Discrimination of Icequakes and Earthquakes in Southeast Alaska using Random Forest and Principal Component Analysis

Kharita Akash[1]

[1]Indian Institute of Technology Roorkee

November 16, 2022

Abstract

Seismic event classification can be challenging in the regions where different types of seismicity overlap in space, time, and magnitude. In this paper, I evaluate the performance of a supervised machine learning technique called Random Forest for the discrimination of icequakes and earthquakes in southeast Alaska at 15 stations surrounding the region. I train the Random Forest on about 3000 icequakes and earthquakes that occurred in the region over the last 17 years. For each event, absolute frequency spectrum values are considered as input features. The accuracies at different stations range from 75 to 95% with an average of about 90%. I conducted tests for selecting the optimum number of decision trees in the RF model and compared the results obtained by applying bandpass filters of different frequency bands on input waveforms. I further experiment by reducing the dimensions of input features by applying Principal Component Analysis (PCA), and conducted test for selecting the minimum number of components and the frequency band that gives the best results. The application of PCA resulted in slightly better results and a final model that gave the best results among all the tests was chosen. The accuracy results of the final model were further analyzed with respect to the amount of available dataset, the average distance of a station from all the glaciers, and the local geology.

Discrimination of Icequakes and Earthquakes in Southeast Alaska using Random Forest and Principal Component Analysis

Akash Kharita[1,*]
1. Indian Institute of Technology Roorkee
* - akharita1999@gmail.com

Abstract

Seismic event classification can be challenging in the regions where different types of seismicity overlap in space, time, and magnitude. In this paper, I evaluate the performance of a supervised machine learning technique called Random Forest for the discrimination of icequakes and earthquakes in southeast Alaska at 15 stations surrounding the region. I train the Random Forest on about 3000 icequakes and earthquakes that occurred in the region over the last 17 years. For each event, absolute frequency spectrum values are considered as input features. The accuracies at different stations range from 75 to 95% with an average of about 90%. I conducted tests for selecting the optimum number of decision trees in the RF model and compared the results obtained by applying bandpass filters of different frequency bands on input waveforms. I further experiment by reducing the dimensions of input features by applying Principal Component Analysis (PCA), and conducted test for selecting the minimum number of components and the frequency band that gives the best results. The application of PCA resulted in slightly better results and a final model that gave the best results among all the tests was chosen. The accuracy results of the final model were further analyzed with respect to the amount of available dataset, the average distance of a station from all the glaciers, and the local geology.

Keywords – Icequake, Earthquake, Random Forest, Principal Component Analysis, Alaska

1. Introduction

Since the last decade, the amount of available seismic data has grown exponentially in volume and variety. Both the coverage and density of many seismic networks have increased. For example – According to Incorporated Research Institutions in Seismology Data Management Centre's (IRIS DMC) statistics report (https://ds.iris.edu/data/distribution/), the seismic data archive of IRIS DMC has

registered a growth from 100 Tebibytes (TiB) in 2009 to 800 TiB in 2022. As the technology has progressed, instruments with better sensitivity and comparatively low cost are being developed and deployed (e.g. Anthony et al., 2019). This increasingly available data has been very advantageous for improving our understanding of earthquake source processes as well as the earth's structure. However, it also comes with a plethora of challenges. As the sensitivities of recording instruments have been improved, comparatively more quantity and types of events are being detected, which makes the task of manual processing, and interpretation difficult. Seismological observatories around the world have begun to use automatic/semi-automatic methods to facilitate the detection, processing and interpretation of the available data. These methods often take advantage of growing computational infrastructure and are developed with the aim of being time and cost-effective, accurate, and suitable for big datasets. However, there are often different kinds of challenges that hamper the accuracy and effectiveness of these methods.

One of the main problems that seismologists face is of discriminating tectonic events such as earthquakes from non-tectonic events that have different source processes but generate similar seismic waveforms. These non tectonic events could be either natural (e.g. icequakes, volcanic tremors) (Dahm & Brandsdóttir, 1997; Qamar, 1988b) or man-made (e.g. nuclear explosions, quarry blasts, mining induced earthquakes)(Gitterman et al., 1998; Tibi et al., 2018; Zhao et al., 2015) or both. While these events are usually small in magnitude and shallow in depth, naturally occurring earthquakes can be of any size and occur at any depth ranging from near surface to the base of the upper mantle (700 Km) depending on the seismo-tectonic settings and the orientation of prevailing stress fields in the area. It becomes a complicated task to discriminate non-tectonic events from earthquakes in the regions where they overlap in space, time, and magnitude. If incorrectly classified, non-tectonic events may accumulate in the earthquake catalogs that contaminate their quality and result in erroneous estimates

of rates of seismicity and consequently, of seismic hazard (Astiz et al., 2014; Gulia, 2010; Mackey et al., 2003; Marzen et al., 2021)

Discriminating non-tectonic events from tectonic events has been a subject of intensive studies for a long time. In particular, the topic of discrimination of man-made explosions from earthquakes of comparable magnitudes has garnered special attention (Koper et al., 2016; Kuyuk et al., 2011; O'Rourke et al., 2016; Stump et al., 2002; Zeiler & Velasco, 2009). Considering the seismic signatures of nuclear tests, mining explosions and quarry blasts are very similar, It is important to discriminate and identify the origin of a shallow seismic event both at local and teleseismic distances to ensure the effective implementation of the Comprehensive Nuclear-Test-Ban Treaty(Bowers & Selby, 2009). Similarly, in volcano seismology, the ways for accurate classification of different types of volcanic seismicity have been researched for a long time (Hibert et al., 2014; Maggi et al., 2017b). In view of this, both the statistical and machine learning based methods have been developed and applied; these methods usually involve deriving a set of features from the event waveforms as a first step and then applying different classification techniques to these features. Statistical approaches include discriminating the events based on the ratio of amplitudes of different seismic phases (Rodgers & Walter, 2002; Taylor, 1996; Kim, 1997; Walter et al., 2018), high and low frequency spectral amplitudes (Walter, 1995; Wang et al., 2021), misfits of P-wave spectra to standard earthquake source model(Allmann et al., 2008), and different kind of magnitudes(Holt et al., 2019; Koper et al., 2016; Wang et al., 2021). While statistical approaches have shown a great promise for the events at teleseismic distances and the events with moderate to high magnitudes, their efficiency plunges at local to regional distances and for events of small magnitude (O'Rourke et al., 2016; Pyle & Walter, 2019).
Recently, Machine Learning methods have been increasingly applied to seismic event discrimination problems. Both the supervised methods and unsupervised methods have been applied. These methods commonly involve the extraction of features from the waveforms that distinguish different types of

events as a first step and then finding a classification boundary governed by the chosen machine learning algorithm. The extracted features can be physics-based (e.g. Magnitude, P/S Spectral ratio etc.) (Falcin et al., 2021; Dowla, 1990; Hammer et al., 2013; Maggi et al., 2017a) or automatically extracted features with no apparent physical meaning (Kong et al., 2019; Linville et al., 2019; Tibi, 2021). Examples of supervised machine learning methods include the use of Convolutional and Recurrent Neural Networks (Beyreuther & Wassermann, 2008; Lara et al., 2021; Linville et al., 2019; Tiira, 1996), Self Organizing Maps and Support Vector Machines(Hammer et al., 2013; Kortström et al., 2016; Masotti et al., 2006) while examples of unsupervised methods include the use of k-means, Principle Component Analysis and Gaussian Mixture Models(Kuyuk et al., 2011, 2012).

Typically the performance of a machine learning technique improves with an increased training dataset. This can cause challenges in the region with sparsely available datasets. Further, if the models are trained on the automatically extracted features they cannot be generalized and may be effective only in a specific region of interest (Zeiler & Velasco, 2009).

In this study, I evaluate the performance of a supervised machine learning algorithm called Random Forest (RF) for discriminating icequakes and earthquakes located within 50 km from the Columbia glacier in southeast Alaska at 15 broadband seismic stations located within 100 km from the glacier (Fig. 1).

RF algorithm has proven to be very effective in the problems of seismic event detection and classification. Previously, It has been used for detection and classification of different kind of volcanic events (Dempsey et al., 2020; Clément Hibert et al., 2017; Maggi et al., 2017b), landslides(Rubin et al., 2012), geysers (Yuan et al., 2019)and aftershocks(Aden-Antoniów et al., 2022). I used the values in the absolute frequency spectrum as input features and experimented by applying filters in different frequency bands to select the frequency band that gives the best results. I further experiment by reducing the dimensions of the features using Principal Component Analysis (PCA) to compare the

model performance with the original model. My analysis revealed that application of the PCA improved the performance of our original model and significantly decreased the computation time. The accuracy at different stations ranges from 75 to 96% and it is influenced by the amount of training dataset, distance from the glaciers, and the effects of the local geology between the event and the receiver. Since I choose the values in the absolute frequency spectrum as input features, my model can be used as a general effective tool to discriminate icequakes from earthquakes at any glacier.

2. Icequakes in Southeast Alaska

Icequakes at four glaciers in southeast Alaska are considered in this study – the Columbia, Meares, Yale, and the Harvard Glacier. Columbia glacier is a 51 km long, temperate, tidewater glacier, located approximately 30 km west of Valdez in southeastern Alaska (Fig. 1(a)). It descends from the height of about 3050 meters down the flanks of Chugach Mountains and ultimately flows into Prince William Sound via a narrow inlet. It consists of two initial branches – one smaller branch that lies west of the Great Nunatak peak and the bigger main branch that lies east to the peak. These branches merge to form one larger branch which terminated near the northern edge of Heather Island until 1980, following which it began retreating rapidly (Meier & Post, 1987). Before the 1980s, the glacier was held at a stable position by the shoreline on one end and the underwater moraine – the accumulated debris carried and deposited by the glaciers, on the other end. As the glacier retreated off the moraine, probably due to the initial nudge provided by climate changes, it freed from the moraine, and the icebergs started calving off the glacier. The retreat of this glacier continues to the present day, though at an uneven pace. Satellite images and airborne altimetric measurements (https://earthobservatory.nasa.gov/world-of-change/ColumbiaGlacier) show that there has been a huge loss in volume of this glacier as it retreated. This single glacier is considered to be responsible for about 50% of snow loss in the Chugach mountains. Snow is mainly lost through the shedding of large chunks

of icebergs caused by the calving events. Over the years, the retreating terminus progressively thinned and the Columbia glacier is now split into two separate glaciers corresponding to two initial branches with calving now occurring at two different fronts. (Enderlin et al., 2018; Post et al., 2011)

Meares Glacier is a 10 km long tidewater glacier at the head of the Unakwik Inlet that connects Chugach National forest to Prince William Sound (Fig. 1(a)). Unakwik Inlet is often known to be nearly ice-free which makes it an ideal location for glacier visitors. Meares is currently one of the two advancing glaciers that flow in the Prince William Sound, the other one being the Harvard glacier (Trabant et al., 2002). Yale glacier is a 32 km long glacier immediately west of the Meares and is separated from the Harvard glacier by the College fjord (Figure 1(a)). While Yale glacier has been retreating since the early nineteenth century with varying retreating rates, Harvard Glacier has been advancing since 1905 and possibly earlier (Sturm et al., 1991). The striking contrast between the terminus behavior of the Yale and Harvard glacier that parallel the same fjord and derive from the same snowfield, suggest that the terminus behavior is more likely the result of dynamic controls related to the fjord depth, ice thickness, and calving rate with climate change playing a secondary role (Sturm et al., 1991).

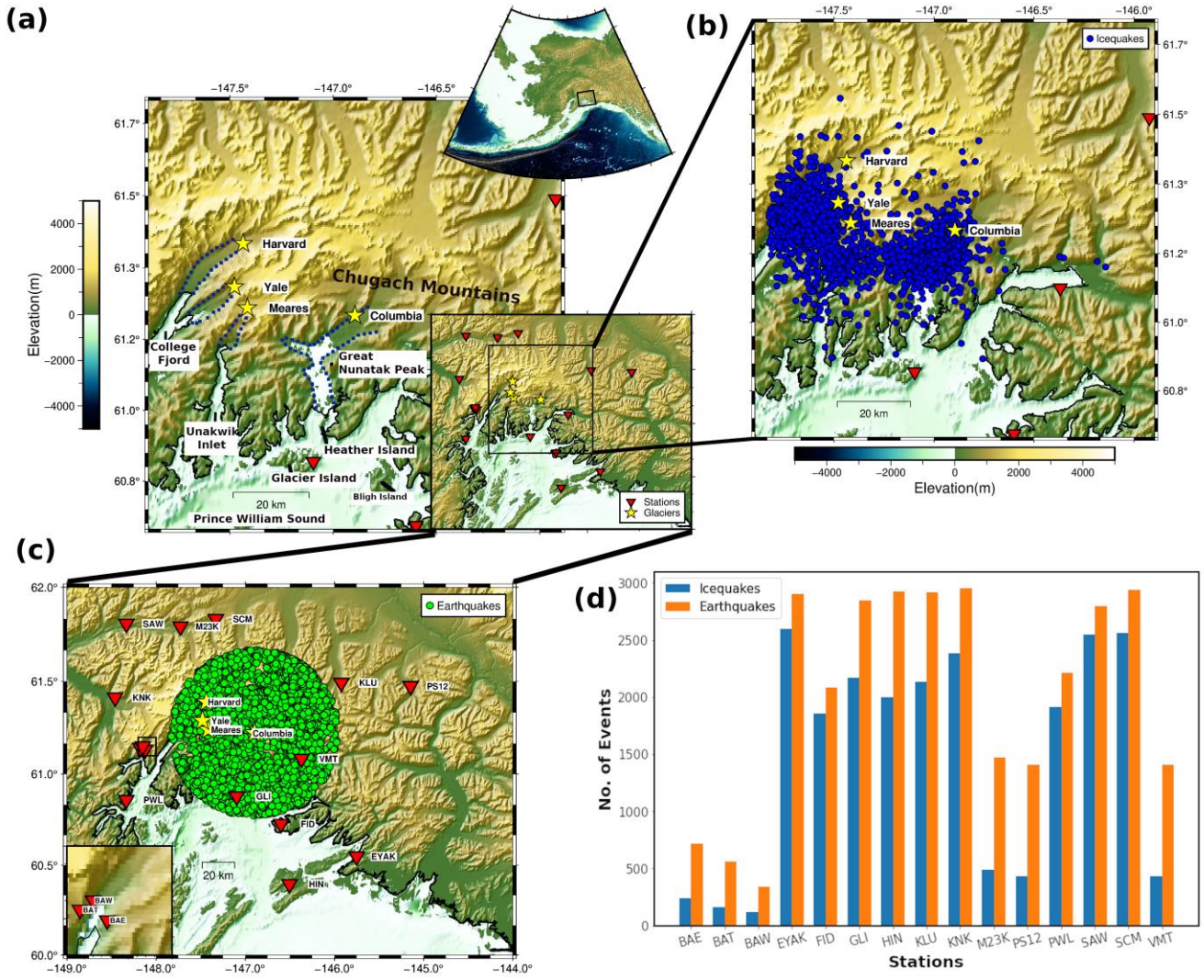

Figure 1: (a) Map of the glaciers along with their flowing paths, (b) location of all the icequakes, (c) location of all the earthquakes and the recording seismic stations considered in the study and, (d) number of icequakes and earthquakes available at each station for the period considered in the study.

Over the last decade, the study of Icequakes has emerged as an important tool for monitoring the terminus behavior of glaciers and the rise in seawater levels (Bartholomaus et al., 2012; Podolskiy & Walter, 2016; West et al., 2010). Icequakes in southeast Alaska have been very well documented and they generate from a wide variety of mechanisms including basal sliding, calving, and surface

crevassing (Neel et al., 2007; Podolskiy & Walter, 2016; Qamar, 1988b). The icequakes caused by basal sliding or calving events contain mostly low dominant frequencies and are readily distinguishable from the surface crevassing events that generate high frequency waveforms. In this research, I focus on low frequency icequakes mostly caused by the calving of the icebergs at the terminal of a retreating glacier where it enters the ocean although recent research has demonstrated that significantly big icequakes can even occur inland due to calving in the presence of lakes formed by the meltwater trapped between glaciers and their moraines. Icequakes show high seasonal dependence with their numbers ramping up during the summer season and falling in the winter season. The duration and the size of these events vary depending on the duration of the calving and the size of the calved icebergs. A calving event and correspondingly its seismic signal can last anywhere between the 30s to over 20 minutes(Bartholomaus et al., 2012; Neel et al., 2007; Podolskiy & Walter, 2016). However, regardless of style, size and duration, calving events show some common spectral characteristics such as (i) emergent mostly monochromatic waveforms, (ii) weakly developed P and S phases, and (iii) low dominant frequencies between 1-3 Hz compared to earthquakes of similar magnitudes (Fig. 2) (Neave & Savage, 1970; Neel et al., 2007; Qamar, 1988a; West et al., 2010).

Earthquakes are often automatically detected by the ratio of the short-term and long-term average amplitude of the seismogram. The arrival of an earthquake is often marked by a sharp increase in amplitude in the form of a P-wave and therefore whenever the STA-LTA ratio exceeds a user-defined threshold, it would indicate the detection of an event. However, in the presence of emergent waveforms such as those produced by icequakes, the STA-LTA detection may not work. A frequency domain power spectral density based icequake detector was developed by Neel et. al, (2007) to detect icequakes. The detector consists of the computation of power spectral densities from windows of time series overlapped by 50%, filtered in different frequency bands i.e. 10-20 Hz, 1-3 Hz, and 0.0833-0.033 Hz, and then a statistic (mean, median or standard deviation) is computed, if that statistic passes a user-

defined value, the event is detected. This detection method is validated by the evidence from the correlation of the detected event with the changes in terminus geometry, correspondence of seismic data with visual records of calving, and location of hypocentres of detected events near the glacier terminus. Icequakes were mostly detected in the 1-3 Hz filter band. However, even detections in this band are often contaminated by atleast10-20% from the events that are not generated by calving. These could be local, regional, or teleseismic earthquakes, events caused by basal sliding and hydraulic transients.

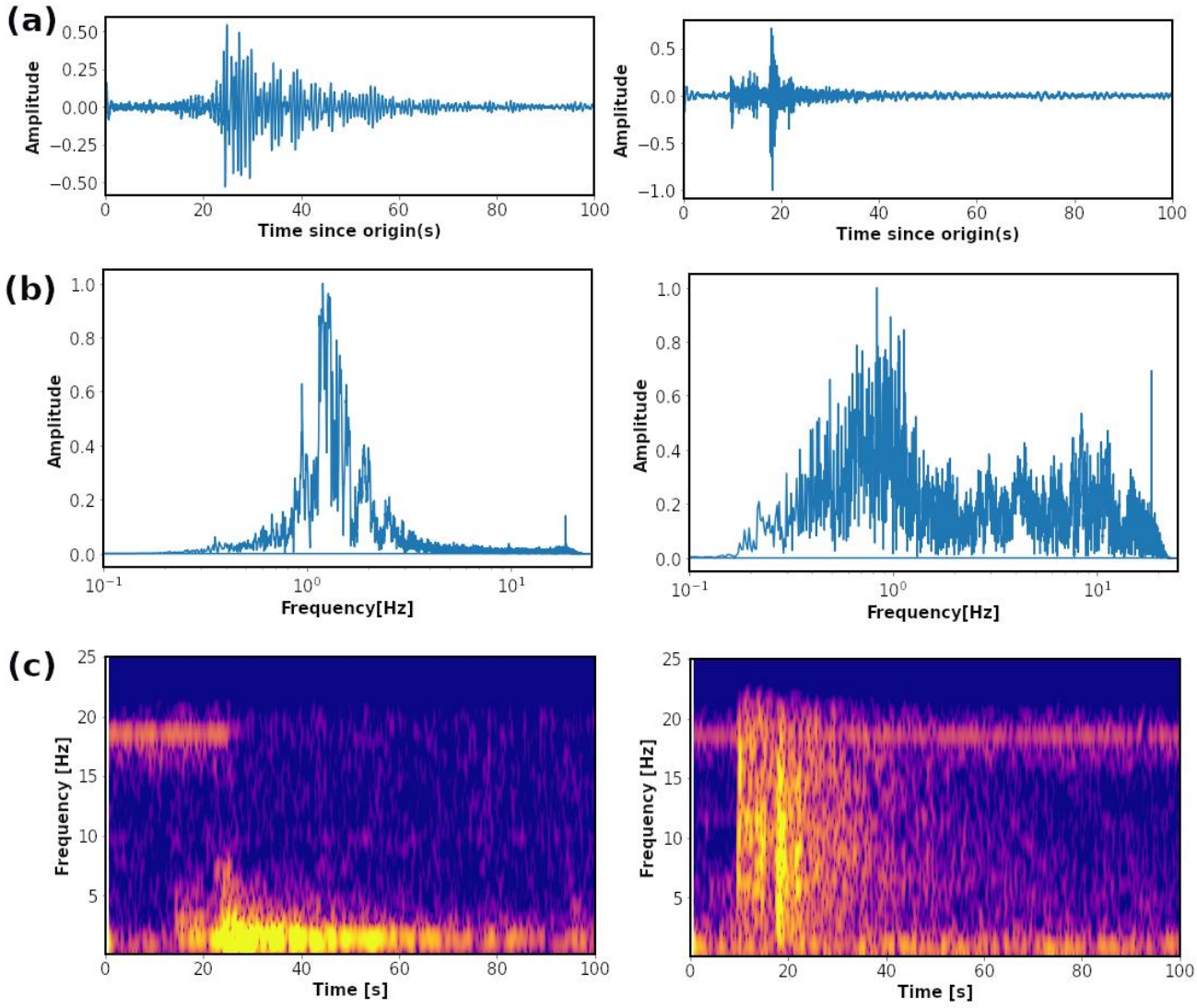

Figure 2: Major differences between icequake and earthquake. (a) Normalized waveforms, five minutes in duration since the origin of the event (b) Normalized Frequency spectrum and (c)

Spectrogram of both the events. The events are of same magnitude and occurred at similar distance from the recording station.

In this study, all the available waveforms, five minutes in duration since the origin time of 2650 icequakes that occurred within the radius of 50 km from the center of the Columbia glacier in the period of about 17 years between 2005-04-07 and 2022-04-22 were downloaded from the IRIS Data Management Centre for 15 broadband seismic stations lying within one degree radius from the centre of the Columbia glacier (Fig. 1(b)). To maintain a well-balanced dataset, similar data was downloaded for the latest 3000 earthquakes out of a total of 12283 earthquakes that were obtained in the catalog provided by the United States Geological Service Advanced National Seismic System service (Fig. 1(c)). The magnitude of earthquakes was kept in the range of zero to three, similar to the magnitude range of icequakes. The number of icequakes and earthquakes varies at different stations due to uneven availability and period of operations (Fig. 1(d)). All the seismic stations are broadband with a sampling rate of 50 Hz and are part of 'AK' network which is maintained by the Alaska Earthquake Center (Fig. 1(c)).

3. Methods

The problem statement of this research study is simple – each automatically detected event needs to be classified either as an earthquake or an icequake i.e. it is a binary classification problem. My approach to solving this problem is - At each seismic station, firstly, train a random forest classifier on certain features extracted from the waveforms, after training, the random forest algorithm will learn the boundaries in the feature space that distinguish the two type of the events and secondly, apply the trained model on the test dataset and analyze the model performance under various conditions.

I use the input features as all the values obtained in the normalized absolute frequency spectrum since it is one of the important characteristics that distinguish icequakes from earthquakes. So if a waveform is

of duration '*t*' seconds and the sampling rate is '*n*' Hz. Then for each waveform, we have *nxt* features. if we have a total of *l* events. Then the feature matrix to the random forest classifier algorithm can be written as a matrix of shape (*l*, *nxt*) and the output (label) matrix can be written as a vector of shape (*l*, 1). Priliminary experiments showed that accuracy results were approximately the same for different components, I chose to proceed with the waveforms from the E component because it effectively captures surface waves and have higher signal-to-noise ratio. Further, I chose the duration of the waveforms used to be of five minutes from the origin of the event as It accounts for varying event-station distances and ensures most of the seismic energy has arrived at the stations. Given the sampling frequency at each station considered in this study was 50 Hz, we have 300x50 = 15000 features for each event (Figure 3).

A summary of the workflow is shown in Figure 3. For a given station, the total amount of available data was split into training and test data with a ratio of 7:3. I trained a random forest classifier on these features and analyze the performance for different number of decision trees in the random forest, under application of bandpass filter in different frequency bands i.e. 1-20, 1-10 and 3-10 Hz, these frequency bands were chosen after considering the spectral properties of icequakes and earthquakes, ambient and anthropogenic sources of noise (Neel et al., 2007; Podolskiy & Walter, 2016). Model performance here refers to the accuracy which is defined as the number of event predicted correctly divided by the total number of events. After determining the optimum number of decision trees and selecting the frequency band that gives the most accuracy. Since 15000 features are large number of features and many of these features may not be important at all, I repeat the same experiment but by reducing the dimensions of features using PCA to see whether using the reduced number of features give comparable results to the original model or not. After selecting the best performing model from the above experiments, I compared its performance at different stations with different amount of training datasets, with the average distance from the center of the four glaciers considered in the study and with the propagation medium between events and station.

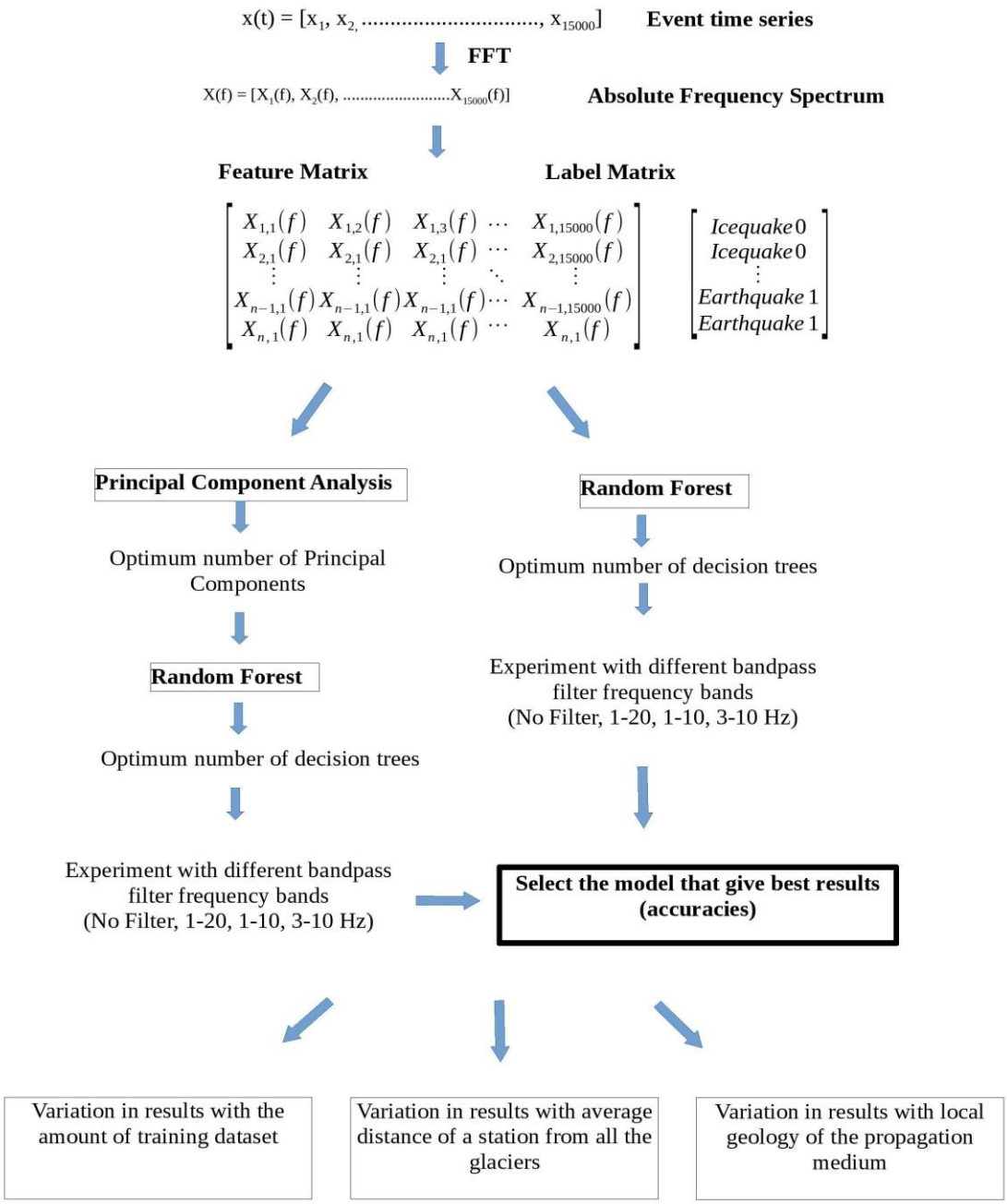

Figure 3: Workflow illustrating the steps for the experiments performed for the selection of the final model. Once the final model is selected, Results are analyzed with respect to the average distance of station from the glaciers, amount of the available dataset, and the local geology.

3.1 Random Forest

Random forest classification involves taking the most frequent decision of an ensemble of the decision trees to avoid overfitting issues (Breiman, 2001). A decision tree is a model that classifies the data by greedy selection of the best split point at each step. however, individual decision trees suffer from the problem of high variance which makes them work well only for the specific dataset. To reduce this variance, multiple models from the random subsets of the dataset can be build and their average decisions can be taken, a technique called bootstrap aggregation or bagging, but the trees that form can be highly correlated meaning very similar splitting points can be chosen in each tree, making different trees very similar which defies the original purpose (Hastie et al., 2001). The Random Forest algorithm involves a tweak in the bagging algorithm by constraining the number of features that decision trees can evaluate at each point. This ensures the trees are uncorrelated. RF can also determine the relative importance of the features based on how often they appeared in decision trees (Breiman, 1996). The RF algorithm can be summarised in the following steps. First, subsets of the samples are randomly selected with replacements from the training dataset, a decision tree is trained for each subset. Only a random subset of features are selected. Then each decision tree predicts the random subset of the testing dataset and then for each sample in the test dataset, the most frequent prediction is selected as the final prediction.

3.2 Principal Component Analysis

Principal Component Analysis is one of the most commonly used feature extraction technique. Feature extraction refers to extracting the new set of features from the existing ones such that the newly formed set of features hierarchically captures most of the information stored in the original set of features. Sometimes, when there are large number of features, it may lead to the problem of overfitting. Moreover, many features may be correlated or redundant and may not necessarily contribute in the classification. In these cases, it is often advantageous to reduce the number of features (also known as

Dimensionality) by applying feature extraction techniques (or Dimensionaliy Reduction) and selecting the small number of features that capture most of the information stored in the original dataset.

In principal, PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system in a way that each new coordinate successively inherit maximum variance of the data. (Wold et al., 1987)

Let **X** be a *n*x*s* matrix consisting of '*n*' series, each series containing '*s*' elements. The rows of **X** are individual series and i^{th} column contains the i^{th} element of each series, Then PCA involves the transformation of matrix **X** into matrix **T** of size *n*x*l* such that

$$\mathbf{T} = \mathbf{Xw} \qquad \text{(i)}$$

where **w** represents a transformation matrix of size *s*x*l*, each column of **w** represents a weightage vector and is constrained to be of unit magnitude. Equation [1] can be written in the matrix form as

$$\begin{bmatrix} T_{11}T_{12}.T_{1l} \\ T_{21}T_{22}.T_{2l} \\ \dots. \\ T_{n1}T_{n2}.T_{nl} \end{bmatrix}_{nxl} = \begin{bmatrix} X_{11}X_{12}.X_{1s} \\ X_{21}X_{22}.X_{2s} \\ \dots. \\ X_{n1}X_{n2}.X_{ns} \end{bmatrix}_{nxs} \begin{bmatrix} w_{11}w_{12}.w_{1l} \\ w_{21}w_{22}.w_{2l} \\ \dots. \\ w_{s1}w_{s2}.w_{sl} \end{bmatrix}_{sxl} \qquad \text{(ii)}$$

The first column of **T** is known as first principle component, the second column represents second principle component and so on. The weights (w_{ij}) are chosen such that first principle component inherits the maximum variance of the data, the second principle component inherits second maximum variance and it goes on till the last column of **T**.

Since '*l*' is usually kept lower than '*s*', after transformation of **X** to **T**, the dimension of each series is reduced from higher dimension '*s*' to lower dimension '*l*', this way PCA can be used to reduce the dimensions of a series while preserving as much variance as possible

4. Results and Discussion

4.1 Deciding the number of trees.

The number of decision trees in a random forest classifier can significantly impact the model performance. Since the random forest algorithm uses bagging technique i.e., only a small subset of samples and a small subset of features are evaluated by each tree, if the number of observations is very large and the number of decision trees is very small, then some of the samples may be missed. It is beneficial to have the large number of decision trees as it improves the predictive power of the model, however, the computational costs are also significantly increased. To determine the optimal number of trees in a random forest model, I compared the accuracies obtained for different numbers of decision trees from 50 to 500 at two stations with the smallest (BAW) and largest (SCM) amount of available training dataset (Figs 4a and b). Results for both stations indicate that the accuracies did not vary significantly with the number of trees hence the number of decision trees (Figs 4a and b) in the model was chosen to be 50.

4.2 Deciding the frequency band

Since each value in the absolute frequency spectrum is considered as a feature in our model, the classification results will likely vary with the application of filters of the different frequency bands. Accuracy results for three frequency bands 1-20, 1-10, and 3-10 Hz (Figure 4c) were compared with each other and with the results obtained for the case without the application of any filter. Results for different frequency bands did not vary significantly. A maximum difference in the results of five percent was observed for stations EYAK, HIN, and PS12 (Fig. 4c). Overall, accuracies for the case of no filter appeared to be highest at all the stations except BAW, BAE, and PS12.

Our simple Random Forest classifier based on the absolute frequency spectrum as features show appreciable results for different stations which are located at different distances away from the events and were trained on different amounts of the data (Fig. 4c). The highest accuracy of 95 percent was obtained for the station SCM and the lowest accuracy of 75 percent for the station EYAK. Several factors appear to influence the accuracy, however, before analyzing those factors, I experiment with the

ways to improve the model. First, I apply PCA to reduce the dimensions of features and compare the results. If the lesser number of features extracted from the original features using PCA gave comparable results, it will be considered a better model because of the lesser computational time and costs associated with it.

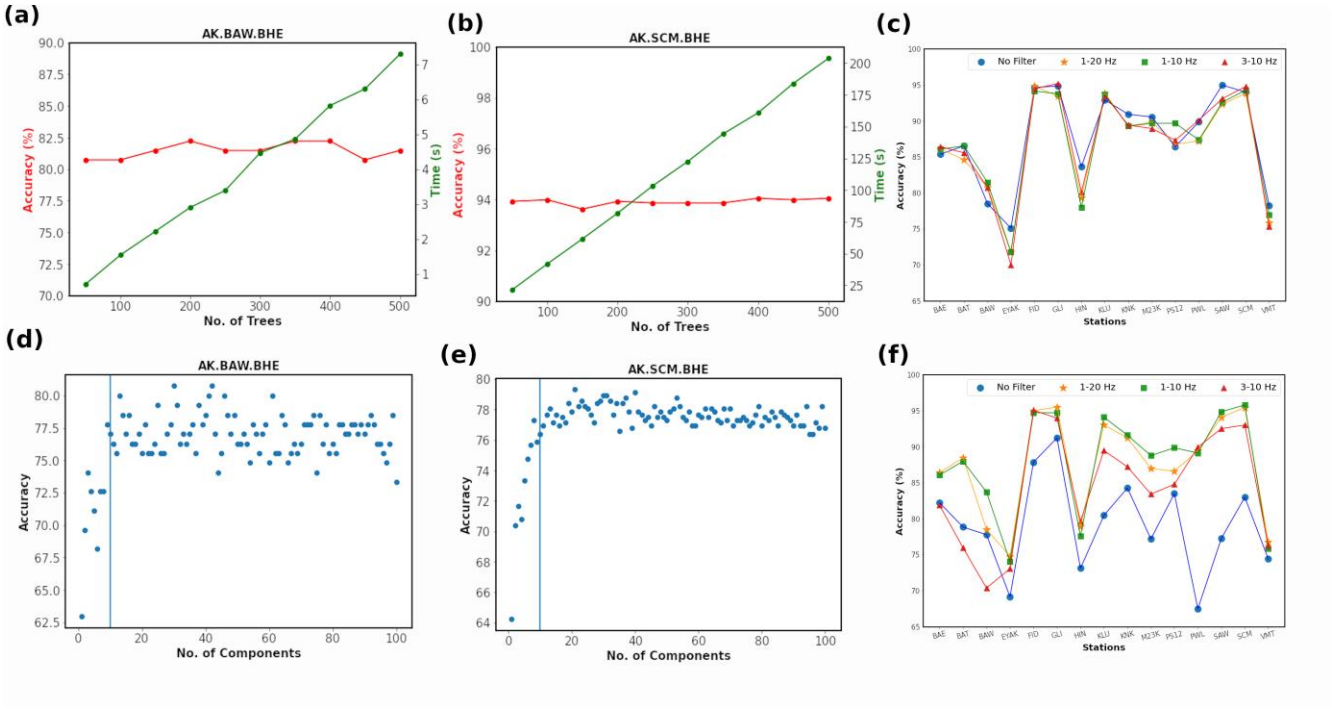

Figure 4: Variation of accuracies and time took with the number of decision trees in the RF model for stations (a) BAW with the least amount of training dataset and (b) SCM with the most amount of training dataset. (c) Accuracies at different stations when RF is directly applied on the features bandpass filtered under different frequency bands. Variations of accuracies was also measured when RF was trained on the different number of principal components for stations (d) BAW and (e) SCM. The vertical line is drawn at 10 principal components. (f) Accuracies at different stations when RF was trained on 10 principal components that were obtained by applying PCA on the original features filter under different frequency bands

4.3 Deciding the minimum number of components

The minimum number of principal components required were determined by comparing the accuracies obtained for the different number of principal components from 1 to 100 for the station BAW and SCM without the application of any filter as done previously (Figs 4d and e). Results for the station BAW show that the accuracies jump steeply by 18% after 10 components and then become almost constant, oscillating up and down by two percent. For station SCM, accuracies show a steeper jump of about 30% after five components, become maximum for 10 components, and then decrease modestly by about five percent from 10 to 100 components (Figs 4d and e). This suggests that at least ten principal components are required to capture sufficient information from the original features. Hence, further experiments were done by reducing the original features into ten principal components.

4.4 Deciding the frequency band

The same experiment with the application of filters with different frequency bands as before but with new reduced 10 principle components as features was repeated to determine the frequency band that gives the best results (Fig. 4f). Unlike previous results, Accuracies for different frequency bands vary significantly in this case. For example, accuracy range from about 65% for no filter case to 90% for a filter band of 1-10 Hz at station PAW (Fig. 4f). The best results were obtained for the frequency band of 1-10 Hz, followed by 1-20, 3-10, and no filter (Fig. 4f).

Application of PCA to extract the features improved the original random forest classifier. A comparison of accuracies from the best results obtained in (section 4.2) and (section 4.4) revealed that applying random forest classification on the 10 components obtained after PCA with a filter band of 1-10 Hz frequency produced very similar and in some cases, even better results than the random classifier applied on the original features (Fig. 5a). Another advantage is the significantly reduced computational time. In the next section, I will analyze the performance of this improved model as a function of the mean distance of a station from the center of all the considered glaciers, amount of dataset, and local

geology. The uncertainties in the accuracies of the model are computed using the normal approximation interval method (Raschka, 2018).

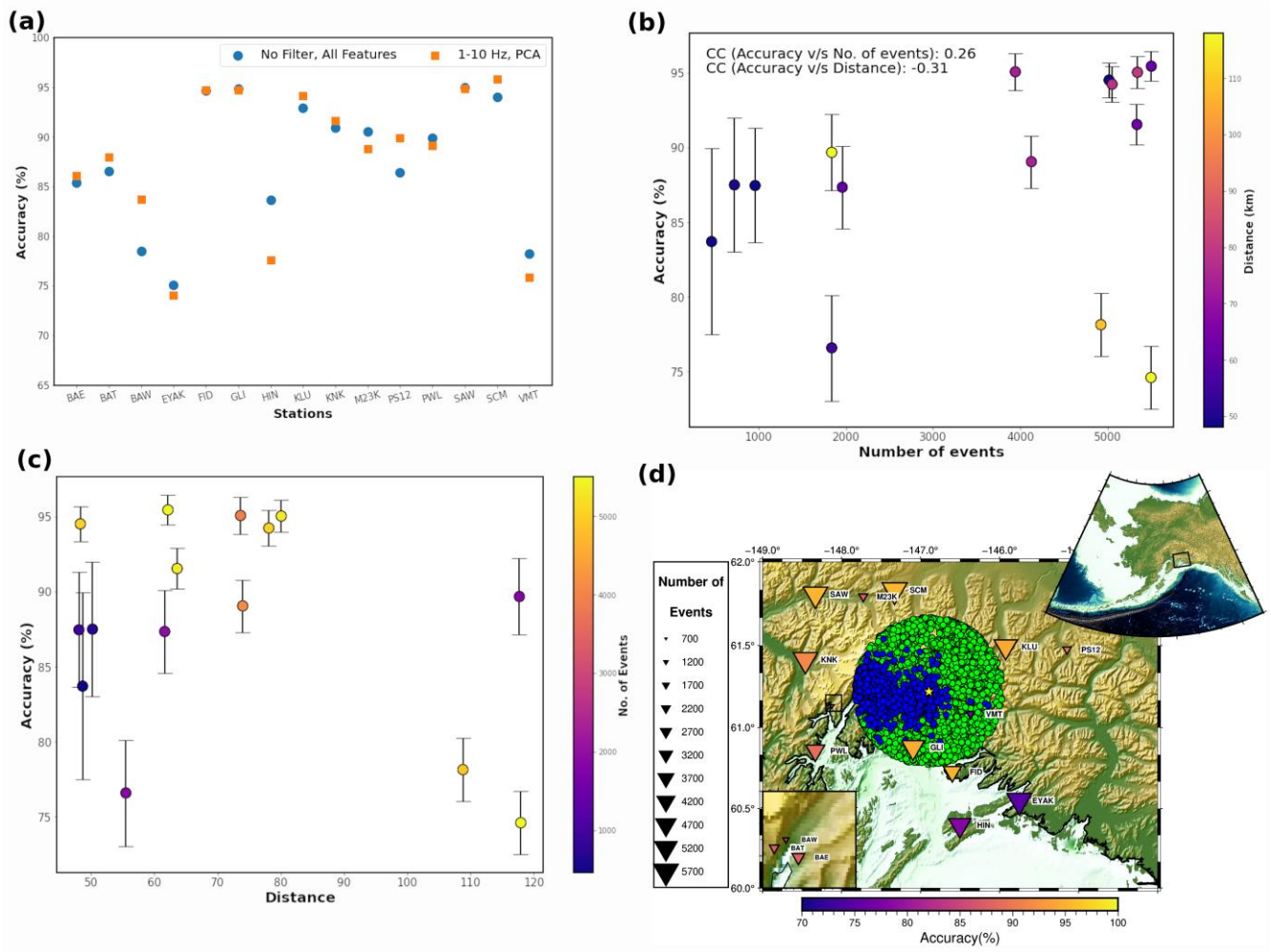

Figure 5: (a) Comparison of accuracies among the best results obtained from experiments 1 and 2. (b) Analysis of the accuracies with number of events and (c) with the average distance of a station from the glaciers. (d) Map of the study area along with all the events and stations considered in the study. Each station is color-coded according to the accuracies and sized according to the amount of available dataset.

A comparison of accuracies obtained for different stations revealed that it shows moderately positive dependence on the amount of training dataset used (Fig. 5b). The Pearson Correlation Coefficient (CC) between the accuracy and the amount of the total dataset came out to be 0.26 (Fig. 5b). Accuracies

gradually increase from 82% at station BAW with the least amount of dataset to 96% at station SCM with the most amount of dataset except at stations VMT, EYAK, and HIN where accuracies deviate from the observed trend. The positive dependence of accuracy on the amount of training data used is not unusual. It is obvious that increasing the training dataset should increase the predictive power of the model.

The deviation of the accuracies at VMT, EYAK, and HIN could be the effect of distance or local geology, or both. The performance of the model is dependent on the differences in the absolute frequency spectrum of the icequakes and earthquakes. Greater the difference between the absolute frequency spectrum of the events, the easier it will be for the model to predict accurately. If the station is located at larger distances, high frequencies will be attenuated by the time waves reach the station with the amount of attenuation depends on the nature of the propagating medium. This will make the frequency content of the icequakes and earthquakes more similar and impact the performance of the model. The accuracies show a moderately negative correlation (CC = -0.31) with the average distances of a station from all the glaciers (Figs 5b, c, and d). For a given average distance, accuracies are higher for the higher amount of training dataset and vice versa (Figs 5b, c, and d).

However, station PS12, despite being located at similar average distances as EYAK and HIN and having about half the amount of training dataset as the two, shows a much higher accuracy of 90% compared to the other two (74 and 76% at EYAK and HIN respectively). This could be related to the differences in the local geology. Station PS12 is located at a pump station no. 12 of the Trans-Alaska Pipeline System (TAPS) which lies in the northeast of the glaciers while stations EYAK and HIN are located on Cordova and the Hinchinbrook Island respectively which lie southeast of the glaciers (Fig. 5d). The seismic waves have to pass through the oceanic crust to reach stations EYAK and HIN while the waves will reach the PS12 by traveling through the continental crust only. The difference in the

properties of the propagating medium and how they affect the frequency content of the propagating waves could be the reason for the observed differences in accuracies.

5. Conclusion

In this work, the performance of the Random Forest algorithm with absolute frequency spectrum values as input features is evaluated for discrimination of earthquakes and icequakes in southeast Alaska at 15 seismic stations. Experiments were performed to determine the optimum number of decision trees, frequency band, and the minimum number of principal components. Results suggest that there is a moderately positive correlation between the amount of training data and the accuracy. This implies that higher training data would lead to higher accuracies. Further, accuracies at different stations show a similar but negative correlation with average distances from the glacier. Some stations show deviations in accuracies from the commonly observed trend, these deviations are probably caused by the effect of differences in the local geology. Overall, accuracies at most stations are close to 90% indicating the robustness of the model. One of the main advantages of this model is that it is not limited to a specific region. Future studies would need to be conducted to evaluate the performance of this model in the discrimination of other types of event that occur in different regions.

6. Code Availability

All the code used for this study is available on request from the author in the form of Jupyter notebooks.

7. Acknowledgements

Waveform data for icequakes and earthquakes were downloaded from Incorporated Research Institutions in Seismology (IRIS) Data Management Center. All the maps included in this study were created from PyGMT (Uieda et al., 2021). All the processing was done using ObsPy(Beyreuther et al., 2010; Krischer et al., 2015) and scikit-learn(Pedregosa et al., 2011).

8. References

Aden-Antoniów, F., Frank, W. B., & Seydoux, L. (2022). An Adaptable Random Forest Model for the Declustering of Earthquake Catalogs. *Journal of Geophysical Research: Solid Earth*, *127*(2), e2021JB023254. https://doi.org/10.1029/2021JB023254

Allmann, B. P., Shearer, P. M., & Hauksson, E. (2008). Spectral discrimination between quarry blasts and earthquakes in southern California. *Bulletin of the Seismological Society of America*, *98*(4), 2073–2079. https://doi.org/10.1785/0120070215

Astiz, L., Eakins, J. A., Martynov, V. G., Cox, T. A., Tytell, J., Reyes, J. C., Newman, R. L., Karasu, G. H., Mulder, T., White, M., Davis, G. A., Busby, R. W., Hafner, K., Meyer, J. C., & Vernon, F. L. (2014). The array network facility seismic bulletin: Products and an unbiased view of united states seismicity. *Seismological Research Letters*, *85*(3), 576–593. https://doi.org/10.1785/0220130141

Bartholomaus, T. C., Larsen, C. F., O'Neel, S., & West, M. E. (2012). Calving seismicity from iceberg-sea surface interactions. *Journal of Geophysical Research: Earth Surface*, *117*(4). https://doi.org/10.1029/2012JF002513

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010). ObsPy: A python toolbox for seismology. *Seismological Research Letters*, *81*(3), 530–533. https://doi.org/10.1785/GSSRL.81.3.530

Beyreuther, M., & Wassermann, J. (2008). Continuous earthquake detection and classification using discrete hidden markov models. *Geophysical Journal International*, *175*(3), 1055–1066. https://doi.org/10.1111/J.1365-246X.2008.03921.X

Bowers, D., & Selby, N. D. (2009). Forensic seismology and the comprehensive nuclear-test-ban treaty. In *Annual Review of Earth and Planetary Sciences* (Vol. 37, pp. 209–236). https://doi.org/10.1146/annurev.earth.36.031207.124143

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *26*(2), 123–140.

Breiman, Leo. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dahm, T., & Brandsdóttir, B. (1997). Moment tensors of microearthquakes from the Eyjafjallajökull volcano in South Iceland. *Geophysical Journal International*, *130*(1), 183–192. https://doi.org/10.1111/J.1365-246X.1997.TB00997.X

Dempsey, D. E., Cronin, S. J., Mei, S., & Kempa-Liehr, A. W. (2020). Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand. *Nature Communications*, *11*(1). https://doi.org/10.1038/S41467-020-17375-2

Enderlin, E. M., O'Neel, S., Bartholomaus, T. C., & Joughin, I. (2018). Evolving Environmental and Geometric Controls on Columbia Glacier's Continued Retreat. *Journal of Geophysical Research: Earth Surface*, *123*(7), 1528–1545. https://doi.org/10.1029/2017JF004541

Falcin, A., Métaxian, J. P., Mars, J., Stutzmann, É., Komorowski, J. C., Moretti, R., Malfante, M., Beauducel, F., Saurel, J. M., Dessert, C., Burtin, A., Ucciani, G., de Chabalier, J. B., & Lemarchand, A. (2021). A machine-learning approach for automatic classification of volcanic seismicity at La Soufrière Volcano, Guadeloupe. *Journal of Volcanology and Geothermal Research*, *411*, 107151. https://doi.org/10.1016/J.JVOLGEORES.2020.107151

FU Dowla, S. T. R. A. (1990). Seismic discrimination with artificial neural networks: preliminary results with regional spectral data. *Bull Seismol Soc Am*, *80*(5), 1346–1373.

Gitterman, Y., Pinsky, V., & Shapira, A. (1998). Spectral classification methods in monitoring small local events by the Israel seismic network. *Journal of Seismology*, *2*(3), 237–256. https://doi.org/10.1023/A:1009738721893

Gulia, L. (2010). Detection of quarry and mine blast contamination in European regional catalogues. *Nat Hazards*, *53*(2), 229–249. https://doi.org/10.1007/s11069-009-9426-8

Hammer, C., Ohrnberger, M., & Fäh, D. (2013). Classifying seismic waveforms from scratch: A case study in the alpine environment. *Geophysical Journal International*, *192*(1), 425–439. https://doi.org/10.1093/GJI/GGS036

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-21606-5

Hibert, C., Mangeney, A., Grandjean, G., Baillard, C., Rivet, D., Shapiro, N. M., Satriano, C., Maggi, A., Boissier, P., Ferrazzini, V., & Crawford, W. (2014). Automated identification, location, and volume estimation of rockfalls at Piton de la Fournaise volcano. *Journal of Geophysical Research: Earth Surface*, *119*(5), 1082–1105. https://doi.org/10.1002/2013JF002970

Hibert, Clément, Provost, F., Malet, J. P., Maggi, A., Stumpf, A., & Ferrazzini, V. (2017). Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest algorithm. *Journal of Volcanology and Geothermal Research*, *340*, 130–142. https://doi.org/10.1016/j.jvolgeores.2017.04.015

Holt, M. M., Koper, K. D., Yeck, W., D'Amico, S., Li, Z., Hale, J. M., & Burlacu, R. (2019). On the Portability of ML–Mc as a Depth Discriminant for Small Seismic Events Recorded at Local Distances. *Bulletin of the Seismological Society of America*, *109*(5), 1661–1673. https://doi.org/10.1785/0120190096

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2019). Machine Learning in Seismology: Turning Data into Insights. *Seismological Research Letters*, *90*(1), 3–14. https://doi.org/10.1785/0220180259

Koper, K. D., Pechmann, J. C., Burlacu, R., Pankow, K. L., Stein, J., Hale, J. M., Roberson, P., & McCarter, M. K. (2016). Magnitude-based discrimination of man-made seismic events from naturally occurring earthquakes in Utah, USA. *Geophysical Research Letters*, *43*(20), 10,638-10,645. https://doi.org/10.1002/2016GL070742

Kortström, J., Uski, M., & Tiira, T. (2016). Automatic classification of seismic events within a regional seismograph network. *Computers and Geosciences*, *87*, 22–30. https://doi.org/10.1016/j.cageo.2015.11.006

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., & Wassermann, J. (2015). ObsPy: A bridge for seismology into the scientific Python ecosystem. *Computational Science and Discovery*, *8*(1). https://doi.org/10.1088/1749-4699/8/1/014003

Kuyuk, H. S., Yildirim, E., Dogan, E., & Horasan, G. (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 Science*, *11*(1), 93–100. https://doi.org/10.5194/nhess-11-93-2011

Kuyuk, H. S., Yildirim, E., Dogan, E., & Horasan, G. (2012). Application of k-means and Gaussian mixture model for classification of seismic activities in Istanbul. *Nonlinear Processes in Geophysics*, *19*(4), 411–419. https://doi.org/10.5194/npg-19-411-2012

Lara, F., Lara-Cueva, R., Larco, J. C., Carrera, E. V., & León, R. (2021). A deep learning approach for automatic recognition of seismo-volcanic events at the Cotopaxi volcano. *Journal of Volcanology and Geothermal Research*, *409*. https://doi.org/10.1016/J.JVOLGEORES.2020.107142

Linville, L., Pankow, K., & Draelos, T. (2019). Deep Learning Models Augment Analyst Decisions for Event Discrimination. *Geophysical Research Letters*, *46*(7), 3643–3651. https://doi.org/10.1029/2018GL081119

Mackey, K. G., Fujita, K., Gounbina, L. V., Koz'min, B. M., Imaev, V. S., Imaeva, L. P., & Sedov, B. M. (2003). Explosion contamination of the northeast Siberian seismicity catalog: Implications for natural earthquake distributions and the location of the Tanlu Fault in Russia. *Bulletin of the Seismological Society of America*, *93*(2), 737–746. https://doi.org/10.1785/0120010196

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017a). Implementation of a multistation approach for automated event classification at Piton de la Fournaise volcano. *Seismological Research Letters*, *88*(3), 878–891. https://doi.org/10.1785/0220160189

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017b). Implementation of a Multistation Approach for Automated Event Classification at Piton de la Fournaise Volcano. *Seismological Research Letters*, *88*(3), 878–891. https://doi.org/10.1785/0220160189

Marzen, R. E., Gaherty, J. B., Shillington, D. J., & Kim, W. Y. (2021). Shaking in the southeastern united states: Examining earthquakes and blasts in the central georgia-south carolina seismic region. *Seismological Research Letters*, *92*(5), 3145–3164. https://doi.org/10.1785/0220210029

Masotti, M., Falsaperla, S., Langer, H., Spampinato, S., & Campanini, R. (2006). Application of Support Vector Machine to the classification of volcanic tremor at Etna, Italy. *Geophysical Research Letters*, *33*(20). https://doi.org/10.1029/2006GL027441

Matthew Sturm, V., Hall, D. K., Benson, eARL S., & Field, W. O. (1991). Non-climatic control of glacier-terminus fluctuations in the Wrangell and Chugach Mountains, Alaska, U.S.A. *Journal of Glaciology*, *37*(127), 348–356. https://doi.org/10.3189/S0022143000005785

Meier, M. F., & Post, A. (1987). Fast tidewater glaciers. *Journal of Geophysical Research*, *92*(B9), 9051–9058. https://doi.org/10.1029/JB092IB09P09051

NEAVE KG, & SAVAGE JC. (1970). ICEQUAKES ON THE ATHABASCA GLACIER. *J Geophys Res*, *75*(8), 1351–1362. https://doi.org/10.1029/JB075I008P01351

Neel, O., Marshall, H. P., Mcnamara, D. E., & Pfeffer, W. T. (2007). Seismic detection and analysis of icequakes at Columbia Glacier, Alaska. *Alaska, J. Geophys. Res*, *112*, 3–23. https://doi.org/10.1029/2006JF000595

O'Rourke, C. T., Baker, G. E., & Sheehan, A. F. (2016). Using P/S amplitude ratios for seismic discrimination at local distances. *Bulletin of the Seismological Society of America*, *106*(5), 2320–2331. https://doi.org/10.1785/0120160035

Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In *Journal of Machine Learning Research* (Vol. 12, Issue 85). http://scikit-learn.sourceforge.net.

Podolskiy, E. A., & Walter, F. (2016). Cryoseismology. *Reviews of Geophysics*, *54*(4), 708–758. https://doi.org/10.1002/2016RG000526

Post, A., O'Neel, S., Motyka, R. J., & Streveler, G. (2011). A complex relationship between calving glaciers and climate. *Eos*, *92*(37), 305–306. https://doi.org/10.1029/2011EO370001

Pyle, M. L., & Walter, W. R. (2019). Investigating the effectiveness of P/S amplitude ratios for local distance event discrimination. *Bulletin of the Seismological Society of America*, *109*(3), 1071–1081. https://doi.org/10.1785/0120180256

Qamar, A. (1988a). Calving icebergs: A source of low-frequency seismic signals from Columbia Glacier, Alaska. *Journal of Geophysical Research*, *93*(B6), 6615. https://doi.org/10.1029/JB093IB06P06615

Qamar, A. (1988b). Calving icebergs: A source of low-frequency seismic signals from Columbia Glacier, Alaska. *Journal of Geophysical Research*, *93*(B6), 6615. https://doi.org/10.1029/JB093iB06p06615

Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. https://doi.org/10.48550/arxiv.1811.12808

Rodgers, A. J., & Walter, W. R. (2002). Seismic discrimination of the May 11, 1998 Indian nuclear test with short-period regional data from station NIL (Nilore, Pakistan). *Pure and Applied Geophysics*, *159*(4), 679–700. https://doi.org/10.1007/S00024-002-8654-6

Rubin, M. J., Camp, T., Herwijnen, A. Van, & Schweizer, J. (2012). Automatically detecting avalanche events in passive seismic data. *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, *1*, 13–20. https://doi.org/10.1109/ICMLA.2012.12

Stump, B. W., Hedlin, M. A. H., Pearson, D. C., & Hsu, V. (2002). Characterization of mining explosions at regional distances: Implications with the international monitoring system. *Reviews of Geophysics*, *40*(4), 2-1-2–21. https://doi.org/10.1029/1998RG000048

Taylor, S. (1996). Analysis of high-frequency Pg/Lg ratios from NTS explosions and western US earthquakes. *Bull Seismol Soc Am*, *86*(4), 1042–1053.

Tibi, R. (2021). Discrimination of seismic events (2006-2020) in North Korea Using P/Lg amplitude ratios from regional stations and a bivariate discriminant function. *Seismological Research Letters*, *92*(4), 2399–2409. https://doi.org/10.1785/0220200432

Tibi, R., Koper, K. D., Pankow, K. L., & Young, C. J. (2018). Discrimination of anthropogenic events and tectonic earthquakes in Utah using a quadratic discriminant function approach with local distance amplitude ratios. *Bulletin of the Seismological Society of America*, *108*(5), 2788–2800. https://doi.org/10.1785/0120180024

Tiira, T. (1996). Discrimination of nuclear explosions and earthquakes from teleseismic distances with a local network of short period seismic stations using artificial neural networks. *Physics of the Earth and Planetary Interiors*, *97*(1–4), 247–268. https://doi.org/10.1016/0031-9201(95)03132-4

Trabant, D., March, R., & Molnia, B. (2002). Growing and advancing calving glaciers in Alaska. *Eos*, *83*.

Uieda, L., Tian, D., Leong, W. J., Jones, M., Schlitzer, W., Toney, L., Grund, M., Yao, J., Magen, Y., Materna, K., Newton, T., Anant, A., Ziebarth, M., Wessel, P., & Quinn, J. (2021). *PyGMT: A Python interface for the Generic Mapping Tools*. https://doi.org/10.5281/ZENODO.5607255

W-Y Kim, V. A. A. L.-L. P. R. (1997). Discrimination of earthquakes and explosions in southern Russia using regional high-frequency three-component data from the IRIS/JSP Caucasus network. *Bull Seismol Soc Am*, *87*(3), 569–588.

W. R. Walter, K. M. M. H. J. P. (1995). Phase and spectral ratio discrimination between NTS earthquakes and Explosions. Part I: Empirical observations. *Bull. Seis. Soc. Am.*, *85*, 1050–1067.

Walter, W. R., Dodge, D. A., Ichinose, G., Myers, S. C., Pasyanos, M. E., & Ford, S. R. (2018). Body-Wave Methods of Distinguishing between Explosions, Collapses, and Earthquakes: Application to Recent Events in North Korea. *Seismological Research Letters*, *89*(6), 2131–2138. https://doi.org/10.1785/0220180128

Wang, R., Schmandt, B., Holt, M., & Koper, K. (2021). Advancing Local Distance Discrimination of Explosions and Earthquakes With Joint P/S and ML-MC Classification. *Geophysical Research Letters*, *48*(23), e2021GL095721. https://doi.org/10.1029/2021GL095721

West, M. E., Larsen, C. F., Truffer, M., O'Neel, S., & LeBlanc, L. (2010). Glacier microseismicity. *Geology*, *38*(4), 319–322. https://doi.org/10.1130/G30606.1

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Yuan, B., Tan, Y. J., Mudunuru, M. K., Marcillo, O. E., Delorey, A. A., Roberts, P. M., Webster, J. D., Gammans, C. N. L., Karra, S., Guthrie, G. D., & Johnson, P. A. (2019). Using machine learning to discern eruption in noisy environments: A case study using CO 2 -driven cold-water geyser in Chimayó, New Mexico. *Seismological Research Letters*, *90*(2 A), 591–603. https://doi.org/10.1785/0220180306

Zeiler, C., & Velasco, A. A. (2009). Developing local to near-regional explosion and earthquake discriminants. *Bulletin of the Seismological Society of America*, *99*(1), 24–35. https://doi.org/10.1785/0120080045

Zhao, G. Y., Ma, J., Dong, L. J., Li, X. B., Chen, G. H., & Zhang, C. X. (2015). Classification of mine blasts and microseismic events using starting-up features in seismograms. *Transactions of Nonferrous Metals Society of China (English Edition)*, *25*(10), 3410–3420. https://doi.org/10.1016/S1003-6326(15)63976-0