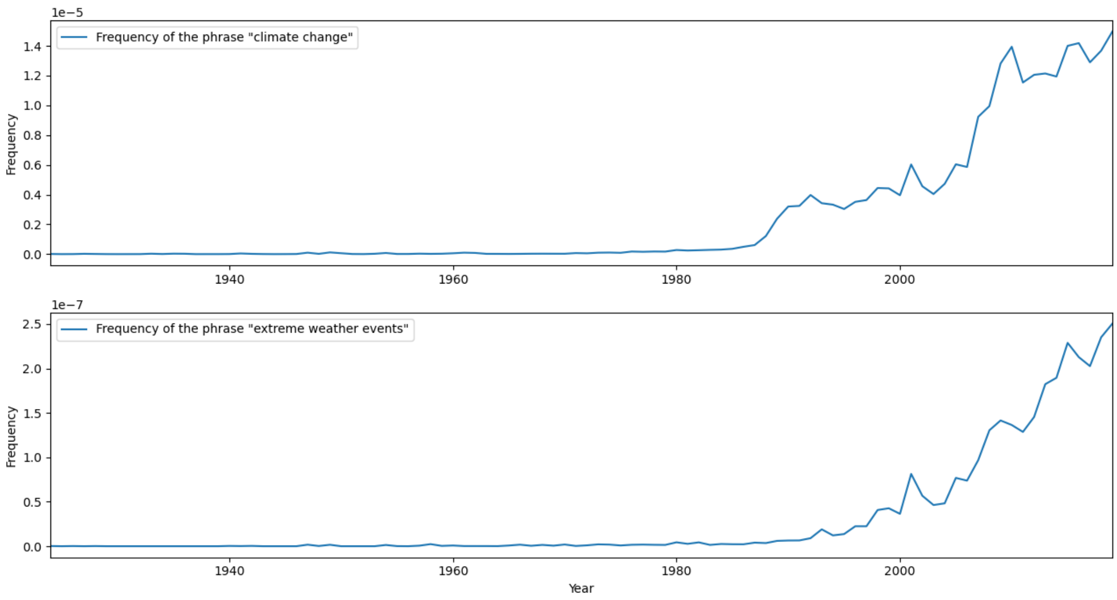

Figure 4.1: Time evolution of the appearance of the phrases "climate change" and "extreme weather events" in publications since 1924 (*Michel et al.*, 2011).

low discharges that can lead to droughts (*Seybold et al.*, 2023); and high flows that can lead to flooding (*Arnell and Gosling*, 2016; *Ward et al.*, 2018) among others. Similarly, extreme weather and climate leads to an increase in the occurrence of mass movement events such as landslides and debris-flows (*Chiarle et al.*, 2021).

British Columbia is the westernmost province of Canada and its climate is significantly influenced by the Pacific Ocean as well as by the belt of mountains on the western margin of the Americas. As such, it has regions that exhibit some of the wettest and driest climates in Canada. A notable feature from the mountains of the province is that they possess roughly between 13,000 and 15,000 glaciers (*Clague et al.*, 2011; *Pfeffer et al.*, 2014; *Bevington and Menounos*, 2022) which contribute to the flow in streams and rivers especially in summer, as these glaciers melt at a heightened rate. The rugged geology of the region is due in part to a combination of extensive volcanic activity and repeated periods of continental glaciation, making the area prone to mass movement events of great magnitude.

On November 2021, British Columbia, Canada, experienced one of its most devastating historical extreme weather event, when the intense rain from an atmospheric river generated floods and landslides in the southwest of the province; the severity of these hazards is bound to worsen with time (*Gillett et al.*, 2022).

Similarly, although less catastrophic, are continuous debris flow events occurring on the flanks of Mount Cayley (Figure 4.2a), that have caused repeated damage to the Squamish River Forest Service Road used for logging and for accessing various recreational activities in the region. At least 4 such events have occurred in recorded history, one presumably in July 1963 (*Clague and Souther*, 1982), another in June 1984 (*Cruden and Lu*, 1992), and a third during the night of June 28^{th}, 2014 that wiped out the road and stranded about 300 campers from the nearest town of Squamish. The most recent event occurred on the night of October 16^{th}, 2023, and stranded logging personnel that managed to walk across Mud Creek to return to town.

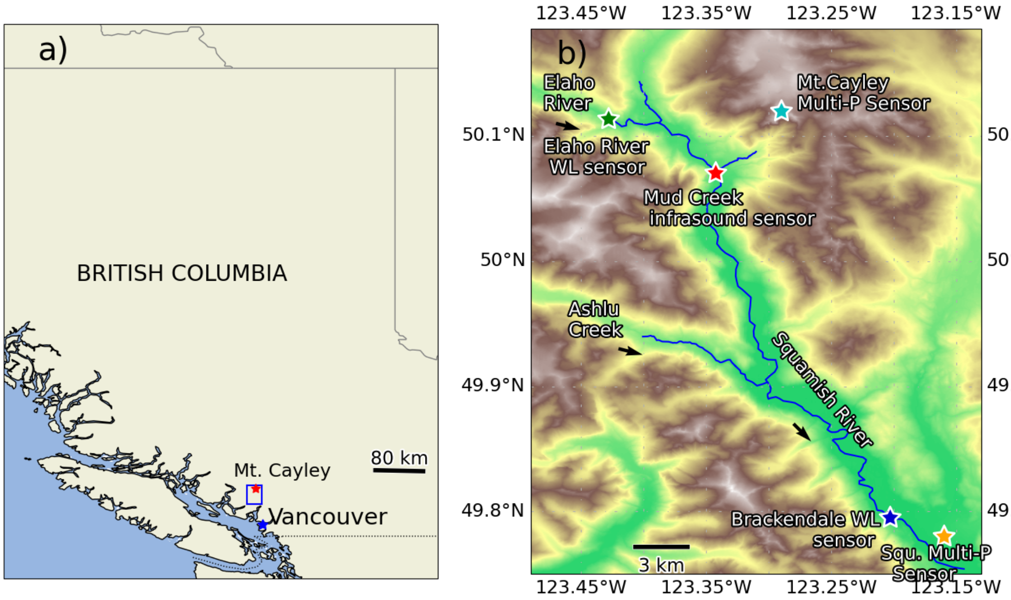

Figure 4.2: a) Location of Mount Cayley (red star), Vancouver (blue star), and region of interest (blue square), B.C., Canada. b) Zoomed region of interest. Stars show locations of instrumentation used in this study. Black arrows show direction of flow.

Improving decision-making and emergency response to increasing hazards such as these requires a constant increase of effective monitoring capabilities (*Krichen et al.*, 2024) although the effectiveness of any strategy must be compared with a baseline knowledge of the setting of interest (*Kumar et al.*, 2021).

Given that rivers are a primary source of freshwater essential for societal sustainability, monitoring streamflow is a widely practiced and sought-after activity globally, in addition to its significance in mitigating hazards and understanding ecological implications (*Dobriyal et al.*, 2017; *Zhu et al.*, 2021). However, the collection of historical hydrological data and installation of monitoring stations remain as challenging tasks for the creation of flood early warning systems (*Perera et al.*, 2020). While there are many ways to perform streamflow measurements and no unique classification of methods exists (e.g., *Dobriyal et al.*, 2017; *Iukhno et al.*, 2021; *Llaban and Ella*, 2022), the methods are generally categorized based on their application principles: direct methods, hydraulic structure methods, non-contact methods, and sensor-based methods (*Llaban and Ella*, 2022). These methods have different features in terms of cost, operational and implementation efficiency, accuracy, time effectiveness, and ecological significance and can go from classical current meters that use rotors inside the stream, to indirect particle image velocimetry using cameras on UAVs or radar and ultrasonic-based sensors (*Dobriyal et al.*, 2017; *Llaban and Ella*, 2022). Nevertheless, given that there is no universally applicable method for measuring streamflows across different settings, ongoing research is dedicated to developing, refining, and expanding existing methods, as well as exploring new approaches for use in remote areas (e.g., *Hundt and Blasch*, 2019; *Zhao et al.*, 2019; *Zhu et al.*, 2021). Several state of the art water discharge sensors require proper and sometimes complex installation conditions (*Dobriyal et al.*, 2017; *Llaban and Ella*, 2022), and as an example, a typical hydrometric station for the Regional District of Nanaimo, British Columbia costs around 15000-20000 CAD for installation and $10000-$20000 CAD for ongoing (yearly) maintenance (*Sutherland*, 2015).

The continuous research and development of sensors has allowed for the appearance of cheaper and more portable instruments, in particular acoustic ones (*Anderson et al.*, 2017). Acoustic sensors are becoming ubiquitous as a geophysical tool because of their versatility to detect of a wide range of phenomena that include natural explosions from active volcanoes or man-made ones from mining, debris-flows, rockfalls and avalanches at the slopes of mountains and volcanoes, and streamflows from rivers and dams, among others (*Allstadt et al.*, 2018; *Anderson et al.*, 2019; *Dannemann Dugick et al.*, 2023). Acoustic sensors, either combined with seismic sensors (*Belli et al.*, 2022) or used in multiple-sensor arrays (e.g., *Marchetti et al.*, 2019; *Marshall et al.*, 2019; *Johnson et al.*, 2021; *Sanderson et al.*, 2021; *Marchetti and Johnson*, 2023), have proven useful to detect mass-movement events such as lahars, avalanches and debris-flows. Although single infrasound sensors are less effective for monitoring purposes (*Marchetti and Johnson*, 2023), they can still be useful for site characterization before, during and after debris flows for the creation of early warning systems (e.g.,*Kogelnig et al.*, 2014; *Liu et al.*, 2015) as well as provide information about changes in the atmosphere (e.g., *Ortiz et al.*, 2021). Due to this versatility, in this study we use single-sensor acoustic data to develop a preliminary framework in British Columbia. In order to monitor continuous fluid movement and for future potential hazard monitoring two tasks are pursued simultaneously: 1) Establish an effective discharge rate monitoring setup at a specific region near the Squamish River to define a baseline acoustic behaviour of the area. 2) Use the inferred baseline information from the local discharge rate to detect potential deviations/anomalies to support the development of future early warning systems. These tasks are derived using recorded low-frequency ($\leq 50Hz$) acoustic signals from before, during and after the extraordinary 2021 Western North America heat wave. More specifically, due to Mud Creek's historical record of persistent and recent damaging debris flows, and at the request of the BC Ministry of Forests, we focus our study at the confluence point of Mud Creek and Squamish River on the western flank of Mount Cayley (Figure 4.2b). Additionally, third party regional weather parameters are compared with the acoustic recordings to define a baseline for the acoustic behaviour of different phenomena in the area, and track their time evolution for the duration of the study period (March 13th to July 9th, 2021).

4.2 Datasets and Methods

4.2.1 Datasets

To track stream variations and potential mass movement events, we installed a temporary low-frequency acoustic sensor at the streams' convergence spot at a distance of a few tens of meters from both Squamish River and Mud Creek (Figure 4.3a). We used one of the relatively newly developed "Gem Infrasound Logger" (Gem) low-frequency acoustic sensors by *Anderson et al.* (2017) because of their portability, autonomy and small size that made them ideal for rapid installation, concealment, and operation in such a remote area (Figure 4.3c). These sensors have already shown their prowess in diverse extreme settings such as active volcanoes (*Anderson et al.*, 2017; *Rosenblatt et al.*, 2022) and weather balloon experiments (*Bowman and Albert*, 2018; *Brissaud et al.*, 2021), but also in less complicated regions for water discharge characterization such as in *Tatum et al.* (2023). The Gems are differential pressure transducer-based sensors that record 100 samples per second and have a flat response in the $0.039 - 27.1 Hz$ frequency band making them suitable for detecting a wide range of phenomena. The sensor was installed near the border of a small forest (Figure 4.3b) to ensure good quality low-noise data using the trees to reduce the effect from strong winds (*Tatum et al.*, 2023).

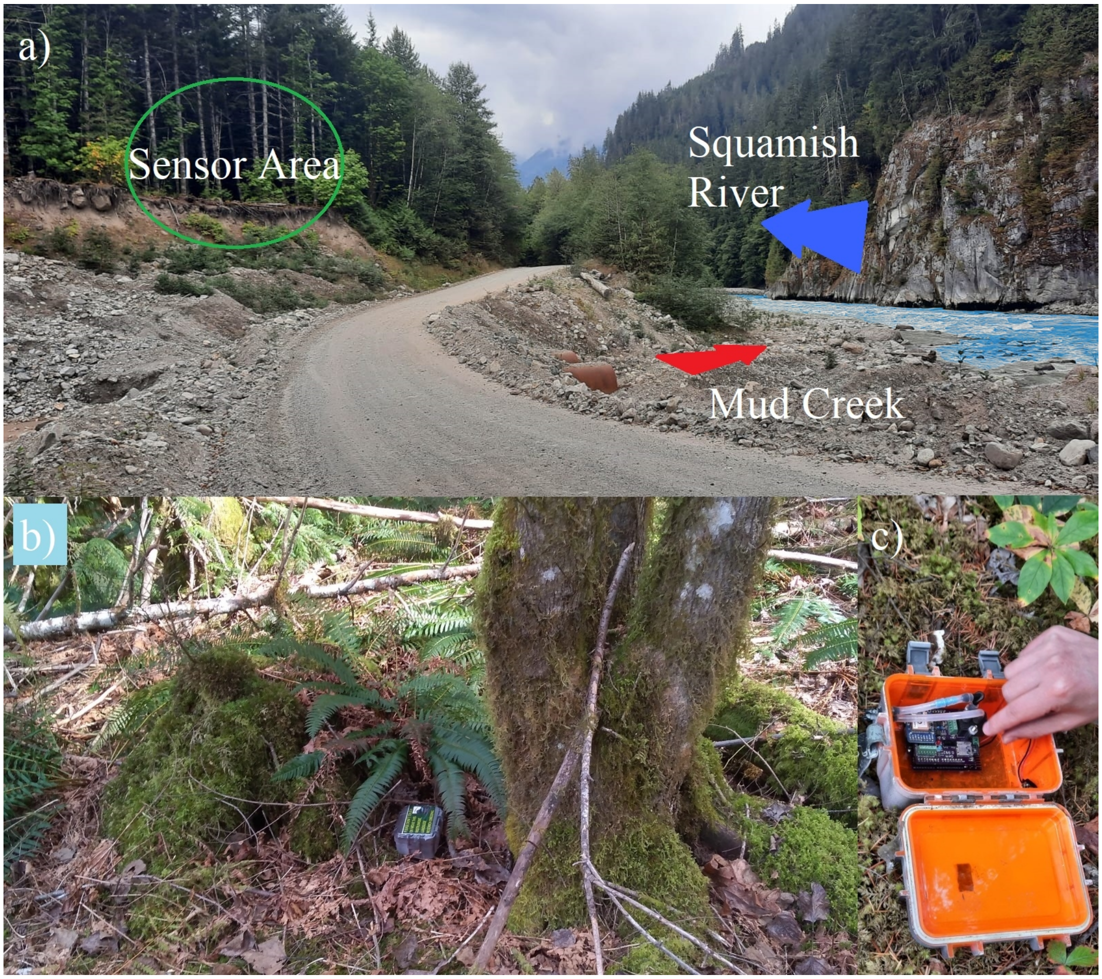

Figure 4.3: Collage of images showing the installation setting of the Gem acoustic sensor. See Figure 4.2 for site location.

The rift in this confluence point has rocks of various sizes (some probably carried from past debris flows down Mud Creek), and this forms irregularities in the channel that produce diverse simultaneous turbulent currents generating noise year-round (Figure 4.4).

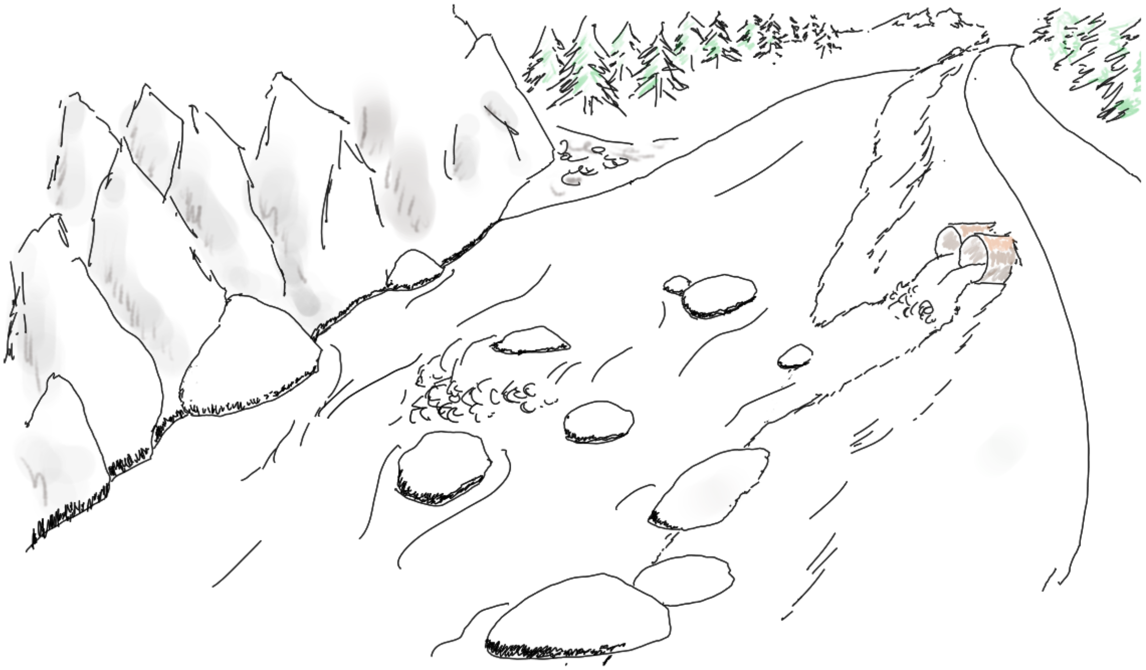

Figure 4.4: Depiction of the area and some of the Roughness Elements (RE) that can create noise related to discharge rates.

In order to both associate and discriminate the information recorded with the low-frequency acoustic sensor to local and regional signals, we also gathered several environmental parameters from weather stations and water level sensors in proximity (within $\sim 40\,km$) of the junction (Figure 4.2b, red star). The main water level sensor collects the information from Squamish River relatively near Brackendale, close to the Squamish Airport weather station and is $39\,km$ downstream (Figure 4.2b, dark blue star) from the installed acoustic sensor. The secondary water level sensor is located upstream on the Elaho River, one of the major tributaries, and $9.5\,km$ northwest of the sensor location (Figure 4.2b, green star). The discharge rates are derived empirically from continuous water level measurements. A WSC hydrometric technologist takes direct measurements of discharge and water level by deploying special instruments, from a bridge, by wading in a stream, by boat, or using a cable-way strung across a river several times a year to stablish stage-discharge curves. Both values are made publicly available in real-time every 5 minutes for up to 18 months retrospectively on the Environment and Climate Change Canada Real-time Hydrometric Data web site (*ECCC-Real-time Hydrometric Data web site*, 2024). One weather station sits at the top of Mount Cayley (Figure 4.2b, light blue star), provides data from higher atmospheric phenomena, and is relatively close to the sensor, with the data provided by the BC Ministry of Forests. The other weather station at the Squamish Airport (Figure 4.2b, orange star) is $\sim 40\,km$ to the South-Southeast of the station and provides weather parameters from lower regions of the atmosphere; its data is also publicly available at *ECCC-Historical Climate Data web site* (2024). The weather stations provide data on an hourly basis and, along with the water discharges, were collected during the full study period (Figure 4.5).

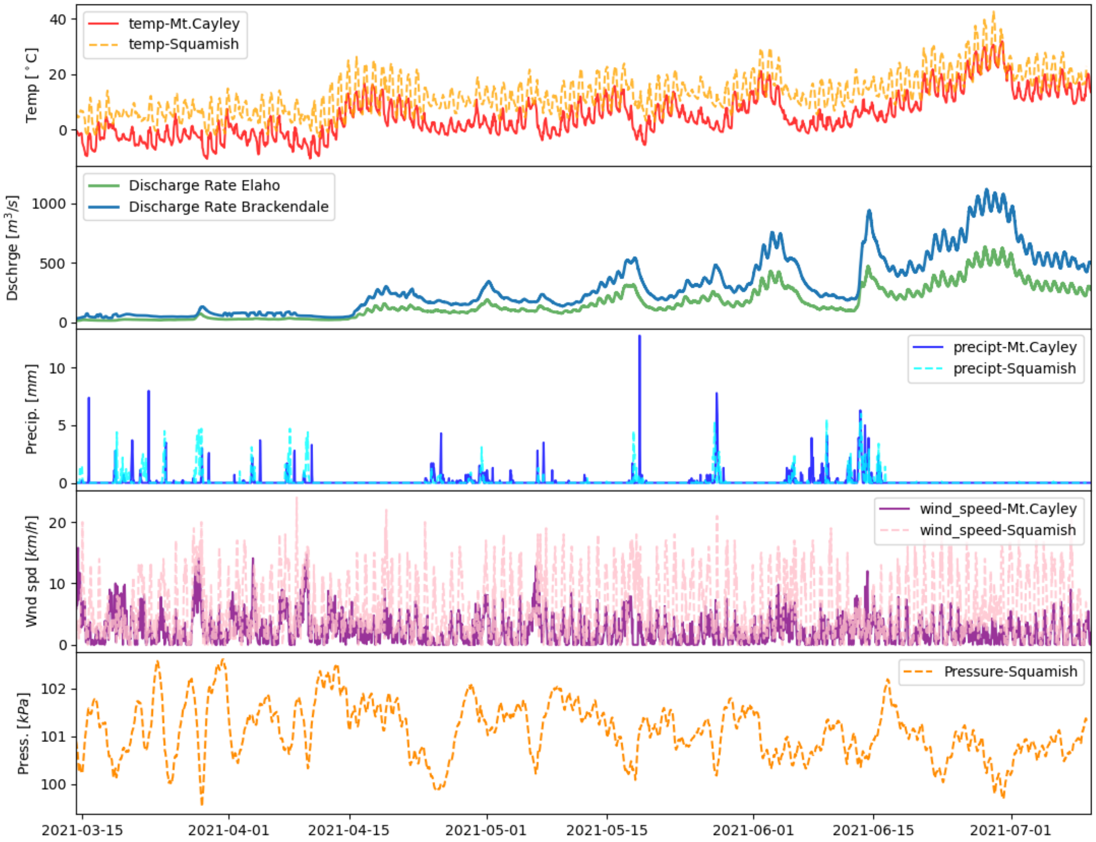

Figure 4.5: Weather parameters in the area of interest.

Diurnal cycles due to sun power are clearly visible in the air temperature and wind speed time series (Figure 4.5). Discharge rates at Squamish River are highly seasonal and driven by precipitation and snowmelt patterns: 1) discharge rates are low from January to March;

2) Freshet flows are large and sustained during late spring through summer due to snowmelt; and 3) There are high but sporadic peak discharges due to rain on snow storms during fall (*Fath*, 2014). Thus, there is a high correlation between air temperature and water discharge rate at both rivers, illustrating the influence of high temperatures on the streams in the region, specially during the heatwave that produced large and fast temperature changes. There were also periods when significant flows did not fully correlate with temperature, specifically around March 28, and after June 15th. Those flows are most likely related to regional parameters other than temperature because they appear in the Squamish as well as the Elaho water discharge measurements.

The daily wind cycles generate strong noise in the raw acoustic recordings (Figure 4.6). However, since the Squamish River is a persistent, identifiable sound producer in the Gem sensor's area, we focus on the discharge rate datasets from the Squamish River as a main control, and from the Elaho River as a secondary one, to create a strong local baseline dataset. Acoustic measurements will be compared to those discharge rates from hereon, and different environmental phenomena will be appraised with respect to this information when relevant.

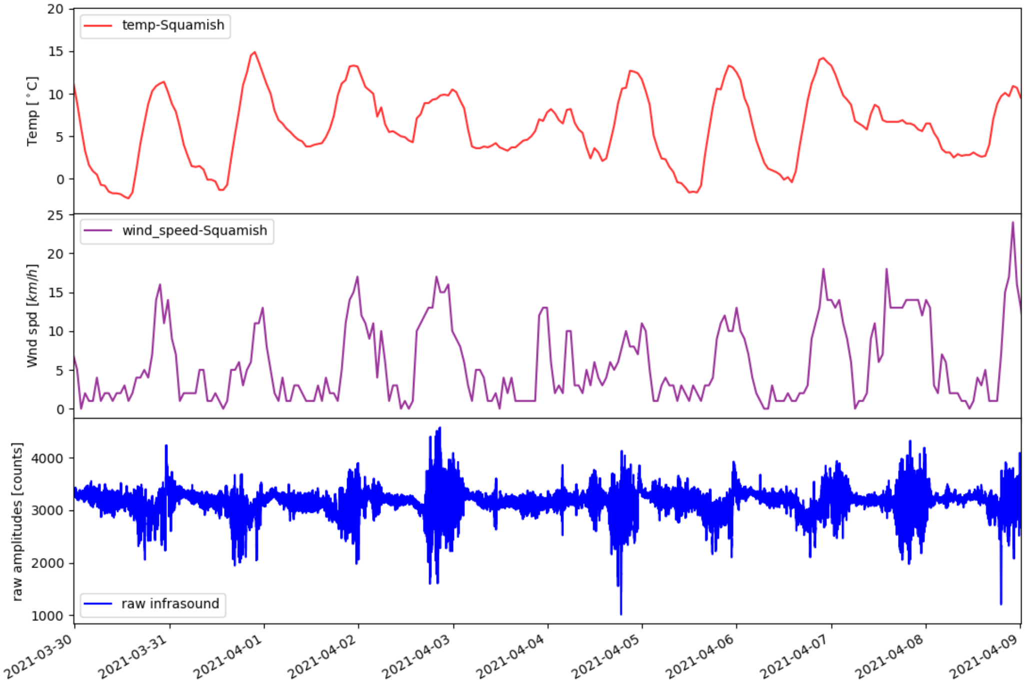

Figure 4.6: Temperature, Wind and Infrasound series.

4.2.2 Methods - Acoustic Streamflow Monitoring

A characterization of the water discharge rate from the Squamish River at the junction with Mud Creek is made by relating it to local noise levels captured from the Gem Infrasound Logger, based on approaches in the recent work by *Ronan et al.* (2017), *Anderson et al.* (2019), *Osborne et al.* (2021) and *Tatum et al.* (2023). Note that it is important not to confuse this *atmospheric* acoustic method with the traditional "acoustic velocity meters" and "acoustic doppler current profiler" methods which are *underwater* acoustic techniques that probe velocities inside the stream (*Dobriyal et al.*, 2017; *Kästner et al.*, 2018; *Zhu et al.*, 2021), nor with the ultrasonic sensor method that work outside of the stream like reflected-beam water gauges (*Llaban and Ella*, 2022).

Ronan et al. (2017) investigated many different types of waves and their sounds in a controlled setting, and linked them to different Froude numbers (Fr) that allow distinguishing between subcritical ($Fr < 1$) and supercritical ($Fr > 1$) flows. Froude numbers are useful because they can be used to investigate and infer various flow parameters under certain assumptions, and settings that extend beyond rivers (e.g., *Le Moigne et al.*, 2020). The Froude number is defined in Equation 4.1 as:

$$Fr = \frac{v}{\sqrt{gh}} \tag{4.1}$$

with v the velocity of the flow, $g = 9.8m/s^2$ the acceleration of gravity, and h the depth of the flow. Additionally, *Ronan et al.* (2017) showed that the different flow jumps show different average spectral power for specific frequency bands. *Anderson et al.* (2019) noted a correspondence of acoustic amplitude with discharges on a seasonal flood cycle, a flash flood, and a regulated dam spill-over was hinted. On the other hand, *Tatum et al.* (2023) performed another controlled study in the same area as *Ronan et al.* (2017), and demonstrated that acoustic power and its relation to wave configuration is consistent within a given year, but not over the years, and that it does not necessarily correlate solely to discharge rate. As such wave morphology is a crucial factor to take into account in studies linking low-frequency sound power and discharge rates. However, as opposed to the experiments of *Ronan et al.* (2017) and *Tatum et al.* (2023), our setting possesses varied obstacles or *roughness elements* (RE) that can be expected to cause different types of flows almost irrespective of discharge rate (Figure 4.4). Finally, *Osborne et al.* (2021) used microphones recording in an audible frequency band ($\geq 50Hz - 20kHz$) to show that river *stage* can be estimated from sound pressure when appropriate frequency bands are selected and signals are properly filtered.

In this study, it was not possible to apply the Froude equation because both local river velocity and depth are unknowns at the confluence spot. Nevertheless, in order to develop a method that could be potentially implemented locally to enable future hazard warnings in a timely and robust manner, we combined different aspects from previous studies (*Ronan et al.*, 2017; *Anderson et al.*, 2019; *Osborne et al.*, 2021; *Tatum et al.*, 2023) into a simplified procedure.

Instead of working in the frequency domain (e.g., *Osborne et al.*, 2021) or calculating acoustic power (e.g., *Tatum et al.*, 2023), we directly operated on the recorded amplitudes in the time domain (as suggested in *Anderson et al.*, 2019) and then obtained data statistics (similar to *Osborne et al.*, 2021) with minor signal pre-processing. First and foremost, as described in *Anderson et al.* (2023), software is provided to pre-process the data acquired with the Gems into more standard readable files. We then used the Obspy package for Python (*Beyreuther et al.*, 2010) and standard libraries to perform subsequent analyses. After this initial step, data streams were divided into non-overlapping windows of 150 seconds. The time windows were detrended and filtered on an appropriate frequency band to be described. Finally, the mean, standard deviation, and different percentiles of the absolute amplitudes of the windowed signals were calculated and explored as proxies for river discharge rate.

The filtering frequency band is not arbitrary; following the guidelines and past experiences from the mentioned studies, a frequency inspection was performed which found that wide frequency bands can become excessively excited by specific flow regimes or unimportant noise (like that produced by wind). For that reason, the effect of changes in discharge rates was investigated across many different frequency bands, and found that good correlation occurs at specific bands that are most likely site-dependent. For instance, in our location a frequency band of $3 - 8Hz$ was determined to be suitable for tracking discharge changes with relatively low impact from other sources. Conversely, by using the open dataset from *Tatum et al.* (2023) to validate this method, a frequency band from $40 - 50Hz$ was found to provide good results. The specific band for a region can be found empirically by filtering and trying many different frequency values, and comparing them to water discharge rates, or can also be found by performing a spectral analysis of the setting and searching for an adequate band in a similar fashion as suggested by *Osborne et al.* (2021). The simplicity of this approach permits this exploration to be fast.

Comparison of the proposed method with another study

By filtering the open data-sets from *Tatum et al.* (2023) on the appropriate band ($40-50Hz$) and using their open code, it is shown how the tracking capabilities from the method proposed here (Figures 4.7c and 4.9c) compares to the processing scheme by *Tatum et al.* (2023) (Figures 4.7b and 4.9b) for the water discharge rate (Figures 4.7a and 4.9a). Additionally, Figures 4.8 and 4.10 show that by using the same frequency band as we proposed ($40 - 50Hz$), the results of applying their processing scheme is nearly identical to ours (Figures 4.7c and 4.9c), albeit with more computational burden. Our results display the filtered amplitudes for 1 hour long windows between 9-10 AM as in *Tatum et al.* (2023) to be compared directly, as well as for 2.5 minutes windows in those times.

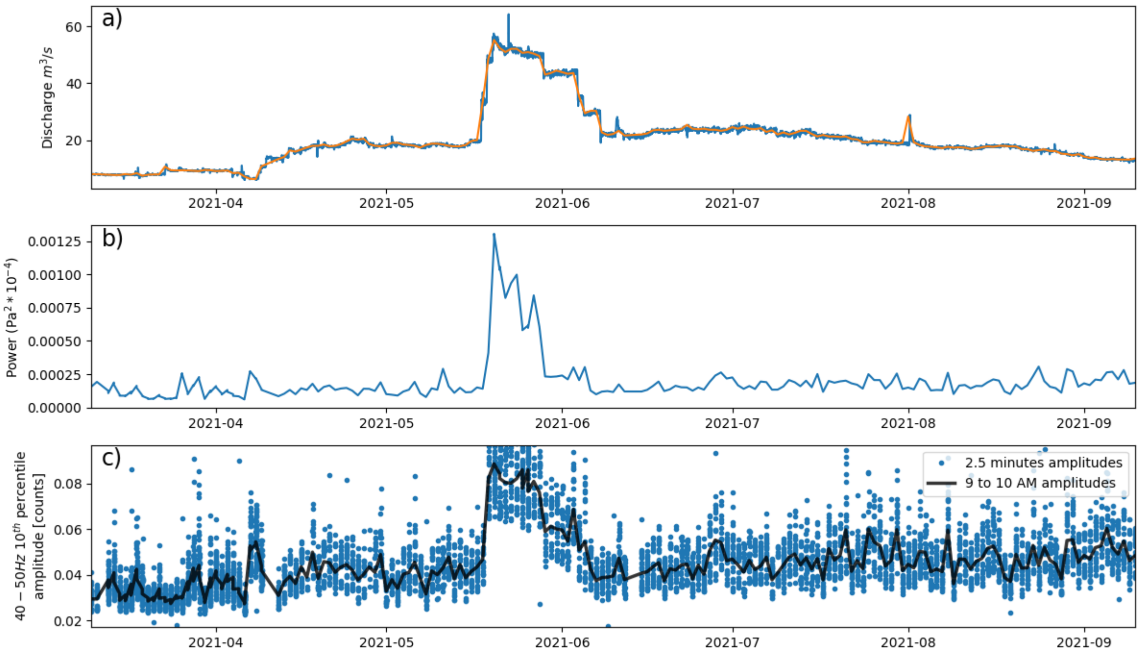

Figure 4.7: a) 2021 Discharge Rate. b) Acoustic Power by *Tatum et al.* (2023). c) Proposed acoustic amplitude at different time window sizes (1-hour and 2.5 minutes long).

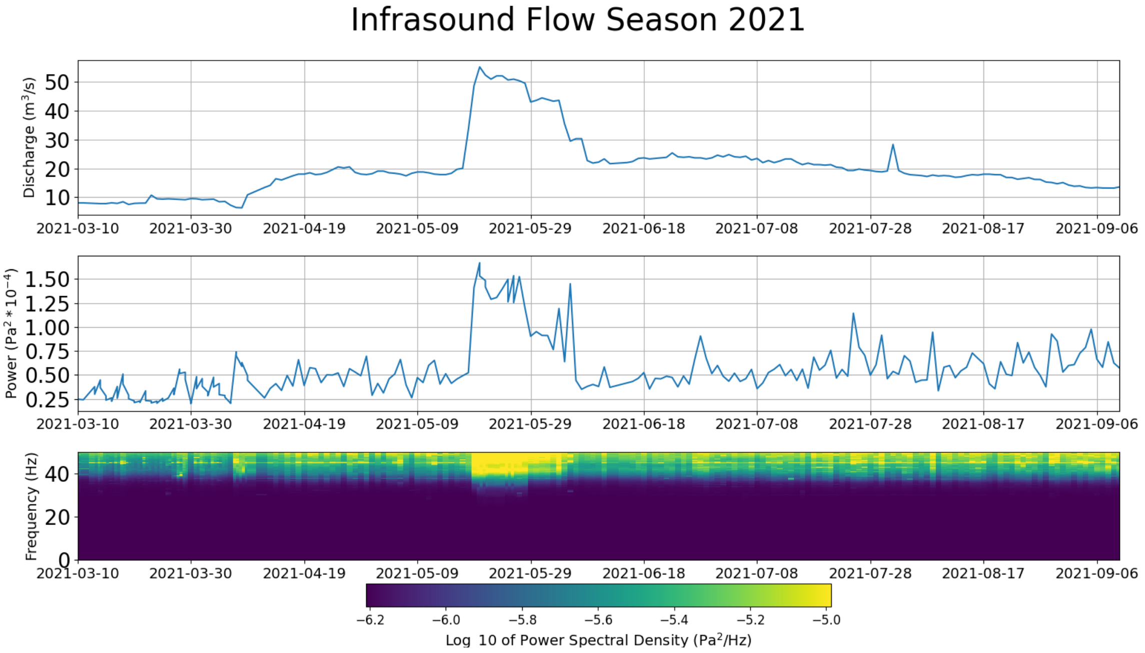

Figure 4.8: Comparison of the 2021 discharge rate with the acoustic power derived as in *Tatum et al.* (2023) restricted to our proposed frequency band $40 - 50Hz$.

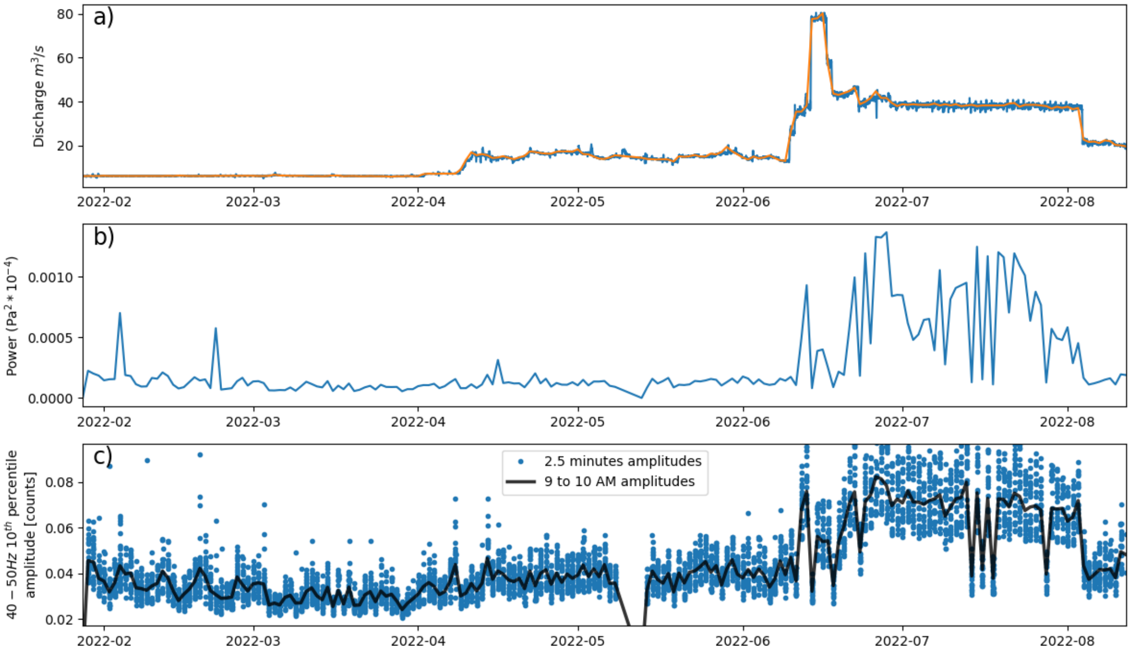

Figure 4.9: a) 2022 Discharge Rate. b) Acoustic Power by *Tatum et al.* (2023). c) Proposed acoustic amplitude at different time window sizes (1-hour and 2.5 minutes long).

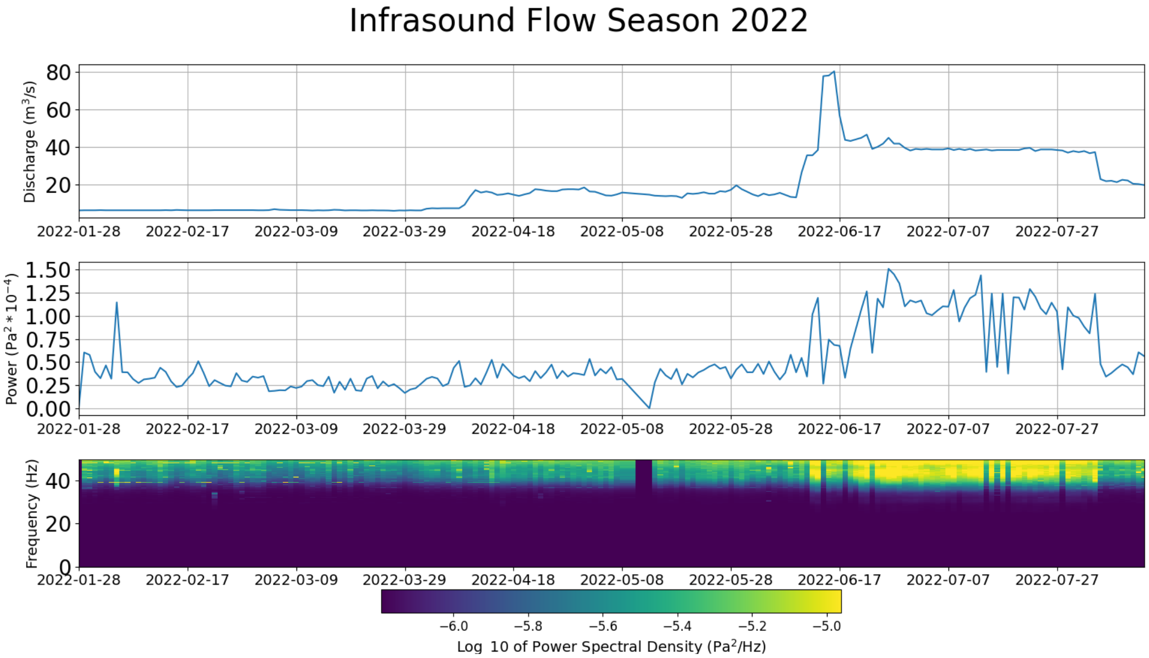

Figure 4.10: Comparison of the 2022 discharge rate with the acoustic power derived as in *Tatum et al.* (2023) restricted to our proposed frequency band $40 - 50Hz$.

4.3 Results

Out of the many statistics derived from the processed acoustic amplitudes analysis, we found that the 10^{th} percentile was the most stable, and show that particular time series for results hereon (a comparison of the derived 10^{th} percentile and other statistics is shown in Appendix D.1). We emphasize that further processing such as instrumental response deconvolution was not applied with an on-site early warning application in mind. Also, because of greater correlation, discharge rates were used instead of water level/stage for all subsequent analysis (a comparison between discharge rates and stages is shown in Appendix D.2). To showcase the effect of the different discharge rates on the frequency content of the signals, semi-diurnal spectrograms filtered above $3Hz$ were calculated from the pre-processed only signals (to observe diurnal trends hinted by parameters shown in Figure 4.6). These time series (processed acoustic amplitudes and spectrograms) were calculated for the whole period and compared to water discharges from the region's sensors described before.

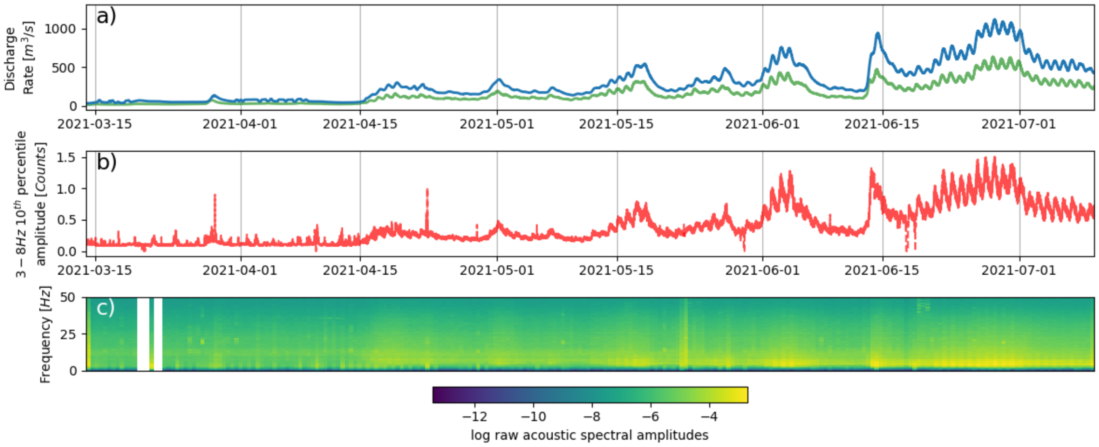

Figure 4.11: Comparison between a) discharge rates and b) filtered amplitudes with the c) masked semi-diurnal spectrogram.

Figure 4.11 shows the full discharge rate (a), the bandpass filtered 10^{th} percentile amplitudes (b), and the spectrogram of the signals filtered above $3Hz$ (c). To visualize important features in the frequency spectrum, the spectrogram was masked for some days between March 23 and 29 because of an unusually powerful signal recorded between those days. The complete unmasked spectrogram is shown in Appendix D.3.

Detailed plots zoomed into periods of interest that displayed extraordinary water discharges that were not directly correlated with temperature increase (Figure 4.12).

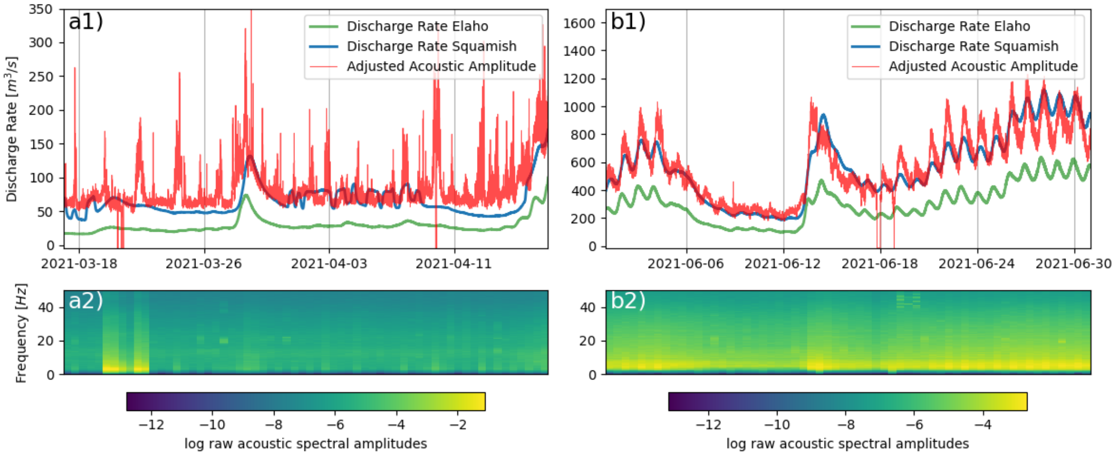

Figure 4.12: (a1,b1) Comparison between discharge rates and filtered amplitudes. (a2,b2) The corresponding unmasked spectrograms for unprocessed amplitudes.

Figure 4.12 shows discharge rate superimposed by re-scaled or adjusted acoustic amplitudes by finding a linear relationship between the 10^{th} percentile amplitudes and the discharge rate at the Squamish River. The relationship was obtained by finding the linear regression equation between the 10^{th} percentile amplitudes and the discharge (Figure 4.13).

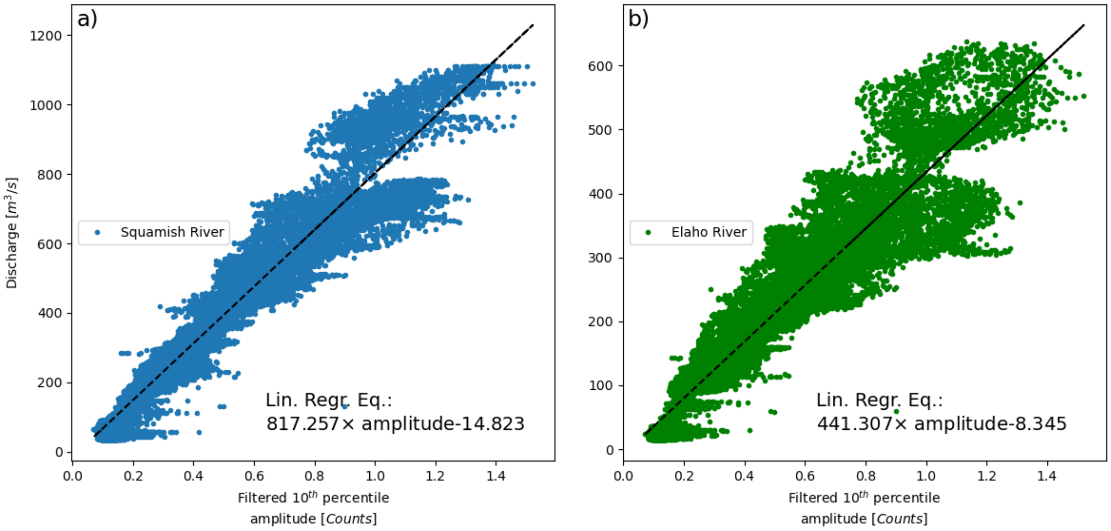

Figure 4.13: Relation between 10^{th} percentile filtered amplitudes and discharge rates at a)the Squamish River and b)the Elaho River.

4.3.1 Cleaning the acoustic records

To achieve a cleaner time series of amplitudes, at least as a first approximation, the 10^{th} percentile amplitudes selected as the baseline series was contrasted with the rest of the series of statistics that were noisier (Appendix D.1). By calculating the ratio between the other statistics versus that of the 10^{th} percentile, variability thresholds were established that define potential anomalies in the data. Figure 4.14a displays the (cropped) ratio of the standard deviation of the $3 - 8Hz$ filtered amplitude to the 10^{th} percentile of the filtered amplitude. This ratio describes how much the values in a given window of time fluctuate compared with the baseline values in the same window, and serves as a tool to identify anomalous behaviour. The effects of applying thresholds to the ratios to identify and discard anomalous data are shown in Figure 4.14b, where the original 10^{th} percentile of the filtered amplitudes along with the cleaned series are plotted.

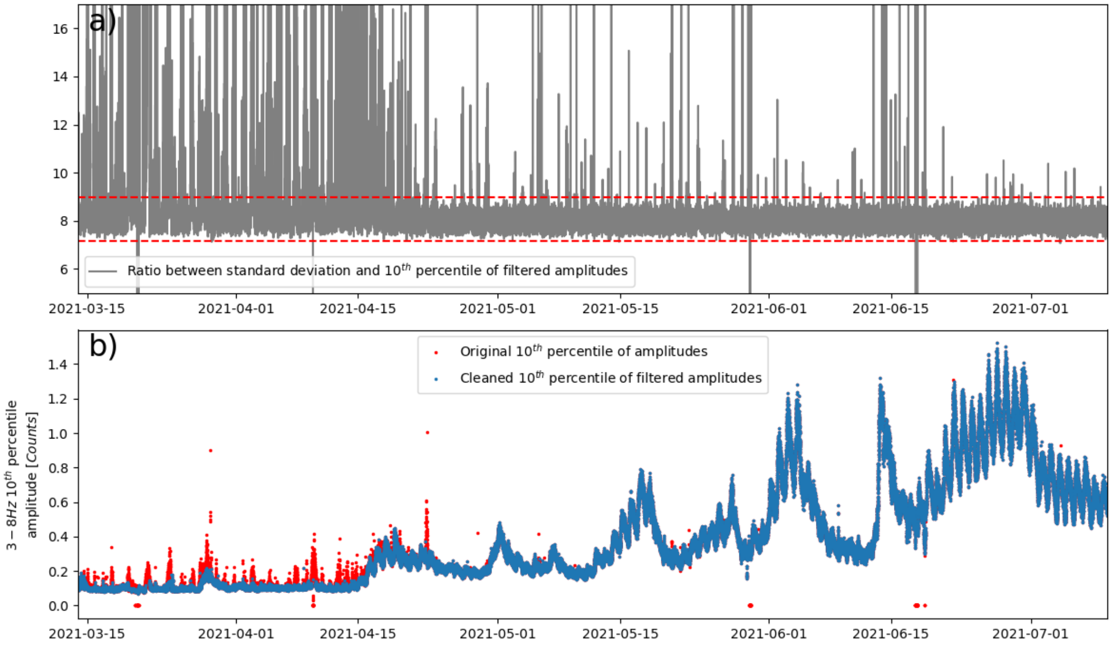

Figure 4.14: a) Ratio between the standard deviation of the filtered amplitudes and the 10^{th} percentile filtered amplitudes. b) Cleaned time series (blue) by discarding points (red) that were outside of thresholds defined on the ratio.

The effects of this preliminary anomalous data cleaning procedure are further detailed in

Figure 4.15, where only amplitudes falling inside the threshold shown in Figure 4.14a are plotted along the discharge rates.

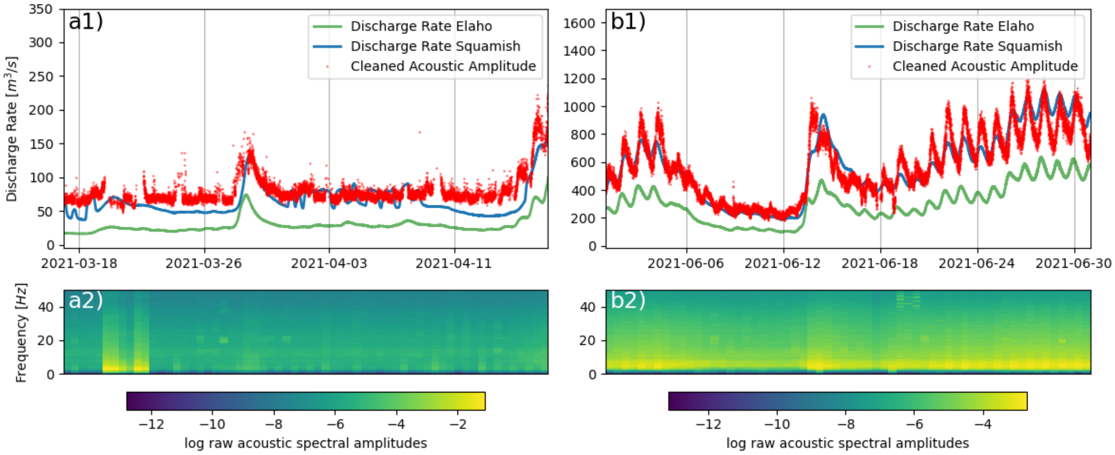

Figure 4.15: (a1,b1) Discharge rates and cleaned filtered amplitudes. (a2,b2) The corresponding unmasked spectrograms for unprocessed amplitudes.

4.4 Discussions

In Figure 4.11 we see the discharge rates time series for both sensors, at Elaho and Squamish, the time series for the 10^{th} percentile of the $3-8Hz$ filtered absolute amplitude, and the semi-diurnal spectrogram for the pre-processed acoustic recordings filtered above 3Hz for whole study period. A high correlation between the discharge rates and filtered amplitudes is evident even when one discharge rate sensor is not in the same river (Elaho) and the other (Squamish) is more than $30\,km$ downstream. The semi-diurnal spectrogram shows that there is a quasi-periodical daily change of power across many frequencies and is more distinguishable during the first half of the study period and related to wind noise (Figure 4.6). Daily fluctuations of the atmospheric temperature and wind have been identified and characterized using infrasound in a relatively stable volcanic setting (*Ortiz et al.*, 2021), and here it becomes superseded by the river noise as discharge rates increase during the second half of the study period (Figure 4.11), due to the proximity of our sensor to the river (Figure 4.3).

In Figure 4.12 we focused on two specific time periods that presented sudden water discharge rate changes at both sensors that were not related to increase in temperature (as seen from comparing temperatures to discharge rates in Figure 4.5). The first surge occurred on March 28^{th} and is most likely related to a short but very strong storm that occurred on that day and that caused heavy rains and winds across the southwest of the province (*Weisgarber, M*, 2021). The other surge is much larger and occurred on June 13 and is arguably related to continuous rains in the region prior to that date (Figure 4.5), although no extraordinary report exists for that specific period other than the heat wave occurring a few days later. Other than the obvious difference in discharge rates due to one surge occurring in spring and the other in summer, only subtle differences exist between the two surges (Figures 4.12 and 4.15).

From the first zoomed image (Figure 4.12a1) we can infer that the diurnal noise due to winds is strong enough to surpass the background river noise. This diurnal noise is observed at other frequency bands as inferred from the semi-diurnal spectrograms (Figures 4.11c and 4.12 b2). The spectrogram on the first period of interest (Figure 4.12a2) has been unmasked to show an anomaly with great power between March 19 and 23. We could not relate this anomaly to any weather parameter in the region (Figure 4.5) but the method was able to bypass part of this anomaly (Figure 4.15a1), and not incorporate it fully into the amplitudes that track the water discharge rate.

On the second zoomed image (Figure 4.12b1), the noise from wind is completely obscured by the Squamish River's discharge increase (i.e., there are many gaps due to anomaly cleaning in Figure 4.15a1 compared to Figure 4.15b1). In this case, diurnal fluctuations in the amplitudes form clear crests and valleys that generally coincide between discharge rates and acoustic amplitudes. These fluctuations are due to diurnal temperature changes that melt the region's snow cover and glaciers at different rates, which in turn change the water discharge rates. We highlight the fact that while crests and valleys generally coincide between the acoustic and discharge rate time series, they do not necessarily track perfectly in time. For instance, for the surges, the acoustic increasing amplitudes appear to arrive at the same time if not earlier than those in the discharge rate series (both at Elaho's sensor –upstream and the one at Squamish –downstream). On the other hand, as seen immediately before the second surge (Figure 4.12b1), the acoustic amplitudes seem to "lag behind" the discharge rate series (most notably when they start to wane). This suggests that the acoustic signal can be used to track sudden discharge increases better than waning periods.

The direct comparisons were possible because acoustic amplitudes were rescaled to track the discharge rates. The rescaling was done using a linear regression between the 10^{th} percentile filtered amplitudes and the discharge rates (Figure 4.13). It is clear that the Squamish River's discharge rate has a better relationship with the acoustic series than the Elaho's discharge rate which is not surprising since the acoustic sensor is measuring roughly the "same river" as the water level $\sim 30\,km$ south. This is only partially true however, since there is a mayor tributary (Ashlu Creek), between the study site and the sensor near Brackendale (Figure 4.2). Due to this large tributary and other minor ones downstream, we presume that the relationship between the acoustic amplitudes and discharge rates would be much closer if there were local discharge rate measurements at the study site. It is also possible that even Mud Creek could have created sensible noise at least during the times of significant snow/glacier melt.

Even with the 10^{th} percentile cleaning of the data (Figure 4.11), there are moments when acoustic amplitudes deviate considerably from discharge rate. There are two cases, when deviations are larger than the expected/average behaviour, and when deviations are below the expected value. For the later case, a quick solution to disregard the bad data is exerting a threshold to discard readings when they fall below the site's baseline noise since there is a clear amplitude in the first few days of the study. For the former case, we used the ratio of the standard deviation of the filtered amplitudes to their 10^{th} percentiles to detect periods when the variability of the data is too large or too small compared to the baseline behaviour. By defining thresholds that separate "stable" data from atypical data, much cleaner acoustic amplitudes could be generated that are less influenced by wind transients (Figure 4.15). The cleaning comes with a cost however, as data continuity is lost, and data loss can be especially significant during times of high noise (Figure 4.15a1-2). It is also plausible that the thresholds are site-dependant and need to be defined in a case-by-case manner. Nevertheless, a clean steady background noise can be established at least as a first-order approximation without incurring heavy calculations.

By following and extending the approaches from *Anderson et al.* (2019) and *Osborne et al.* (2021), i.e., by applying percentile statistics plus the standard deviation to acoustic amplitudes, the effect of transient phenomena can be mitigated due to a persistent and consistent baseline noise produced by the Squamish River at the site that permitted the good correlation with acoustic amplitudes. The relatively stable correspondence of the acoustic amplitude with the many discharge rates during the whole period is due to the somewhat ideal conditions of the site. It is possible that the rock wall in front of the sensor location (seen in Figure 4.3) acts as a reflector that allows the river noise to remain "trapped" in the area despite potential strong winds and weather conditions. Additionally, due to the presence of rocks of various sizes in the river, and the roughness of the rock wall, even when discharge rates increase and surpass the height of the smaller rocks (thus reducing some effective roughness elements), the noise sources from channel obstacles in the water-atmosphere interface remain important enough to produce noise in various frequencies (*Osborne et al.*, 2022) (Figure 4.4). Related to this, while we can not fully explain the important signals in the lower frequencies ($\leq 10Hz$) as opposed to *Ronan et al.* (2017) and *Tatum et al.* (2023), it is likely due to the overall size of the noise producing structures and the geometry of the setting as a whole that led to strong harmonics in the lower frequencies when the discharge rates began to rise after April 15, 2021.

Using water discharge rates likely led to better correlation with acoustic amplitudes than with stages because water levels are strongly conditioned by river channel geometries. This is apparent when both river's discharge rates and levels are compared (Appendix D.2) and while the Elaho River has less streamflow than the Squamish River, at a certain point, the river stages become almost identical as an artefact of local river channel geometries. Thus, we emphasize that discharge rates better characterize flows over a region as well as locally, and should be used in other natural settings for similar studies. Of course, better results are expected if using local discharge rates, and this also implies that periodic calibration might be needed to determine the relationship between acoustic amplitudes and the discharge rates as is customary for some methods (*Dobriyal et al.*, 2017; *Llaban and Ella*, 2022), especially after events such as debris-flows that may alter the local noise sources. Future research may include using other frequency bands for the purpose of identification of anomalies, stemming from this stable/robust way to track the baseline behaviour. Additionally, this is a step towards the future implementation of machine learning based early warning systems that can be hard to implement because sensors may not have been deployed long enough, or because some mass movement events are rare (*Allstadt et al.*, 2018). The addition of co-located seismic sensors along with cameras could help infer a more precise local water discharge rate as well as deepen knowledge on the evolution of mass flow parameters (e.g., *Walsh et al.*, 2023; *Bosa et al.*, 2024).

It was shown that the method seems to be robust enough when taking into account on-site characteristics, since its application to the dataset of an independent study (*Tatum et al.*, 2023) produced comparable performance for acoustic discharge rate tracking.

4.5 Conclusions

In this work we showed that the low frequency region of sound is suitable to track discharge rates, at least for rivers with characteristics similar to the Squamish River by considering data filtered in specific bands of frequencies, and extracting percentile statistics. The method gave similar results when used in an independent dataset. A good empirical correlation between discharge rate and filtered infrasound amplitude could be established in a wide range of values as the 2021 heatwave was a perfect opportunity to test the capabilities to detect very large discharge rate variations. Additionally, atypical surges unrelated to the snow/glacier melting driven by a rise in temperature were successfully identified and their onset was tracked at least as fast as the surge passed near the instrument. The flexibility of the method allowed for the simple extraction of parameters that helped further clean the filtered acoustic amplitudes for a better description of the area's background noise. This represents the first step into the creation of early warning systems that require both simple instrumentation and that allow rapid/local signal processing in remote areas. Furthermore, this approach has an even greater advantage especially since the acoustic sensor is not limited to measuring only river signals, and can eventually be used to capture sound-producing local anomalies besides debris-flows, sudden water surges or floods.

References

Allstadt, K. E., R. S. Matoza, A. B. Lockhart, S. C. Moran, J. Caplan-Auerbach, M. M. Haney, W. A. Thelen, and S. D. Malone (2018), Seismic and acoustic signatures of surficial mass movements at volcanoes, *Journal of Volcanology and Geothermal Research*, *364*, 76–106, doi:https://doi.org/10.1016/j.jvolgeores.2018.09.007.

Anderson, J., T. Ronan, H. D. Ortiz, J. B. Johnson, and D. C. Bowman (2019), Acoustics as a tool for detecting and inferring streamflow, in *AGU Fall Meeting Abstracts*, vol. 2019, pp. A21S–2804.

Anderson, J. F., J. B. Johnson, D. C. Bowman, and T. J. Ronan (2017), The Gem Infrasound Logger and Custom-Built Instrumentation, *Seismological Research Letters*, *89*(1), 153–164, doi:10.1785/0220170067.

Anderson, J. F., K. S. Anderson, and T. Beschorner (2023), gemlog: Data conversion for the open-source gem infrasound logger, *Journal of Open Source Software*, *8*(86), 5256, doi:10.21105/joss.05256.

Arnell, N. W., and S. N. Gosling (2016), The impacts of climate change on river flood risk at the global scale, *Climatic Change*, *134*(Volume 134, 2016), 387–401, doi:10.1007/s10584-014-1084-5.

Belli, G., F. Walter, B. McArdell, D. Gheri, and E. Marchetti (2022), Infrasonic and seismic analysis of debris-flow events at illgraben (switzerland): Relating signal features to flow parameters and to the seismo-acoustic source mechanism, *Journal of Geophysical Research: Earth Surface*, *127*(6), e2021JF006,576, doi:https://doi.org/10.1029/2021JF006576, e2021JF006576 2021JF006576.

Bevington, A. R., and B. Menounos (2022), Accelerated change in the glaciated environments of western canada revealed through trend analysis of optical satellite imagery, *Remote Sensing of Environment*, *270*, 112,862, doi:https://doi.org/10.1016/j.rse.2021.112862.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010), ObsPy: A Python Toolbox for Seismology, *Seismological Research Letters*, *81*(3), 530–533, doi:10.1785/gssrl.81.3.530.

Bosa, A., G. Bejar, G. Waite, J. Mock, A. Pineda, and J. Anderson (2024), Dynamics of rain-triggered lahars and destructive power inferred from seismo-acoustic arrays and time-lapse camera correlation at volcán de fuego, guatemala, *Research Square*, preprint.

Bowman, D. C., and S. A. Albert (2018), Acoustic event location and background noise characterization on a free flying infrasound sensor network in the stratosphere, *Geophysical Journal International*, *213*(3), 1524–1535, doi:10.1093/gji/ggy069.

Brissaud, Q., S. Krishnamoorthy, J. M. Jackson, D. C. Bowman, A. Komjathy, J. A. Cutts, Z. Zhan, M. T. Pauken, J. S. Izraelevitz, and G. J. Walsh (2021), The first detection of an earthquake from a balloon using its acoustic signature, *Geophysical Research Letters*, *48*(12), e2021GL093,013, doi:https://doi.org/10.1029/2021GL093013, e2021GL093013 2021GL093013.

Chiarle, M., M. Geertsema, G. Mortara, and J. J. Clague (2021), Relations between climate change and mass movement: Perspectives from the canadian cordillera and the european alps, *Global and Planetary Change*, *202*, 103,499, doi:https://doi.org/10.1016/j.gloplacha.2021.103499.

Clague, J. J., and J. G. Souther (1982), The dusty creek landslide on mount cayley, british columbia, *Canadian Journal of Earth Sciences*, *19*(3), 524–539, doi:10.1139/e82-043.

Clague, J. J., B. Menounos, and R. Wheate (2011), *Canadian Rockies and Coast Mountains of Canada*, pp. 106–111, Springer Netherlands, Dordrecht, doi:10.1007/978-90-481-2642-2_51.

Cruden, D. M., and Z. Y. Lu (1992), The rockslide and debris flow from mount cayley, b.c., in june 1984, *Canadian Geotechnical Journal*, *29*(4), 614–626, doi:10.1139/t92-069.

Dannemann Dugick, F., C. Koch, E. Berg, S. Arrowsmith, and S. Albert (2023), A New Decade in Seismoacoustics (2010–2022), *Bulletin of the Seismological Society of America*, *113*(4), 1390–1423, doi:10.1785/0120220157.

Dobriyal, P., R. Badola, C. Tuboi, and S. A. Hussain (2017), A review of methods for monitoring streamflow for sustainable water resource management, *Applied Water Science*, *7*(6), 2617–2628, doi:10.1007/s13201-016-0488-y.

ECCC-Historical Climate Data web site (2024), Environment and Climate Change Canada Historical Climate Data web site, `https://climate.weather.gc.ca/historical_data/search_historic_data_e.html`.

ECCC-Real-time Hydrometric Data web site (2024), Environment and Climate Change Canada Real-time Hydrometric Data web site, `https://wateroffice.ec.gc.ca/search/real_time_e.html`.

Fath, K. J. (2014), Late holocene history of squamish river north of brackendale, british columbia, Master's thesis, Simon Fraser University, Burnaby, Canada.

Finkel, J., E. P. Gerber, D. S. Abbot, and J. Weare (2023), Revealing the statistics of extreme events hidden in short weather forecast data, *AGU Advances*, *4*(2), e2023AV000,881, doi:https://doi.org/10.1029/2023AV000881, e2023AV000881 2023AV000881.

Gillett, N. P., A. J. Cannon, E. Malinina, M. Schnorbus, F. Anslow, Q. Sun, M. Kirchmeier-Young, F. Zwiers, C. Seiler, X. Zhang, G. Flato, H. Wan, G. Li, and A. Castellan (2022), Human influence on the 2021 british columbia floods, *Weather and Climate Extremes*, *36*, 100,441, doi:https://doi.org/10.1016/j.wace.2022.100441.

Hundt, S., and K. Blasch (2019), Laboratory assessment of alternative stream velocity measurement methods, *PLOS ONE*, *14*(9), 1–21, doi:10.1371/journal.pone.0222263.

Iukhno, A., S. Buzmakov, and A. Zorina (2021), Water discharge measuring instruments: An up to date overview, *ENVIRONMENT. TECHNOLOGIES. RESOURCES.*

Proceedings of the International Scientific and Practical Conference, *3*, 116–123, doi: 10.17770/etr2021vol3.6613.

Johnson, J. B., J. F. Anderson, H. P. Marshall, S. Havens, and L. M. Watson (2021), Snow avalanche detection and source constraints made using a networked array of infrasound sensors, *Journal of Geophysical Research: Earth Surface*, *126*(3), e2020JF005,741, doi: https://doi.org/10.1029/2020JF005741, e2020JF005741 2020JF005741.

Kogelnig, A., J. Hübl, E. Suriñach, I. Vilajosana, and B. W. McArdell (2014), Infrasound produced by debris flow: propagation and frequency content evolution, *Natural Hazards*, *70*(3), 1713–1733, doi:10.1007/s11069-011-9741-8.

Krichen, M., M. S. Abdalzaher, M. Elwekeil, and M. M. Fouda (2024), Managing natural disasters: An analysis of technological advancements, opportunities, and challenges, *Internet of Things and Cyber-Physical Systems*, *4*, 99–109, doi:https://doi.org/10.1016/j.iotcps.2023.09.002.

Kumar, P., S. E. Debele, J. Sahani, N. Rawat, B. Marti-Cardona, S. M. Alfieri, B. Basu, A. S. Basu, P. Bowyer, N. Charizopoulos, J. Jaakko, M. Loupis, M. Menenti, S. B. Mickovski, J. Pfeiffer, F. Pilla, J. Pröll, B. Pulvirenti, M. Rutzinger, S. Sannigrahi, C. Spyrou, H. Tuomenvirta, Z. Vojinovic, and T. Zieher (2021), An overview of monitoring methods for assessing the performance of nature-based solutions against natural hazards, *Earth-Science Reviews*, *217*, 103,603, doi:https://doi.org/10.1016/j.earscirev.2021.103603.

Kästner, K., A. J. F. Hoitink, P. J. J. F. Torfs, B. Vermeulen, N. S. Ningsih, and M. Pramulya (2018), Prerequisites for accurate monitoring of river discharge based on fixed-location velocity measurements, *Water Resources Research*, *54*(2), 1058–1076, doi: https://doi.org/10.1002/2017WR020990.

Le Moigne, Y., J. Zurek, G. Williams-Jones, E. Lev, A. Calahorrano-Di Patre, and J. Anzieta (2020), Standing waves in high speed lava channels: A tool for constraining lava dynamics and eruptive parameters, *Journal of Volcanology and Geothermal Research*, *401*, 106,944, doi:https://doi.org/10.1016/j.jvolgeores.2020.106944.

Liu, D.-l., X.-p. Leng, F.-q. Wei, S.-j. Zhang, and Y. Hong (2015), Monitoring and recognition of debris flow infrasonic signals, *Journal of Mountain Science*, *12*(4), 797–815, doi:10.1007/s11629-015-3471-4.

Llaban, A. B., and V. B. Ella (2022), Conventional and sensor-based streamflow data acquisition system for sustainable water resources management and agricultural applications: an extensive review of literature, *IOP Conference Series: Earth and Environmental Science*, *1038*(1), 012,040, doi:10.1088/1755-1315/1038/1/012040.

Marchetti, E., and J. B. Johnson (2023), Chapter one - infrasound array analysis of rapid mass movements in mountain regions, in *Advances in Geophysics*, *Advances in Geophysics*, vol. 64, edited by C. Schmelzbach, pp. 1–57, Elsevier, doi:https://doi.org/10.1016/bs.agph.2023.06.001.

Marchetti, E., F. Walter, G. Barfucci, R. Genco, M. Wenner, M. Ripepe, B. McArdell, and C. Price (2019), Infrasound array analysis of debris flow activity and implication for early warning, *Journal of Geophysical Research: Earth Surface*, *124*(2), 567–587, doi:https://doi.org/10.1029/2018JF004785.

Marshall, H. P., J. B. Johnson, J. Anderson, S. De Angelis, R. P. Escobar-Wolf, J. J. Lyons, and A. Pineda (2019), On the capabilities of networked infrasound arrays for investigating rapid gravity-driven mass movements: lahars and snow avalanches, in *AGU Fall Meeting Abstracts*, vol. 2019, pp. V44B–02.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden (2011), Quantitative analysis of culture using millions of digitized books, *Science*, *331*(6014), 176–182, doi:10.1126/science.1199644.

Naveau, P., A. Hannart, and A. Ribes (2020), Statistical methods for extreme event attribution in climate science, *Annual Review of Statistics and Its Application*, *7*(Volume 7, 2020), 89–110, doi:https://doi.org/10.1146/annurev-statistics-031219-041314.

Ortiz, H. D., R. S. Matoza, J. B. Johnson, S. Hernandez, J. C. Anzieta, and M. C. Ruiz (2021), Autocorrelation infrasound interferometry, *Journal of Geophysical Research: Solid Earth*, *126*(4), e2020JB020,513, doi:https://doi.org/10.1029/2020JB020513, e2020JB020513 2020JB020513.

Osborne, W. A., R. A. Hodge, G. D. Love, P. Hawkin, and R. E. Hawkin (2021), Babbling brook to thunderous torrent: Using sound to monitor river stage, *Earth Surface Processes and Landforms*, *46*(13), 2656–2670, doi:https://doi.org/10.1002/esp.5199.

Osborne, W. A., R. A. Hodge, G. D. Love, P. Hawkin, and R. E. Hawkin (2022), The influence of in-channel obstacles on river sound, *Water Resources Research*, *58*(4), e2021WR031,567, doi:https://doi.org/10.1029/2021WR031567, e2021WR031567 2021WR031567.

Perera, D., O. Seidou, J. Agnihotri, H. Mehmood, and M. Rasmy (2020), Challenges and technical advances in flood early warning systems (fewss), in *Flood Impact Mitigation and Resilience Enhancement*, edited by G. Huang, chap. 2, IntechOpen, Rijeka, doi:10.5772/intechopen.93069.

Pfeffer, W. T., A. A. Arendt, A. Bliss, T. Bolch, J. G. Cogley, A. S. Gardner, J.-O. Hagen, R. Hock, G. Kaser, C. Kienholz, and et al. (2014), The randolph glacier inventory: a globally complete inventory of glaciers, *Journal of Glaciology*, *60*(221), 537–552, doi:10.3189/2014JoG13J176.

Ronan, T. J., J. M. Lees, T. D. Mikesell, J. F. Anderson, and J. B. Johnson (2017), Acoustic and seismic fields of hydraulic jumps at varying froude numbers, *Geophysical Research Letters*, *44*(19), 9734–9741, doi:https://doi.org/10.1002/2017GL074511.

Rosenblatt, B. B., J. B. Johnson, J. F. Anderson, K. Kim, and S. J. Gauvain (2022), Controls on the frequency content of near-source infrasound at open-vent volcanoes: a case study from volcán villarrica, chile, *Bulletin of Volcanology*, *84*(103), doi:https://doi.org/10.1007/s00445-022-01607-y.

Sanderson, R. W., R. S. Matoza, R. M. Haymon, and J. H. Steidl (2021), A Pilot Experiment on Infrasonic Lahar Detection at Mount Adams, Cascades: Ambient Infrasound and Wind-Noise Characterization at a Quiescent Stratovolcano, *Seismological Research Letters*, *92*(5), 3065–3086, doi:10.1785/0220200361.

Seybold, E. C., A. Bergstrom, C. N. Jones, A. J. Burgin, S. Zipper, S. E. Godsey, W. K. Dodds, M. A. Zimmer, M. Shanafield, T. Datry, R. D. Mazor, M. L. Messager, J. D. Olden, A. Ward, S. Yu, K. E. Kaiser, A. Shogren, and R. H. Walker (2023), How low can you go? widespread challenges in measuring low stream discharge and a path forward, *Limnology and Oceanography Letters*, *8*(6), 804–811, doi:https://doi.org/10.1002/lol2.10356.

Sutherland, C. (2015), Regional Climate and Hydrometric Monitoring Network Scoping Study, https://www.rdn.bc.ca/dms/documents/dwwp-reports/region-wide-reports/rdn_regional_hydrometric_and_climate_monitoring_scoping_study_-_2015.pdf.

Tatum, T. A., J. F. Anderson, and T. J. Ronan (2023), Whitewater sound dependence on discharge and wave configuration at an adjustable wave feature, *Water Resources Research*, *59*(8), e2023WR034,554, doi:https://doi.org/10.1029/2023WR034554, e2023WR034554 2023WR034554.

Walsh, B., C. Lormand, J. Procter, and G. Williams-Jones (2023), Characterizing the evolution of mass flow properties and dynamics through analysis of seismic signals: insights from the 18 march 2007 mt. ruapehu lake-breakout lahar, *Natural Hazards and Earth System Sciences*, *23*(3), 1029–1044, doi:10.5194/nhess-23-1029-2023.

Ward, P. J., E. C. de Perez, F. Dottori, B. Jongman, T. Luo, S. Safaie, and S. Uhlemann-Elmer (2018), *The Need for Mapping, Modeling, and Predicting Flood Hazard and Risk at the Global Scale*, chap. 1, pp. 1–15, American Geophysical Union (AGU), doi:https://doi.org/10.1002/9781119217886.ch1.

Weisgarber, M (2021), Spring wind storm knocks down trees and leaves tens of thousands in the dark., https://bc.ctvnews.ca/

spring-wind-storm-knocks-down-trees-and-leaves-tens-of-thousands-in-the-dark-1.5366464.

Zhao, C., T. Pan, J. Xia, S. Yang, J. Zhao, X. Gan, L. Hou, and S. Ding (2019), Streamflow calculation for medium-to-small rivers in data scarce inland areas, *Science of The Total Environment*, *693*, 133,571, doi:https://doi.org/10.1016/j.scitotenv.2019.07.377.

Zhu, Z.-N., X.-H. Zhu, C. Zhang, M. Chen, H. Zheng, Z. Zhang, J. Zhong, L. Wei, Q. Li, H. Wang, S. Li, and A. Kaneko (2021), Monitoring of yangtze river discharge at datong hydrometric station using acoustic tomography technology, *Frontiers in Earth Science*, *9*, doi:10.3389/feart.2021.723123.

Chapter 5

Conclusions

Natural hazards have been and will be part of human history for our entire existence. For this reason, our position to face them must be increasingly proactive instead of reactive. The constant increase in analytical capabilities given by better geophysical theories, new instrumentation, and faster and more powerful computing must be used in order to improve our response and mitigate their effects. In this work I have used and combined a group of computational methods and tools of diverse nature to carry out tasks that increased the knowledge of several also diverse data sets. To achieve this new knowledge, the tasks that were executed, along with the applied tools for each case, were conceived, designed, and adapted around the previous information and the nature of each dataset. For the Llaima and Cotopaxi seismic datasets and catalogues -where one of them was believed to be noisy- I applied a catalogue cleaning procedure. For the Tungurahua volcano acoustic dataset, I combined classical data analysis tools, unsupervised, supervised, and deep learning tools (as well as a modification of the catalogue cleaning procedure) to upgrade a previously created human catalogue. For the Mount Cayley dataset, a quick data analysis tool was defined to monitor the Squamish River's local discharge rate.

The specific conclusions for each chapter are described next, followed by final remarks and suggestions for future work.

5.1 Catalogue Cleaning Procedure

The development and application of the catalogue cleaning procedure to locked datasets with multiple classes led to the definition of an important concept pertaining to the level of "trust" one has to the assigned label of an event within a catalogue: the *consistency.* This value helps answer the third question of this project (*Can a degree of uncertainty be assigned to the assumption that each event's label may or may not be mislabelled?*) by definition, because of its "consensus" nature. However, we found the level of uncertainty is highly dependent on the quality of the catalogue. For example, for a very good/clean catalogue, a consistency of 0.5 should be considered bad (half the predictors see this event as the other class!), however if the catalogue is noisy, that value could be biased by an imbalanced class noise (as we found for the Cotopaxi catalogue). Nevertheless, the definition of the consistency alone also answers the second question directly. That is, events with low consistencies can be suspected of being noisy, and should be further reviewed or discarded. The initial question is consequently dealt with −*Can the quality of a given catalogue be assessed?* Yes, after the catalogue cleaning procedure, a statistical analysis of the consistencies within each family or class can inform if there are potential mistakes in the labels of the events. If a catalogue is clean, the large majority of the events within a family should have high consistencies, and more notably, the distribution of consistences should be left tailed which means that fewer and fewer events should be more and more uncertain. Interestingly enough, I found that the family whose distribution does not follow the ideal pattern is not necessarily the one that has the noise, as observed in the Cotopaxi catalogue.

The relevance of this procedure is multiple because not only can it be used for cleaning other catalogues with similar characteristics to the ones in this study, but it can serve as a way to quickly refine newly created catalogues conserving only high quality events as I showed for the subsequent chapter.

A minor additional conclusion is that the whole concept of 'consensus' of various classifiers could be applied only because their training is fast/simple, and thus this work serves as an example of the lingering relevance, potential importance, and discernible applications of less novel tools (compared to Deep Learning).

5.2 Partially known dataset exploration

Unlocked datasets are incredibly rich sources of new knowledge because of the sheer amount of data (and phenomena occurring simultaneously) continuous seismic or acoustic recordings have. For Tungurahua volcano, a human catalogue of explosions was used as the primary source to feed a supervised classifier that helped distinguish explosions from non-explosions obtained by classical algorithm detections. This task alone is noteworthy because a better characterization of the level of explosive activity was achieved. However, finding events with confidence was just the first part of the task, and by using unsupervised learning algorithms, I showed that explosive patterns repeated over the years and were associated to different eruptive activity. Even more so, I showed that deep learning led to similar results as the classical unsupervised algorithms, but in a fraction of the time. I also provided guidance on how to further enhance these already effective techniques within their respective contexts. This chapter served to explicitly test part of the potential of machine learning when applied to unlocked datasets, and much like this thesis itself, it helped to define "a sequence of knowledge gains" for datasets with partial previous information. From detection, to cleaning, to pattern identification, a suite of tools were used/proposed to effectively deal with a very large (10 years) dataset. The explored tools are flexible enough that their application to multi-station datasets can lead to better results. These results can be combined with the seismic signals associated with explosions to gain even deeper knowledge of the explosive activity.

5.3 Extracting knowledge from a fully unknown dataset

Due to the intrinsic uncertainty related to a brand new dataset and location, starting a new monitoring network from zero is quite a challenging task. To successfully monitor sporadic events such as debris flows or avalanches, one must have knowledge of the base noise level in a region. At the same time, if a region is remote, efforts should be made to optimize anomaly detection and early warning alert processes. In the last chapter of this thesis, I proposed a procedure to track the level of the Squamish River that also allowed us to define potentially anomalous data. The construction of this procedure was done both by doing a site-dependent analysis of possible frequency bands of interest to achieve good tracking, and also with the help of multiple regional climate data to give confidence to the results obtained.

The search for adequate frequency bands for discharge rate monitoring was shown to be of crucial importance because it defines the precision of the results derived from the proposed procedure. Additionally, discharge rate as opposed to water level was also found to be a more meaningful magnitude to correlate sound to. The results from this chapter corroborate and highlight low frequency/infrasound sensors as suitable tools for tracking streamflow discharge rates in remote/uncontrolled regions due to the rugged characteristics of the environment where the instrument was installed, and lay the ground for future extensions to natural hazards monitoring. In particular, this study demonstrated the capabilities and relevance of low-cost sensors (in this case the Gem Infrasound Logger) that will serve to power increasingly comprehensive warning systems.

5.4 Concluding Remarks and Future Work

Seismic and acoustic data have been and will remain invaluable tools for monitoring natural hazards, because they provide insights about geophysical processes below the Earth's surface, at the interface with the atmosphere, and once they become totally superficial, in as close as we can to real-time. However, these signals need to be carefully analyzed to be truly useful, both when dealing with large and complex monitoring networks, as in the case of little instrumentation and limited datasets.

Machine learning is a very versatile set of tools, but contrary to popular perception, it cannot be implemented lightly; it must be applied with extensive knowledge of the task to be performed. Furthermore, just as when we apply artificial intelligence tools to everyday life and we do not (should not) use them blindly without vetting the information, in the same way (and even more so) machine learning applied to monitoring natural disasters must be carefully controlled by humans.

Interestingly, the order of the chapters is given according to increasing uncertainty (and chronological completion), but in situations of new monitoring networks, the steps would be executed in reverse, that is, first the baseline knowledge of a region should be established, then exhaustive catalogues would be created describing signals of interest and patterns in the data, and finally the quality of these created catalogues may be evaluated. The fact that these tools have already been created and are available means that they can be now replicated in the proper sequence for said situations.

Specific future work for each of the chapters is varied: For the second chapter it would be interesting to test the procedure with data from other volcanoes (e.g., Mt Saint Helens, Deception Island volcano), and also to look for improvements on synthetic datasets. Tasks remaining for the third chapter include interpretation of the families based on joint information with the co-located seismic sensors and other stations plus other geophysical parameters collected during the activity of Tungurahua volcano such as deformation and gases. Additionally, information of the actual effects of the explosions (e.g., emission of water vapour, mild or strong ash, pyroclastic density currents, etc.) based on monitoring reports by IGEPN investigated to determine links to the different families along with other parameters not used in this study (such as acoustic and seismic amplitudes). In terms of the methodology, other wavelet families or network geometries could be tested more to either further reduce processing time, or increase reconstruction (and thus compression) accuracy, and eventually even modelling. As for the fourth chapter, besides the addition of collocated seismic sensors and cameras, advanced anomaly detection tools based on deep learning could be implemented as more data and knowledge is acquired from the area.

Although this chapter is called ‘Conclusions’, at least based on the results of chapters 3 and 4, these actually seem like the first steps towards more robust systems. Although the need to perform this study for settings with a single station was always in mind, the possibilities of obtaining even more refined results will increase with the use of multiple stations. Furthermore, the possible integration of these methods in conjunction with other types of continuous geophysical signals (e.g., deformation) expands the range of possibilities towards an eventual more powerful monitoring and natural hazards forecasting.

In the same vein, the development of more complex machine learning tools that can not only learn from the data, but also integrate geophysical knowledge will push the limits of what can be achieved even further, but this can be achieved only as long as there are solid foundations, as has been my goal in this work.

Appendix A

Supplementary Information for the Introduction

A.1 Brief description of some tools

Some typical (*James et al.*, 2013; *Géron*, 2019) machine learning supervised and unsupervised tools are:

Supervised Learning

k-Nearest Neighbours: The K-Nearest Neighbours (KNN) algorithm consists of assigning labels to unknown observations based on the majority class of its k most proximal neighbours. The "proximity" will be defined based on a metric. KNN struggles in higher dimensions due to the curse of dimensionality (i.e., 10 points at random in a given length are closer than in a given area, a given volume, and so on). It adds potential bias and variance in an uncontrollable manner and it also is hard to compute (you need to check for points) for large n and p.

Logistic regression: This is a traditional statistical technique designed specifically for binary responses that assumes a linear model for the "logit" function (the log of the oddsratio).

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

The parameters are estimated using maximum likelihood so that the final estimated value is $\hat{p}(x)$ (probability of success or class).

Discriminant Analysis: Linear Discriminant Analysis (LDA) is another linear method for classification where Gaussians for each class are defined (assuming same predictor correlation and variance for each class) and then equal probability density linear boundaries are produced. Logistic regressions are unstable when there is little to no overlap of classes in any dimension (quasi step or step function). The LDA models the distribution of the X given the response Y, i.e. $P(X|Y)$, then uses Bayes' Rule to estimate $P(Y|X)$. The Gaussians have means, variances and correlations for each class. The number of parameters grow as $K + p(p-1)/2$. If covariances are not the same then the boundaries are parabolas. This leads to QDA (Quadratic Discriminant Analysis) which is a more flexible model but needs $(K-1)(p(p+1)/2)$ parameters and that results in extra variance.

Support Vector Machines: Support Vector Machines (SVM) is a classification algorithm based on finding optimal decision boundaries (hyperplanes) when classes are separable and adapting linear boundaries when classes are not separable. The gap between the closest data points divided by boundaries is called MARGIN and the closest points are called Support Vectors. The idea is based on creating a function which assigns positive and negative coding of support vectors, and zero at the boundary, and maximizes the distance to support vectors. When points are not separable, the same approach must be corrected for overlap by calculating the errors (amount of deviance based on decision boundary) from misclassified points to the hypothetical decision boundary. The result is a linear boundary obtained non-parametrically; it does not assume a probability distribution or linear relation of the predictors (i.e., Logistic); it is a geometric idea. This technique can be extended to include "automatic" transformation of the features to polynomial or radial basis functions via the "Kernel trick".

non-linear SVM: In non-linear SVM, the approach is to transform the regressors X into a new set with higher dimensions and then project it into the original space. The new function is a linear combination of basis functions (arbitrarily chosen) of the predictors. These basis functions are called Kernels and the most popular are polynomial kernels and Gaussian Radial Basis functions.

Regression Trees: This algorithm creates simple assignation regions based on step functions as decision boundaries. They assign classes for terminal nodes based on majority membership in the node. The criteria to prune or create children nodes is based on relative error minimization for a partition, minimum parent node size to partition, minimum children node size to be created, and so on.

Ensembles-Voting Classifier: Voting classifier is a technique that uses a small group of trained classifiers, and then assigns a prediction based on a majority-wins voting scheme. The idea of this ensemble is that if sufficiently uncorrelated comparable classifiers tend to get responses right most of the time and vote for the final decision, the prediction tends to be correct more times than each of them separately. For instance, as an approximation, let us assume that 3 comparable independent classifiers have each 0.51 chance of getting a prediction right *ideally for all possible observations.* In a voting scheme, the probability of them getting the right answer is the probability of 2 or 3 getting the answer right simultaneously:

$$P(\text{2 or 3 correctly classified}) = 3C2 \times 0.51^2 \times 0.49 + 3C3 \times 0.51^3 \approx 0.515$$

which is 10% higher than each of the previous classifiers alone. When the probability of being right is greater, this gained percentage increases. If the individual classifiers all output probabilistic outcomes, the voting classifier can also be probabilistic, making it a "soft" voting classifier. On the contrary, the voting classifier assigns a specific class and is a "hard" classifier.

Ensembles-Bagging: This technique consists of averaging a set of bootstrapped (resampled observations and variables) learners to reduce the variability (Bootstrap Aggregation = B-Agging). It works best with low bias-high variance models and is called **ensemble of weak learners**.

Random Forests: Random forests are bagging regression trees with a random subsample of explanatory variables. When the number of parameters is large, the trees tend to be similar if using Bagging. Based on variance of correlated random variables, the idea is to "add random variability" to the trees to then average and reduce it. The subset m is usually $\sqrt{p}$ for classification (and $p/3$ for regression). Since the most important variables might be left out, the technique allows for less important variables to contribute in the explanation of variability. The predicted value of a given x is still the average of the many (tens or hundreds) trees. A drawback is that it loses tree structure (unable to visualize when and where splits occurred). This technique reduces variance of single classification trees.

Artificial Neural Networks: An Artificial Neural Network (or just Neural Network - NN) can be understood as a set of nodes (or *perceptrons*) that in some sense emulate neurons in a biological brain. These nodes form connected "layers" and output responses from the linear combinations of their inputs (+ biases) into other nodes. The linear combinations of inputs are fed into an activation function, for example, a sigmoid function that yields a value between 0 and 1 (to emulate a probability) and becomes the input of further connected nodes. The weights of the connections are optimized (usually by a backwards error propagation method) to minimize a cost of misclassifying each instance. The architectures of the networks are varied, and many versions exist with different problem-solving capabilities (*Géron*, 2019). Deep learning is a branch of machine learning that uses neural networks with many layers. One of the most important properties of these deep neural networks is that there is no need for explicit feature extraction since the systems can find efficient representations by themselves.

Unsupervised Learning

k-Means: k-Means is a popular unsupervised learning technique that consists of iteratively finding "centroids" of groups of unlabelled observations, such that the distance between grouped instances and these centroids and between observations close to this centroid is minimized. This is done by minimizing intra-group variance, equivalently maximizing inter-group variance. The number of groups is a parameter to be decided.

Hierarchical clustering: Hierarchical clustering are a family of procedures to find groups of most similar observations based on a dissimilarity matrix between each element and the rest. Among all the methods, the most common ones are to initially consider all elements disjoint and merging them based on similarity, or by considering all elements initially together and then partitioning them until they are all separated. As in k-Means, the number of groups is a parameter to be decided.

DBSCAN: Density-Based Spatial Clustering of Applications with Noise defines clusters by constructing dense observation regions. It is based on the definition of clusters based on *core* instances with a *minimum number* of neighbours closer than a small distance ϵ. Series of neighbouring cores with their non-core neighbours define a large single cluster. Instances with no neighbours are considered anomalies.

Spectral Clustering: The Spectral Clustering algorithm creates a low-dimensional representation from a similarity matrix between all instances, and uses another clustering algorithm (e.g., k-means) to group instances.

Self Organizing Maps: Self Organizing Maps (SOM) are a particular kind of Artificial Neural Network that iteratively finds and deforms a network of connected vectors (or nodes) such that topologically similar features are located closer together in the network, all observations are near a node in the network, and the network approximates the distribution of observations in the feature space. The number of nodes and geometry of the network have to be set.

Gaussian Mixture Models: Gaussian Mixture Models are a family of probabilistic models that assume that the observations in a given dataset belong to a normally distributed subpopulation, where the number of Gaussian subpopulations is given. Each Gaussian distribution helps defining a cluster or family of instances and has its own *unknown* mean and covariance matrix. To find all the unknown parameters the instances are initially given a set of arbitrary labels (obtained from k-means for example), and initial parameters of the Gaussian distributions are determined. An iterative process is then performed in which the algorithm calculates the probability of each data point belonging to each Gaussian distribution. These probabilities are then used to update the parameters of the Gaussian distributions, and the process is repeated until convergence. When the final parameters are found, the Gaussian Mixture Model is fully determined and now points in the feature space can be assigned to a given class based on the maximum of the probabilities of being generated from all the Gaussian distributions. The general idea is similar to LDA but for the context of unsupervised learning.

Classical data analysis tools

Notable classical data analysis tools are:

Spectrograms: Spectrograms are one type of visual representation of a signal's frequency content evolution over time. In broad terms it is obtained by dividing the signal in small time intervals or windows, calculating the frequency spectrum for each interval using the Fourier transform, and then plotting these spectra in an sequential manner as a function of time. The resulting image shows how the signal's frequency content changes over time by using varying colors indicating the time-frequency distribution of amplitudes. Figure 1.3 shows examples of signals with their spectrograms.

Scalograms - Continuous Wavelet Transforms representations: Scalograms are yet another way to describe a signal's "shape" evolution over time, but based on the Continuous Wavelet Transform that uses localized waves (*wavelets*) as basis functions, as opposed to the sines and cosines from the Fourier Transform, and provides another time-frequency representation of the signal where different frequencies have different resolutions in the representation. It is particularly useful for analyzing signals with abrupt changes, transient events, or non-periodic behaviour (*Kumar and Foufoula-Georgiou*, 1997).

STA/LTA algorithm: One of the earliest and simplest methods for event detection is the Short Term Average - Long Term Average Ratio (STA/LTA) (*Withers et al.*, 1998). The idea behind this algorithm is to calculate successive values of ratios between the absolute or squared signal of short-to-long signal windows, define onset and offset thresholds, and compare the calculated values to these thresholds which, when surpassed, mark the beginning and ending of an event as shown (Figure A.1). The successive ratio values form the *characteristic function* and its shape changes with different choices of window size. The window sizes as well as the threshold values need to be tuned depending on applications and data. This method has undergone many modifications and adaptations through the years since its early use in 1965 (*Withers et al.*, 1998), and is now outperformed by some that use other signal information, such as power spectral densities (*Vaezi and van der Baan*, 2015),

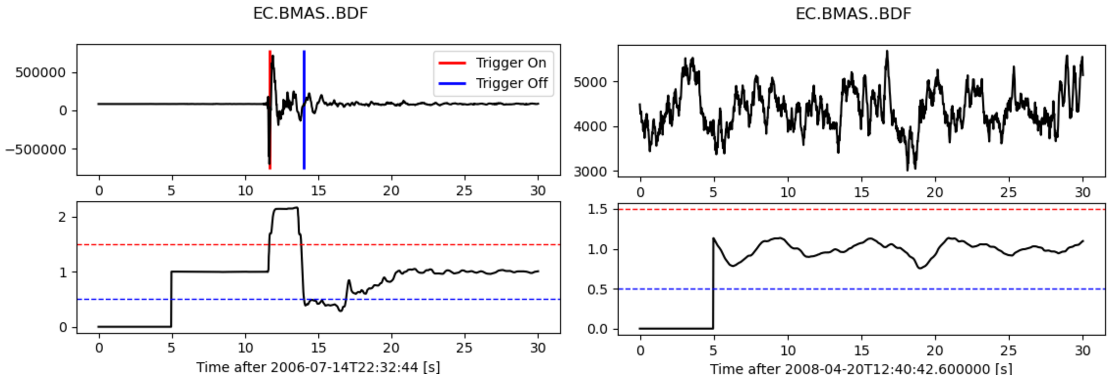

Figure A.1: Example of the STA/LTA algorithm applied to raw acoustic data (upper panels) with their characteristic functions (lower panels). Left: A picked event where the short term average surpassed the long term average. Right: The characteristic function is never higher than the threshold so there is no detection.

in specific contexts. However, it is still used as a first approach for volcano monitoring (e.g., *Spina et al.*, 2021; *Tepp*, 2018; *Malfante et al.*, 2018), landslides (*Bell*, 2018), or for debris flows (*Coviello et al.*, 2019) since it can even be implemented within digitizing units due to its low complexity. Nevertheless, in certain contexts this technique can not be applied and manual counting is still sometimes performed (e.g., *Bell et al.*, 2017; *Schimmel et al.*, 2021).

Dynamic Time Warping: Dynamic Time Warping (DTW) is an algorithm to non-linearly measure the similarity between two sequences by finding the optimal distortion that one sequence has to go through to replicate the other sequence regardless of their onsets, lengths, or varying speeds. The general procedure is to construct a matrix with the distances between each point of a series to each point of the other one (this matrix can be restricted), and finding the optimal path (minimum cumulative distance) so that the first and the last samples coincide between the series by applying a dynamic programming procedure (*Sakoe and Chiba*, 1978).

References

Bell, A. F. (2018), Predictability of Landslide Timing From Quasi-Periodic Precursory Earthquakes, *Geophysical Research Letters*, *45*(4), 1860–1869, doi:10.1002/2017GL076730.

Bell, A. F., S. Hernandez, H. E. Gaunt, P. Mothes, M. Ruiz, D. Sierra, and S. Aguaiza (2017), The rise and fall of periodic 'drumbeat' seismicity at Tungurahua volcano, Ecuador, *Earth and Planetary Science Letters*, *475*, 58–70, doi:10.1016/j.epsl.2017.07.030.

Coviello, V., M. Arattano, F. Comiti, P. Macconi, and L. Marchi (2019), Seismic Characterization of Debris Flows: Insights into Energy Radiation and Implications for Warning, *Journal of Geophysical Research: Earth Surface*, *124*(6), 1440–1463, doi:10.1029/2018JF004683.

Géron, A. (2019), *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, second ed., O'Reilly, CA 95472.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning*, *Springer Texts in Statistics*, vol. 103, Springer New York, New York, NY, doi:10.1007/978-1-4614-7138-7.

Kumar, P., and E. Foufoula-Georgiou (1997), Wavelet analysis for geophysical applications, *Reviews of Geophysics*, *35*(4), 385–412, doi:https://doi.org/10.1029/97RG00427.

Malfante, M., M. Dalla Mura, J.-P. Metaxian, J. I. Mars, O. Macedo, and A. Inza (2018), Machine Learning for Volcano-Seismic Signals: Challenges and Perspectives, *IEEE Signal Processing Magazine*, *35*(2), 20–30, doi:10.1109/MSP.2017.2779166.

Sakoe, H., and S. Chiba (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.

Schimmel, A., V. Coviello, and F. Comiti (2021), Debris-flow velocity and volume estimations based on seismic data, *Natural Hazards and Earth System Sciences Discussions*, *2021*, 1–21, doi:10.5194/nhess-2020-411.

Spina, L., E. Del Bello, T. Ricci, J. Taddeucci, and P. Scarlato (2021), Multi-parametric characterization of explosive activity at Batu Tara Volcano (Flores Sea, Indonesia), *Journal of Volcanology and Geothermal Research*, *413*, 107,199, doi:10.1016/j.jvolgeores.2021.107199.

Tepp, G. (2018), A Repeating Event Sequence Alarm for Monitoring Volcanoes, *Seismological Research Letters*, *89*(5), 1863–1876, doi:10.1785/0220170263.

Vaezi, Y., and M. van der Baan (2015), Comparison of the STA/LTA and power spectral density methods for microseismic event detection, *Geophysical Journal International*, *203*(3), 1896–1908, doi:10.1093/gji/ggv419.

Withers, M., R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, and J. Trujillo (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.

Appendix B

Supplementary Information for "Cleaning volcano-seismic event catalogues: a machine learning application for robust systems and potential crises in volcano observatories"

B.1 Description of the construction of the features proposed in this study

In this appendix we describe how the features proposed in this study are constructed based on Figure B.1 as follows: Frequency Domain (FD) features are extracted calculating the Discrete Fourier Transform (DFT) of the signal, and obtaining the average of the Amplitude Spectrum (red horizontal lines in the leftmost panels) at 7 same-size intervals spanning from 0 Hz to 12.5 Hz. Time Domain (TD) features extraction consists of the following sequence applied to each resulting waveform: 1) Take absolute value of the waveform. 2) Obtain the cumulative absolute sum (CAT). 3) Detrend the standardized CAT. 4) Obtain the envelope using the Hilbert Transform. 5) Extract the 8 coefficients derived from a 7th order Legendre Polynomials fit (red line in the rightmost panels) of the time-normalized envelope for direct comparison between shapes. Add the duration of the events as a separate feature.

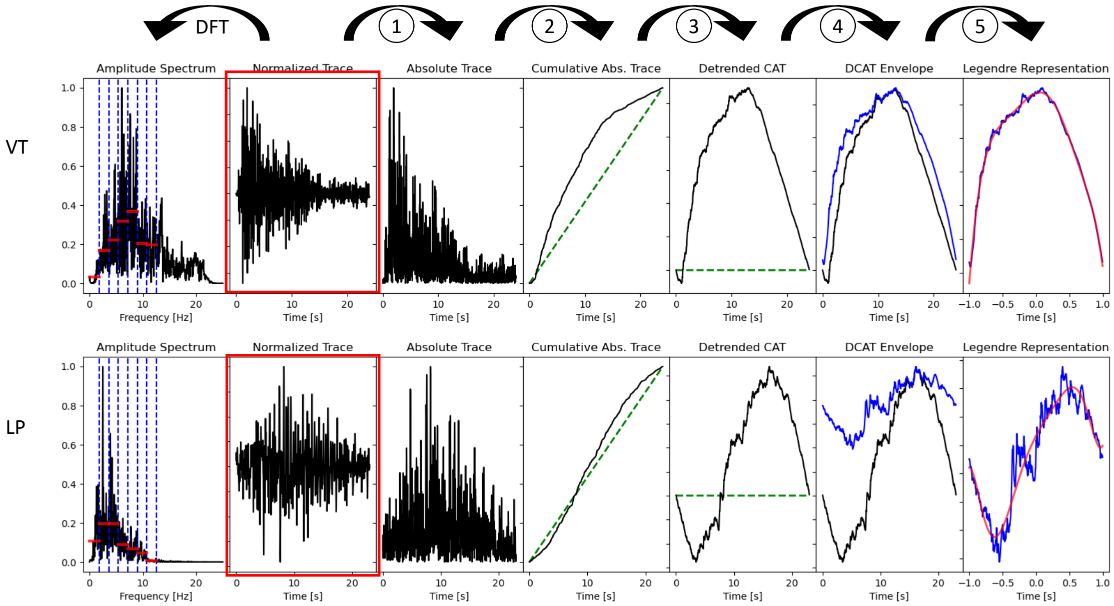

Figure B.1: Schematic describing the steps for calculating the features proposed in this study.

B.2 Correlation between the sets of features used in tis study

In Figure B.2 the absolute values of the Correlation Matrix between features extracted following *Watson* (2020), *Titos et al.* (2018), and those proposed in this study for the Cotopaxi and Llaima volcano datasets are shown. Groups of features from the different studies are separated by white lines. Features that have strong correlations (orange to yellow colours) are usually related because of similar mathematical definitions. For example, the peak and median values in the Frequency Domain (FD) of *Watson* (2020) are strongly correlated with the 20th, 50th and 80th percentiles in the FD of *Titos et al.* (2018) and the mean interval amplitudes proposed in this study. Notice, however, that in general variables are not strongly correlated between groups of features as much as within groups of features. This is specially evident for instance in the linear predicting coefficients (LPCs) of *Titos et al.* (2018) in the Llaima dataset.

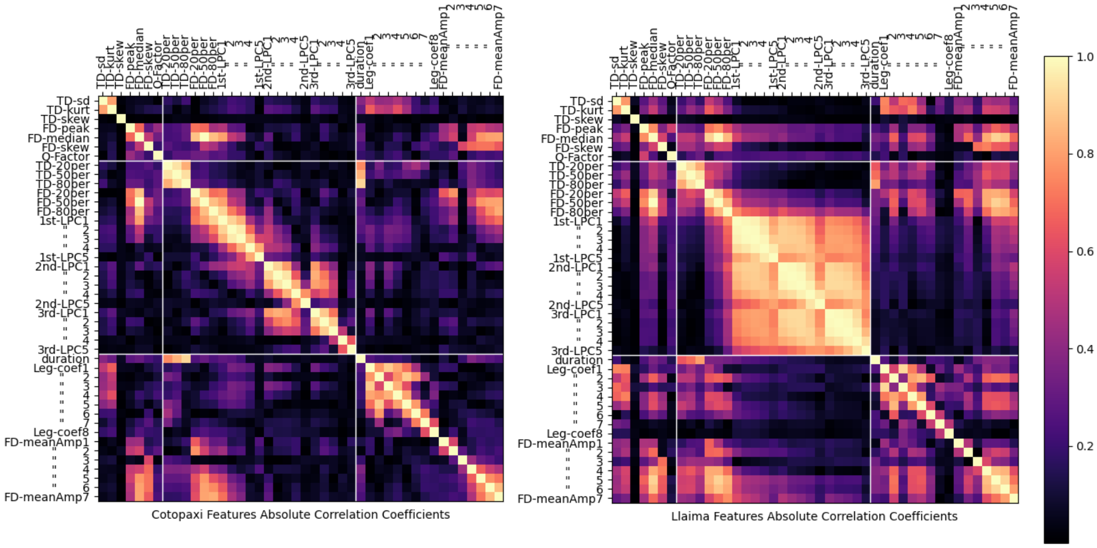

Figure B.2: Absolute correlations between sets of features used in this study.

B.3 Details of the blind re-classification experiment

The results of the blind re-classification experiment are shown in Table B.1 with details of the comparison between Low Quality (LQ) and High Quality (HQ) events' labels with the labels perceived by analysts after providing only raw waveforms. The agreements or disagreements with the original labels are based on the "majority voting" from the analysts' perception. In real world applications only events with low quality should be revised by analysts and changed when in disagreement with the original label.

Event ID	Consistency Score	Original (Quality)	Analyst 1	Analyst 2	Analyst3	RESULT
BREF_0216_LP	0.514286	LP ("LQ")	LP	LP	VT	AGREE
BREF_0547_LP	0.285714	LP ("LQ")	VT	LP	VT	AGREE
BREF_0656_LP	0.657143	LP ("LQ")	VT	LP	VT	DISAGREE
BREF_0449_LP	0.342857	LP ("LQ")	VT	LP	VT	DISAGREE
BREF_0675_LP	0.571429	LP ("LQ")	VT	LP	LP	DISAGREE
BREF_0590_LP	0.542857	LP ("LQ")	LP	LP	VT	AGREE
BREF_0276_LP	0.485714	LP ("LQ")	LP	LP	VT	AGREE
BREF_0489_VT	0.085714	VT ("LQ")	VT	LP	VT	AGREE
BREF_0294_VT	0.257143	VT ("LQ")	VT	LP	VT	AGREE
BREF_0355_VT	0.228571	VT ("LQ")	VT	LP	VT	AGREE
BREF_0439_VT	0.114286	VT ("LQ")	VT	VT	VT	AGREE
BREF_0385_VT	0.542857	VT ("LQ")	LP	LP	VT	DISAGREE
BREF_0303_VT	0.171429	VT ("LQ")	VT	VT	VT	AGREE
BREF_0414_VT	0.257143	VT ("LQ")	VT	LP	VT	AGREE
BREF_0417_VT	0.057143	VT ("LQ")	VT	VT	VT	AGREE
BREF_0001_LP	1	LP ("HQ")	VT	LP	VT	DISAGREE
BREF_0002_LP	0.914286	LP ("HQ")	VT	LP	LP	AGREE
BREF_0005_LP	0.914286	LP ("HQ")	VT	LP	VT	DISAGREE
BREF_0006_LP	0.914286	LP ("HQ")	LP	LP	LP	AGREE
BREF_0003_LP	0.914286	LP ("HQ")	VT	LP	LP	AGREE
BREF_0184_LP	0.914286	LP ("HQ")	LP	VT	VT	DISAGREE
BREF_0072_VT	0.714286	VT ("HQ")	VT	VT	VT	AGREE
BREF_0571_VT	0.742857	VT ("HQ")	VT	VT	VT	AGREE
BREF_0047_VT	0.857143	VT ("HQ")	VT	VT	VT	AGREE
BREF_0343_VT	0.742857	VT ("HQ")	VT	VT	VT	AGREE
BREF_0390_VT	0.8	VT ("HQ")	VT	VT	VT	AGREE
BREF_0455_VT	0.742857	VT ("HQ")	VT	VT	VT	AGREE
BREF_0573_VT	0.885714	VT ("HQ")	VT	VT	VT	AGREE

Table B.1: Results from the blind re-classification experiment.

References

Titos, M., A. Bueno, L. García, and C. Benítez (2018), A deep neural networks approach to automatic recognition systems for volcano-seismic events, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(5), 1533–1544, doi:10.1109/JSTARS.2018.2803198.

Watson, L. M. (2020), Using unsupervised machine learning to identify changes in eruptive behavior at mount etna, italy, *Journal of Volcanology and Geothermal Research*, *405*, 107,042, doi:https://doi.org/10.1016/j.jvolgeores.2020.107042.

Appendix C

Supplementary Information for "Investigating 10 years of volcano acoustic activity at Tungurahua volcano, Ecuador aided by Machine Learning"

C.1 Comparison between different events according to their (dis)similarities

In this section I exemplify how different (dis)similarity criteria may include or exclude similar events if using them separately.

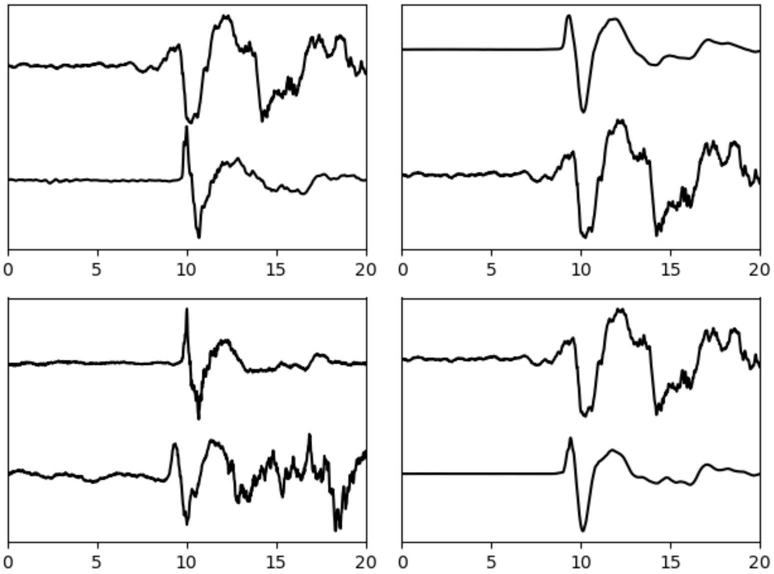

Figure C.1: Examples of events with high correlations (>0.8) yet high Dynamic Time Warping distance.

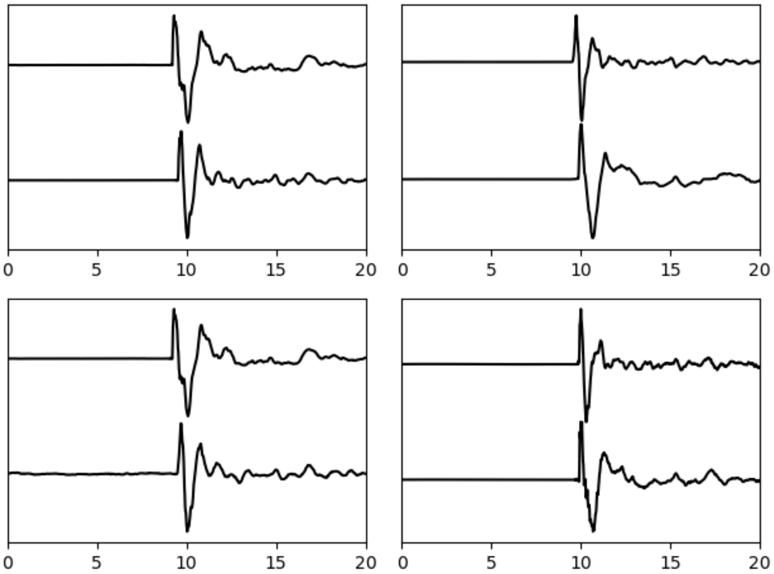

Figure C.2: Examples of events with low correlations (<0.6) yet low Dynamic Time Warping distance.

C.2 Example of a period with complex explosion detection

In this section I show a 20 minute period of acoustic recordings from Tungurahua volcano that demonstrate intense acoustic activity that make explosion detections difficult.

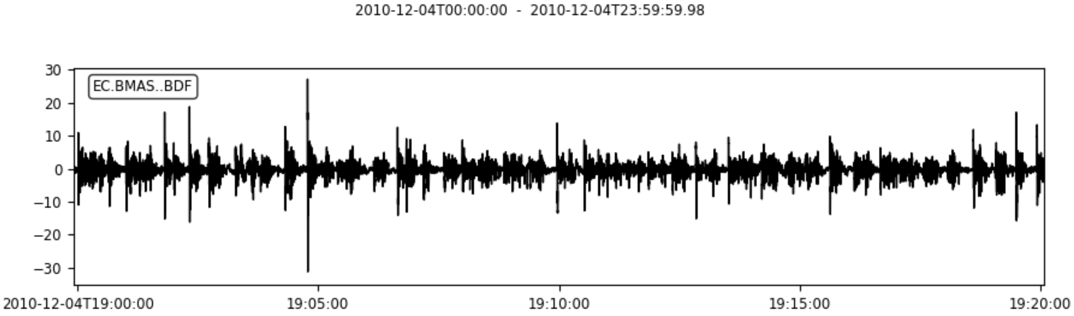

Figure C.3: Example of explosions occurring during a complex emissions period that difficult detection.

C.3 Cross-correlation distance matrix using encoded scalograms

In this section I show how the cross-correlation matrix of the encoded scalograms (Figure C.4) for the same 5 periods studied in this chapter is similar to those obtained from the direct waveforms in Figure 3.7.

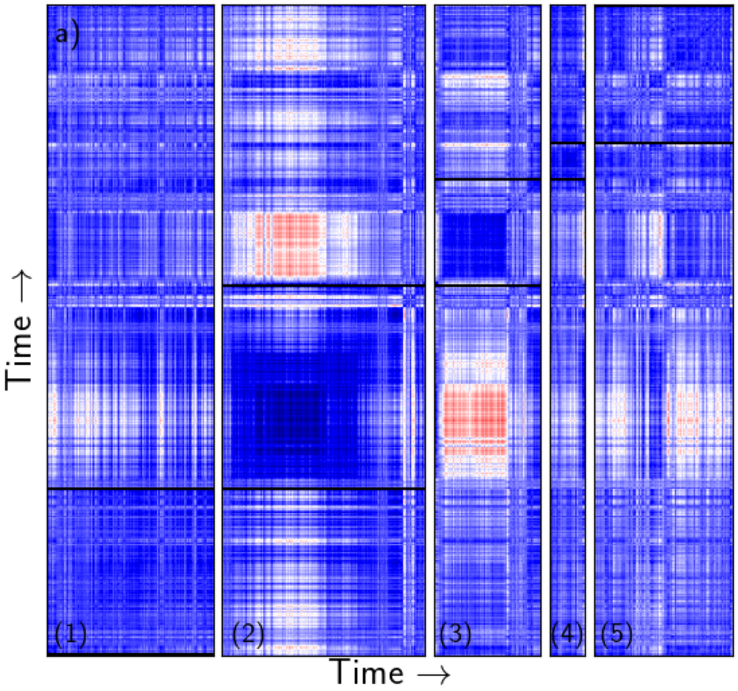

Figure C.4: Cross correlation distance matrix for periods in Table 2 obtained from decoded scalograms.

C.4 Examples of events from different clustered families

In this section I show several examples of events belonging to different clustered families.

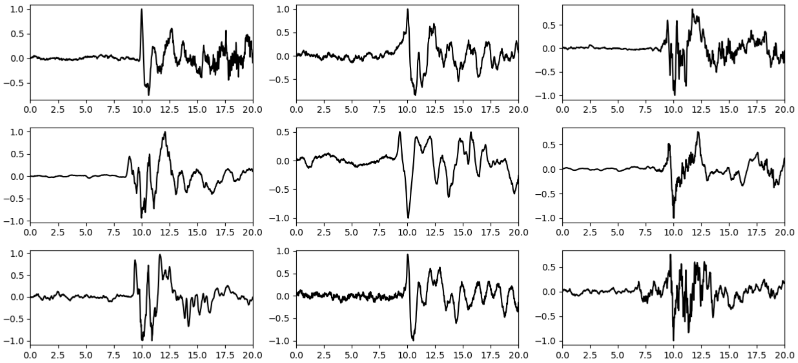

Figure C.5: Examples from family 4A.

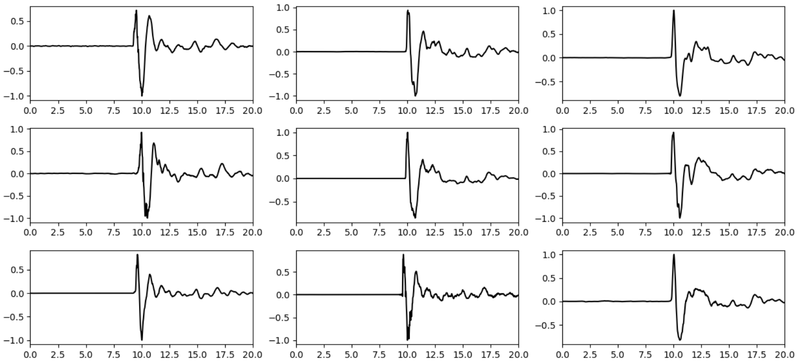

Figure C.6: Examples from family 4B.

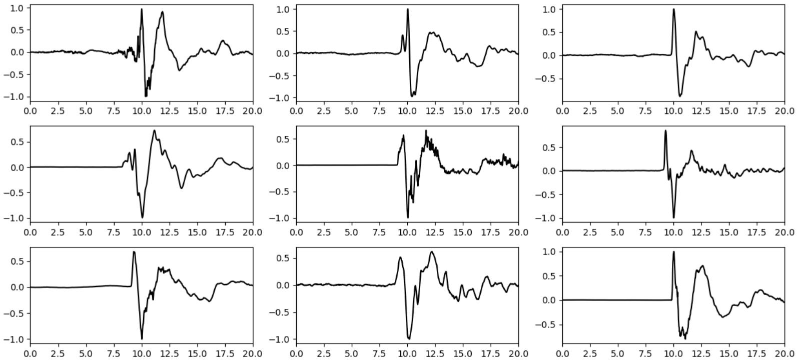

Figure C.7: Examples from family 4C.

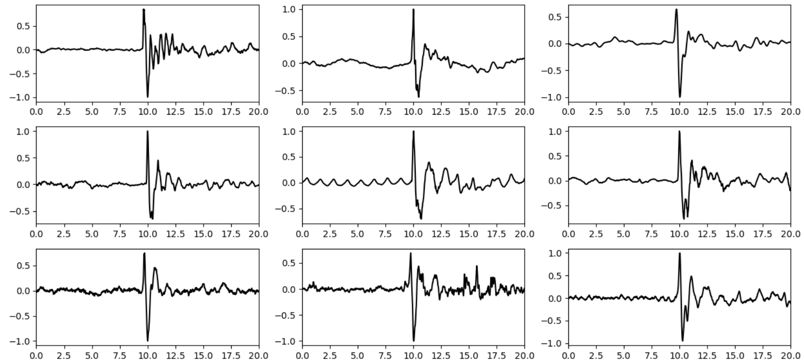

Figure C.8: Examples from family 4D.

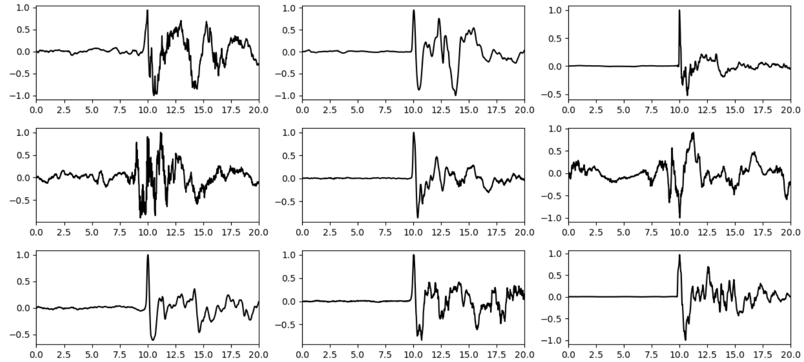

Figure C.9: Examples from family 7A.

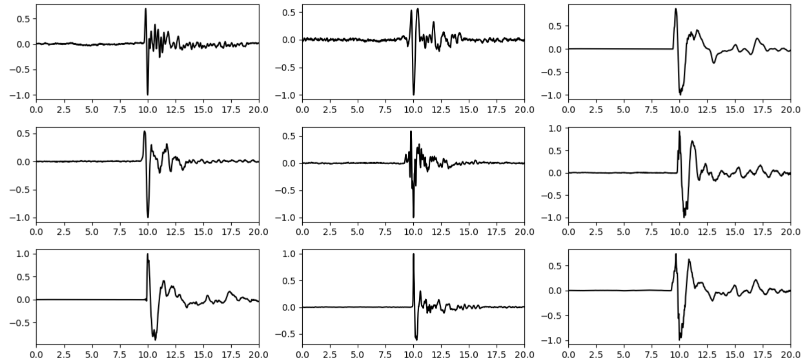

Figure C.10: Examples from family 7B.

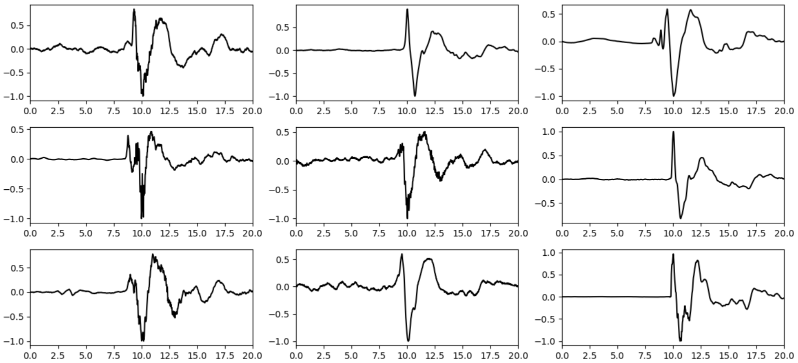

Figure C.11: Examples from family 7C.

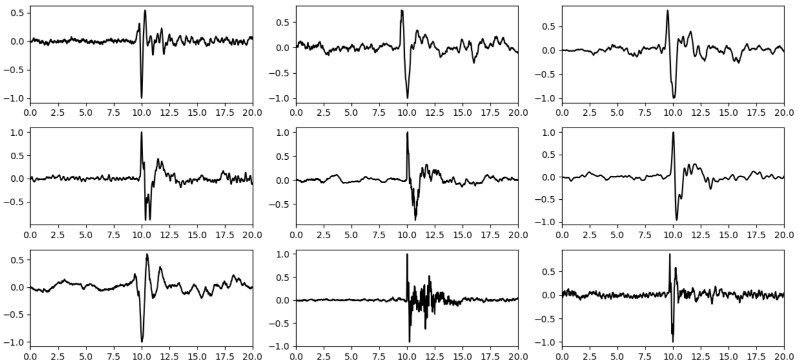

Figure C.12: Examples from family 7D.

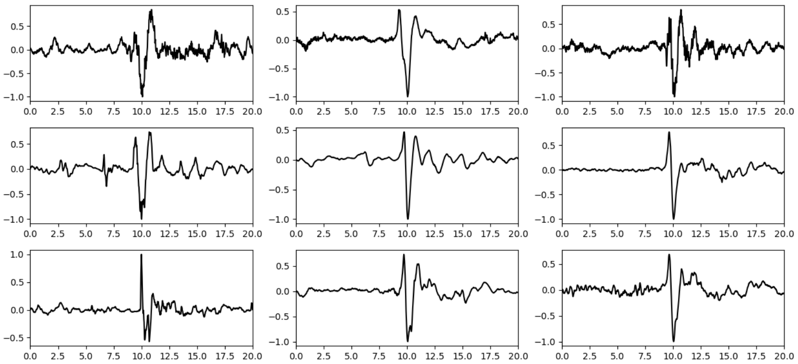

Figure C.13: Examples from family 7E.

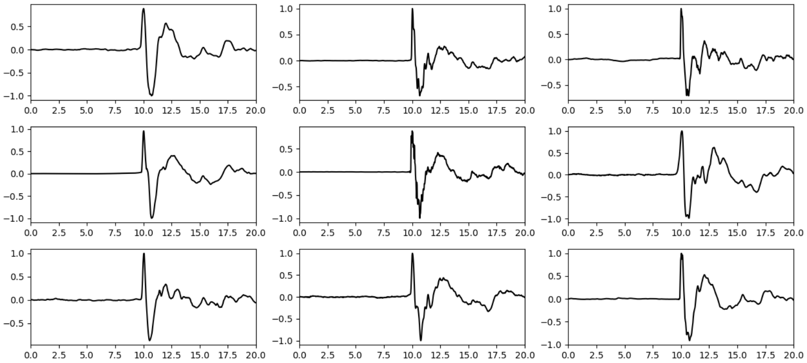

Figure C.14: Examples from family 7F.

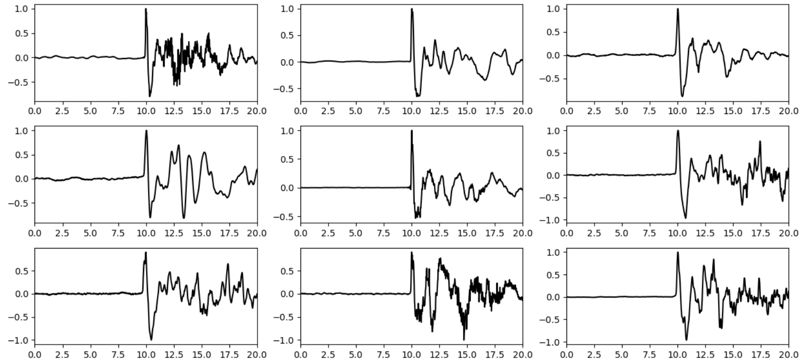

Figure C.15: Examples from family 7G.

Appendix D

Supplementary Information for "Tracking changes in natural streams using acoustic data: The 2021 Western North America heat wave at the Squamish River, BC, Canada"

D.1 Different Statistics for the filtered acoustic amplitude

For the analyses of Chapter 4, we explored various statistics to track water discharge rates. All the statistics show similar behaviour, but the most stable one resembling discharge rates was the 10^{th} percentile. This is seen in Figure D.1 by scaling all the series so that their amplitudes become comparable. The 10^{th} percentile is the less variable one and thus is the selected statistic to track water discharge rates.

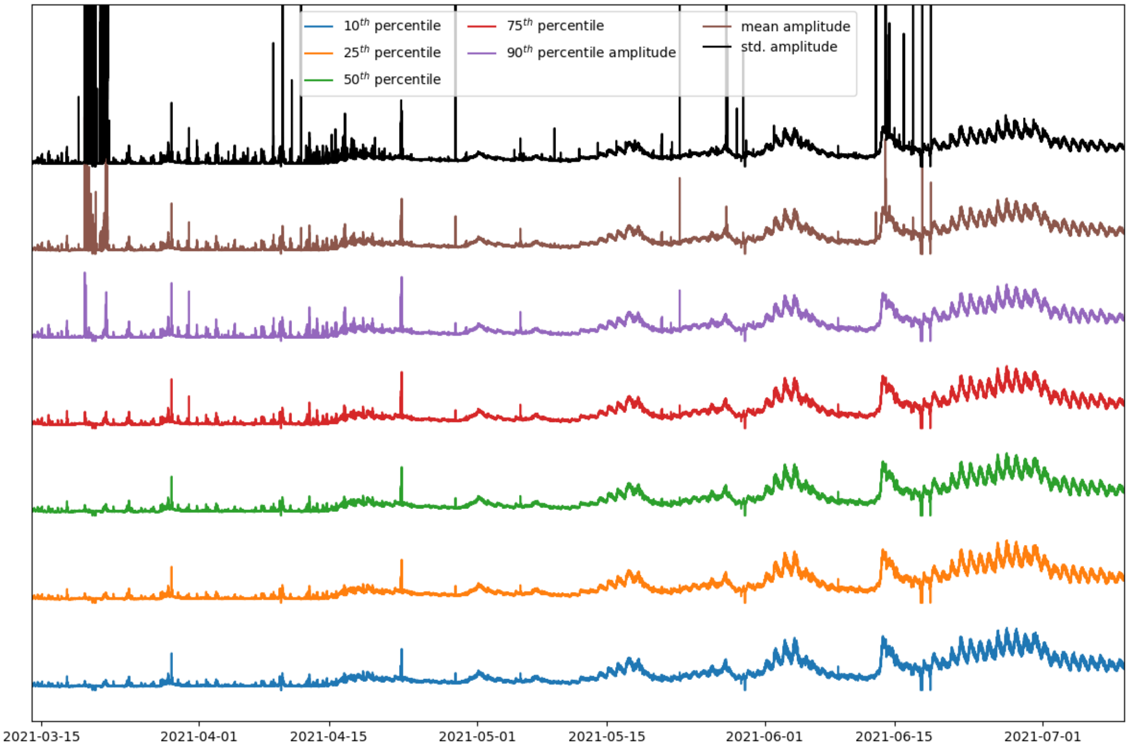

Figure D.1: Different statistics for the filtered $3 - 8Hz$ acoustic amplitudes.

D.2 Comparison of Discharge Rates and Water Levels

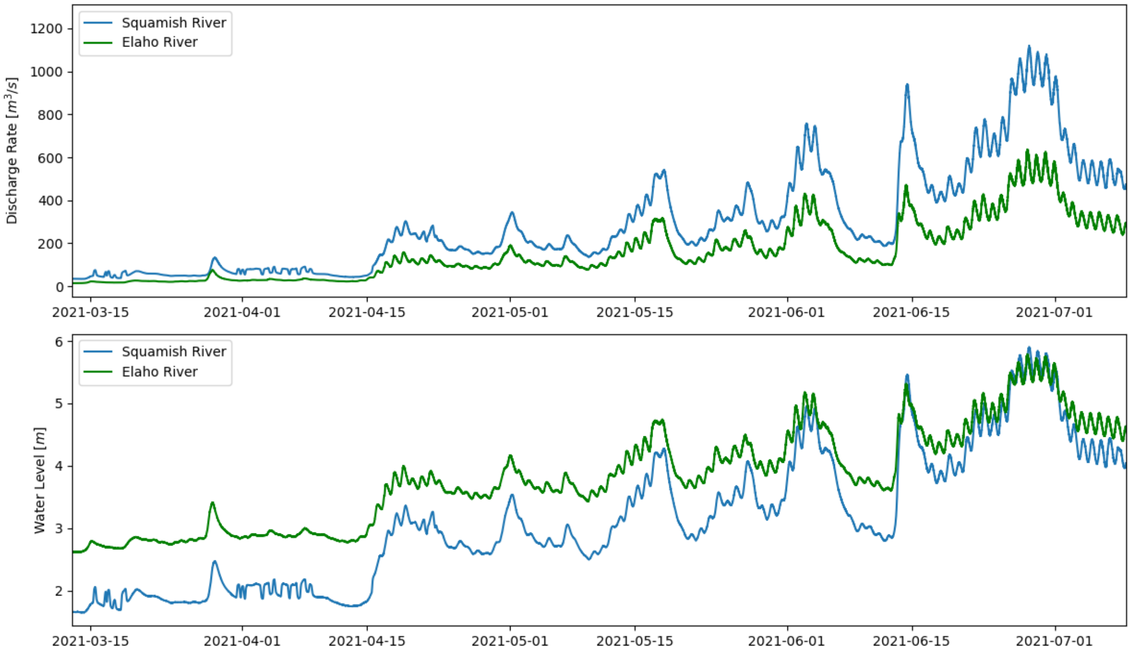

Figure D.2: Comparison between water discharge rates and water stages for the Squamish and Elaho Rivers.

D.3 Unmasked spectrogram during the study period

The spectrogram in Figure 4.11 was masked due to intense noise that prevented the visualization of the spectral intensities. In Figure D.3 we show how the complete figure without masking and the other spectral intensities are dimmed.

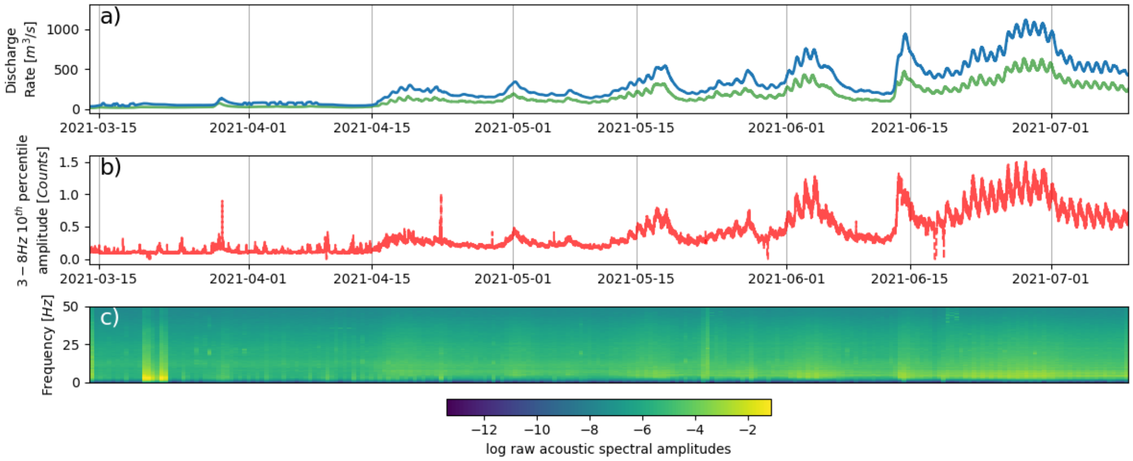

Figure D.3: (a) Discharge Rates and (b) acoustic Amplitudes during the study period with (c) unmasked spectrogram.