SMU Data Science Review

Volume 1 | Number 1 Article 11

2018

Seismology and Volcanology: Exploration of Volcanoes, Long-Periods, and Machines - Predicting Volcano Eruption Using Signature Seismic Data

Kyle Killion
Southern Methodist University, kkillion@mail.smu.edu

Rajeev Kumar
Southern Methodist University, rajeevk@mail.smu.edu

Celia J. Taylor
Southern Methodist University, celiat@mail.smu.edu

Gabriele Morra
University of Louisiana at Lafayette, gabrielemorra@gmail.com

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

 Part of the Geophysics and Seismology Commons, and the Software Engineering Commons

Recommended Citation
Killion, Kyle; Kumar, Rajeev; Taylor, Celia J.; and Morra, Gabriele (2018) "Seismology and Volcanology: Exploration of Volcanoes, Long-Periods, and Machines - Predicting Volcano Eruption Using Signature Seismic Data," *SMU Data Science Review*: Vol. 1 : No. 1 , Article 11.
Available at: https://scholar.smu.edu/datasciencereview/vol1/iss1/11

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.

Seismology and Volcanology: Exploration of Volcanoes, Long-Periods, and Machines - Predicting Volcano Eruption Using Signature Seismic Data

Kyle Killion[1], Rajeev Kumar[1], Celia Taylor[1], Gabriele Morra, Ph.D.[2],

[1] Master of Science in Data Science Capstone Project, Southern Methodist University, 3300 University Blvd, Dallas, TX 75205
[2]Department of Physics, School of Geosciences, University of Louisiana at Lafayette, 104 East University Avenue, Lafayette, LA 70504

Abstract. Seismo-volcanologists manually isolate and verify long-period waves and Strombolian events using seismic and acoustic waves. This is a very detailed and time-consuming process. This project is to employ machine learning algorithms to find models which locate long-period and Strombolian signatures automatically. By comparing the timing of seismic and acoustic waves, clustering techniques effectively isolated big volcanic events and aided in the further refinement of techniques to capture the hundreds of typical daily Strombolian events at Villarrica volcano. Within the research, we utilized the unsupervised machine learning environment to locate a group of signatures for customizing machine learned long-period signature detection.

1 Introduction

Villarrica is a stratovolcano in the southern Andean Cordillera of Chile. This volcano has an active lava lake at its summit crater. Varying sizes of gas bubbles, called slugs, rise through a tubular core of magma from the depths of the volcano and rupture at the surface of the lava lake. The difference of air pressure causes the slugs to burst in Strombolian eruptions and throw magma into the air. Because the size of the gas bubbles ranges from small to very large (millimeters to meters in diameter) and slugs coalesce as they rise through the lava tube, the violence of these eruptions also varies. Even small eruptions are registered on seismographs and in acoustic waves. Eruptions occur in Villarrica many times per hour and add up to hundreds per day. Many Strombolian volcanoes like Villarrica emit 1-2 Hertz (Hz) waves for about one half of a minute after each Strombolian burst. The idea of this research is to find the signature for these long-period seismic waves with an unbiased, unsupervised, and more automatic method. By determining these signatures, the conduit flow and activity can further be tracked and learned. The ultimate goal is to determine and more quickly detect which signatures or signature activities highly correlate to devastating explosions. This knowledge may help protect people and save lives.

2 Data

The seismic and acoustic data for this research are from the Incorporated Research Institutions for Seismology organization or IRIS website [1]. A structured database of the Villarrica volcano data from 2010 through 2013 is in the IRIS Data Management Center (DMC) MetaData Aggregator Network. The Python ObsPy library [2] was utilized to access the structured database associated with the Chilean Villarrica volcano. Various seismic and acoustic collections of stations were deployed from 2010 through 2013 around the volcano, at different distances, varying from some very close to the summit of the volcano to others located at the base of the volcanic edifice.

2.1 Data Scope and Size

The scope of the data used for this research changed during the discovery phase of the project. Originally, we looked at data from the Kilauea volcano in Hawaii. Then we found a very well documented data set of the Villarrica volcano in Chile. This data set is from the research discussed in two papers. The first paper is by Richardson and Waite [3], and the second paper is by Richardson, Waite, and Palma [4]. These three seismo-volcanologists provided data on the manual detection of long-period waves and Strombolian activity. The focus of this project changed from working with data from multiple volcanos to identifying Villarrica volcano long-period seismic waves, acoustic waves, and Strombolian activity through machine learning in this data set.

The size of the data employed for this research depended on the goal of building an algorithm for detecting the long-period seismic form of the Strombolian events. Focusing on one volcano narrowed the scope of the project and reduced the size of data used until the algorithm captured events for the one volcano. One day's worth of data using 10 stations for one volcano is about 1 Gigabyte. We used data from several days and various time periods within those days to capture a Villarrica Strombolian event and its long-period signature.

2.2 Data Description

Table 1 has the different parameters that we used to process the Villarrica seismic and acoustic data. Some parameters are part of the IRIS data set. We defined some parameters to do our own processing.

Table 1: Parameter Definitions

Data Element	Data Description	Reference
Sampling rate	Sampled data rate at (100/second)	[5]
Longitude	Location of station	[5]
Latitude	Location of station	[5]

Proximal Volcano	Volcano the data describes	In this case, the data is from Villarrica volcano.
Network code	"a 1 or 2 character code identifying the network/owner of the data. These codes are assigned by the FDSN to provide uniqueness to seismological data."	[5]
Station code	"a 1 to 5 character identifier for the station recording the data."	[5]
Location ID	"2 character code used to uniquely identify different data streams at a single station. These IDs are commonly used to logically separate multiple instruments or sensor sets at a single station."	[5]
Channel codes	"3 character combination used to identify the 1) band and general sample rate 2) the instrument type and 3) the orientation of the sensor. A convention for these codes has been established."	[5]
FracDelta	Difference of FracHigh and FracLow over seismic Data.	User defined variable to implement algorithms
FracHigh	Rolling window of 600 data points centered to capture a maximum value	User defined variable to implement algorithms
FracLow	Rolling window of 600 data points centered to capture a minimum value	User defined variable to implement algorithms
DeltaDiff	Counts how many times the seismic wave is above and below zero.	User defined variable to implement algorithms
Filtered	Removes all but the hertz frequencies of interest of 'Long-Periods'	User defined variable to identify signals and remove noise.

3 Data Acquisition and Exploration Insights

The IRIS DMC MetaData Aggregator Network and Python library ObsPy were utilized to access a database of the seismic records of numerous stations around the Chilean Villarrica Volcano. These data consist of instrumental data that measured seismic and acoustic activity using a Reftek 130 Datalogger (seismic) and a

Honeywell Differential Pressure Sensor (acoustic). These instruments captured data from different deployments of approximately 30 locations around the Villarrica crater.

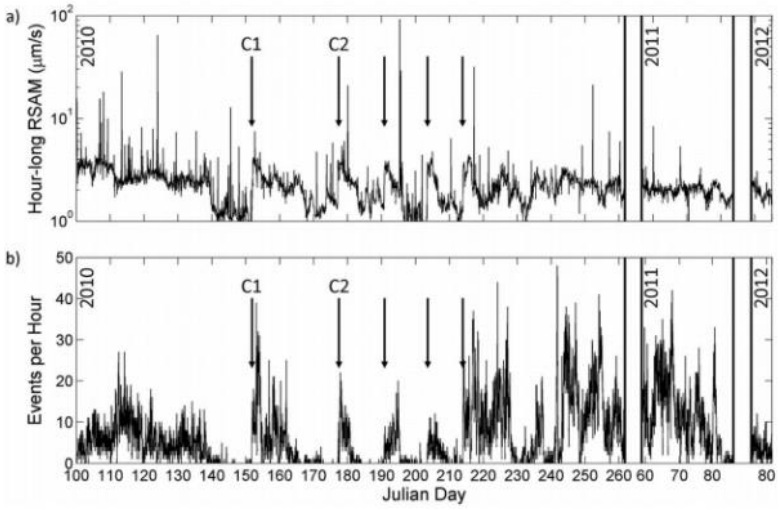

Figure 1. Julian days indicating events per hour [4]

We did not look at absolute magnitude of the seismic and acoustic waves, which required a complicated fine tuning of an algorithm based on local attenuation and noise to be transformed in P and S waves. Instead, we used relative magnitudes and frequency (measured in Hertz=1/s). As shown in Figure 1, the data is from selected Julian days in which prior works [4] recorded the greatest number of events per hour.

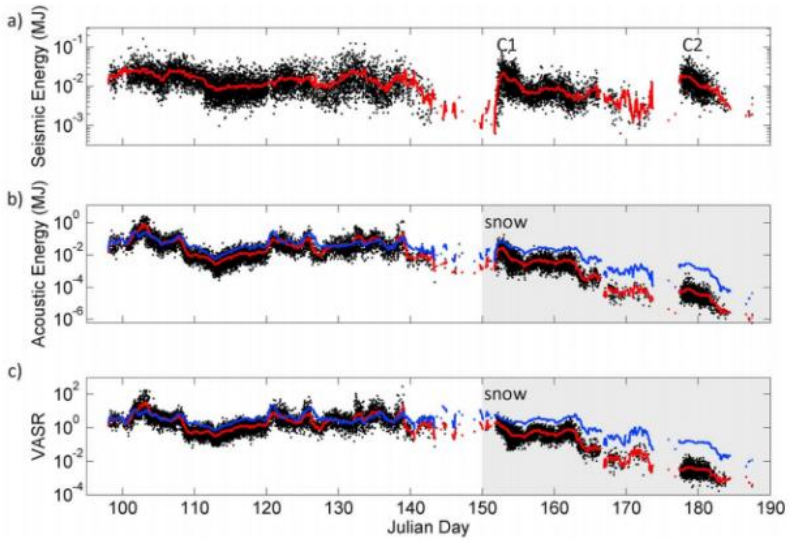

Figure 2. Julian days indicating impacts of snow over the acoustic data [4]

The accuracy and validity of the data also had to be taken into account for their preparation. As indicated in Figure 2, the snowfall upon Villarrica during certain times impacted the quality of the acoustic data set. The break of rolling averages around the Julian day 150 timeframe in the Southern Hemisphere shows this snowfall dampening effect and persuaded us to avoid this data for further analysis [4].

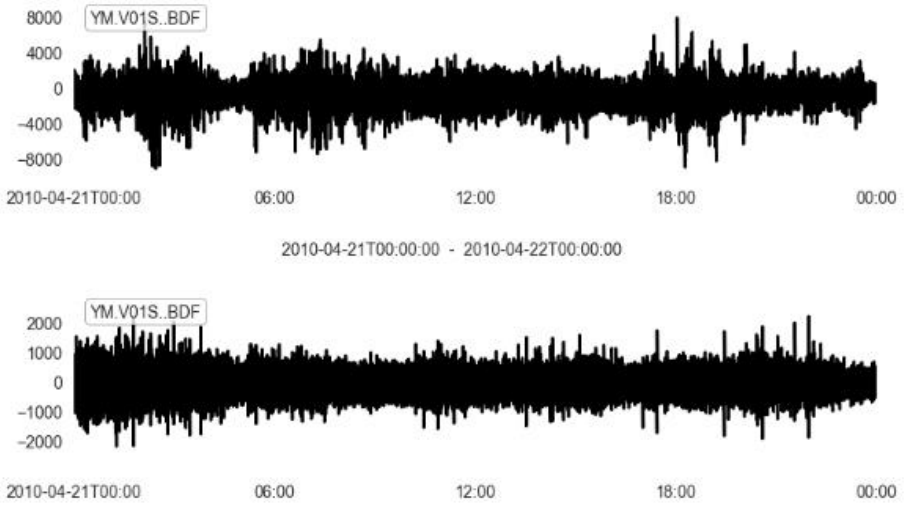

Figure 3. V01S instrument showing the unfiltered acoustic data (top) and then the effects of filtering the acoustic readings (bottom)

Given the nature of acoustic readings and implementation, frequency based filtering was applied to eliminate the noise of airplanes, thunder, lightning, and other aspects. Because of these events, frequencies above 10 Hz and below .4 Hz have been filtered out. Figure 3 shows the difference in the acoustic data before and after the filtering.

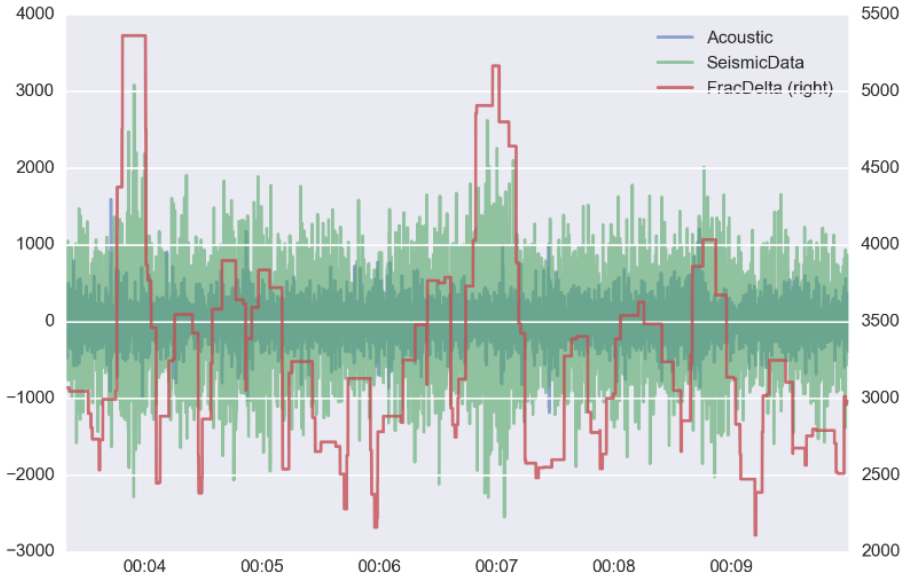

Figure 4. FracDelta overlaid on acoustic and seismic activity Julian day 110

Seismic data were then explored to aid in capturing signatures of long-period events. As seen in Figure 4, a series of rolling centered windows labeled FracDelta, FracHigh, and FracLow were utilized to explore seismic events. A window comprised of 600 data points, equivalent to 6 seconds, would capture a long-period event. Each long-period event typically would last up to 30 seconds. FracHigh has been aimed at capturing the centered maximum value within the specified window while FracLow similarly captures the centered minimum value. The FracDelta calculates the difference of FracHigh and FracLow in order to encapsulate these events of interest.

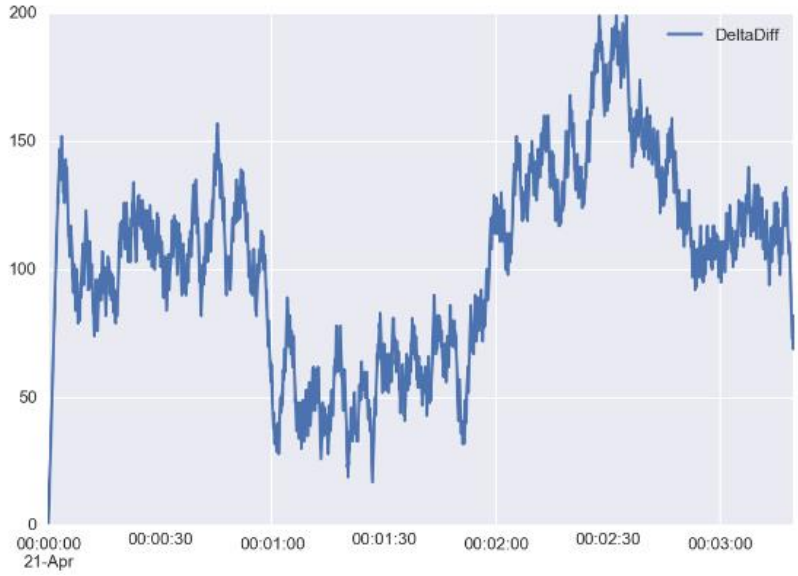

Figure 5. DeltaDiff as it tracks the seismic activity of Villarrica

Figure 5 is the delta calculation tracker that keeps a total of the harmonizing seismic waves as they oscillate between positive and negative values. We are still unsure about this feature and what insights it provides scientifically, but curiosity of its implementation in the study remains. While monitoring the positive and negative energies as they flow within the earth, this is thought to aid in keeping track of the balances of infrasonic energy.

4 Methods and Models

4.1 Supervised vs Unsupervised Machine Learning Algorithms

We implemented and evaluated various machine learning algorithms such as supervised, unsupervised, and semi-supervised machine learning. Supervised machine learning is the machine learning process in which the algorithm learns on the labeled or classified data using a training data set. Unsupervised machine learning algorithms learn on the non-labeled data. Semi-supervised machine learning is when only a portion of the data is labeled. In the seismic dataset, the data is preprocessed within the unsupervised machine learning environment. This is ideal given that no human interaction or expertise is needed in order to accomplish the identification of long-period activities. We chose K-means Clustering with Local Outlier Factor (LOF) from sklearn and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithms to find clustering and associations between seismic and acoustic data. Our goal with the clustering is to identify and estimate the degree of correlation between seismic and acoustic data during the Strombolian events.

4.2 K-Means Clustering

K-Means clustering algorithm uses an iterative refinement technique most commonly referred to as Lloyd's algorithm. In this algorithm, each observation is assigned to a cluster whose mean has the least squared Euclidean distance. Then, the mean of a cluster is recomputed. This process iterates until the assignments of observations to the cluster do not change.

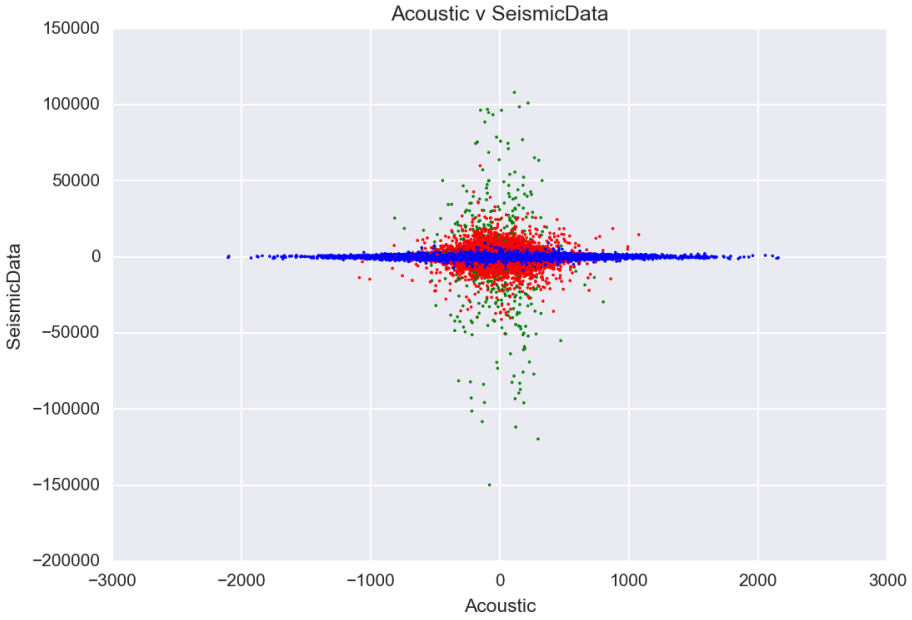

Figure 6. Result of K-Means Clustering. (DeltaDiff)

Within Figure 6, we can see the lagging nature of acoustic to seismic data and clustering. When Strombolian events occur, the seismic stations record the seismic waves through the earth's crust before the acoustic instruments capture the acoustic waves. More distance and conflicting factors in the atmosphere interferes with the acoustic waves.

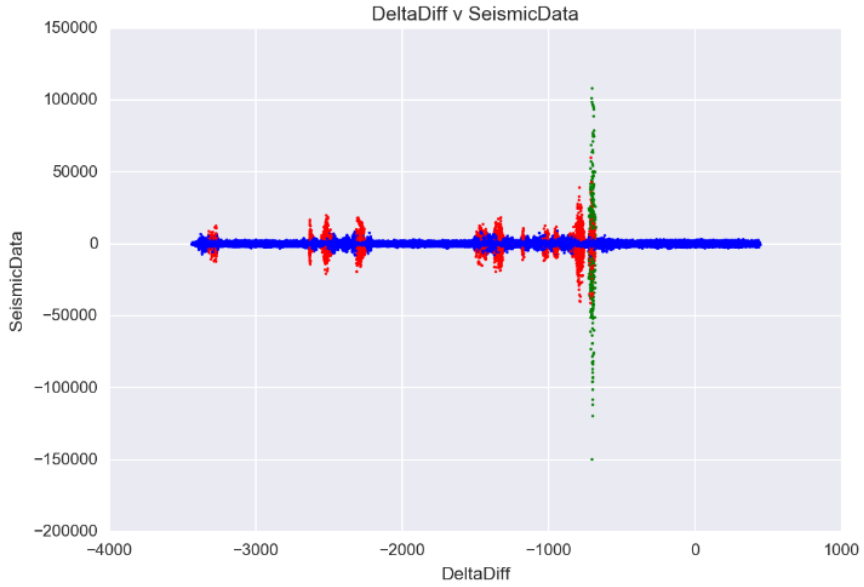

Figure 7. Result of K-Means Clustering. (DeltaDiff)

The Figure 7 shows the result of K-Means clustering using the 3 clusters option. The vertical spikes indicate Strombolian events in conjunction with FracDelta. We notice that we are observing a crescendo data structure within this relationship.

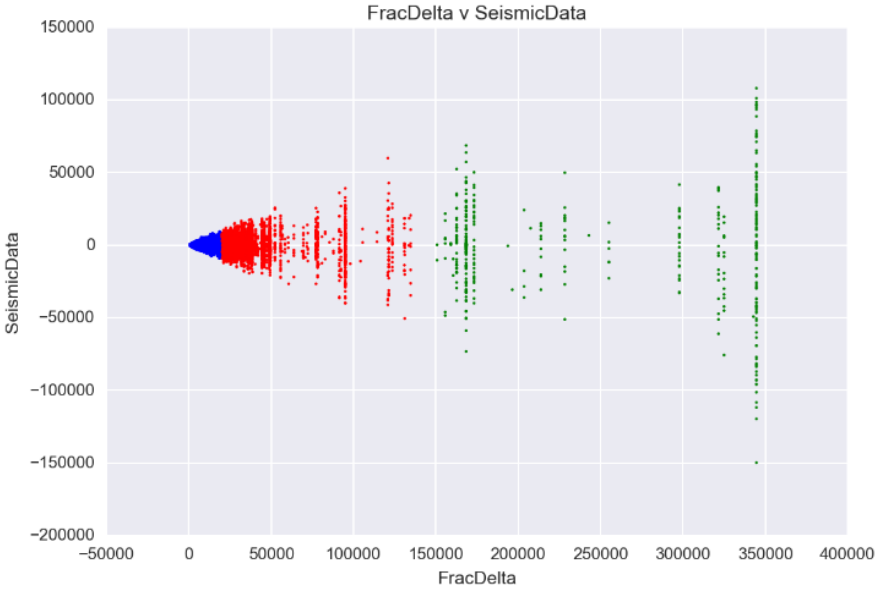

Figure 8. Result of K-Means Clustering (FracDelta)

The Figure 8 shows bands of frequencies. Between 150,000 and 350,000 FracDelta, there are different patches of long-period seismic events. We observed that the acoustic wave data within these periods lagged behind the seismic wave data. The acoustic data affirms lava lake surface activity in result to seismic long-period activity.

Figure 9. Inspections of data within the unsupervised clustered events. Indicated in gold, acoustic data was providing a surface lagging indicator affirming previous Strombolian events.

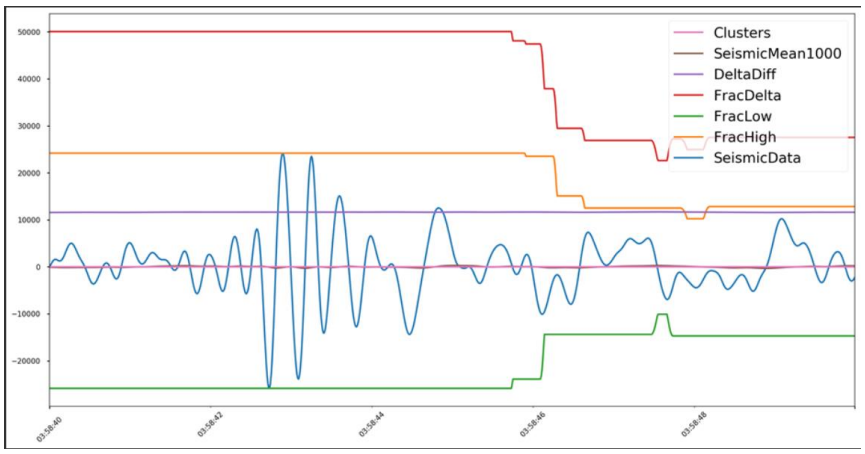

Figure 10. Classic seismic Strombolian signature

Figure 10 shows a classic Strombolian seismic signature that lasts about 10 seconds. This Strombolian event was captured from the VA station, which is the closest station to Villarrica's crater, from data on 02/28/2011. Finding this classic Strombolian seismic event from raw data affirms that machine learning does discern the desired volcanic activity signatures.

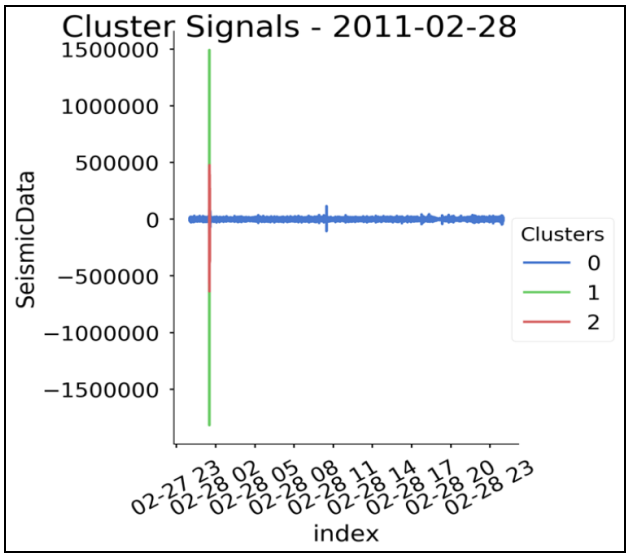

Figure 11. Clustering Detects Major Event

Figure 11 depicts the effort to find signature Strombolian events. We looked at the 24 hour time period of 2/28/2011. We found a major event around midnight. This was a very positive result in our research because it helped us to fine tune our algorithms and variable thresholds to detect smaller signature events. This clustering is also on raw data.

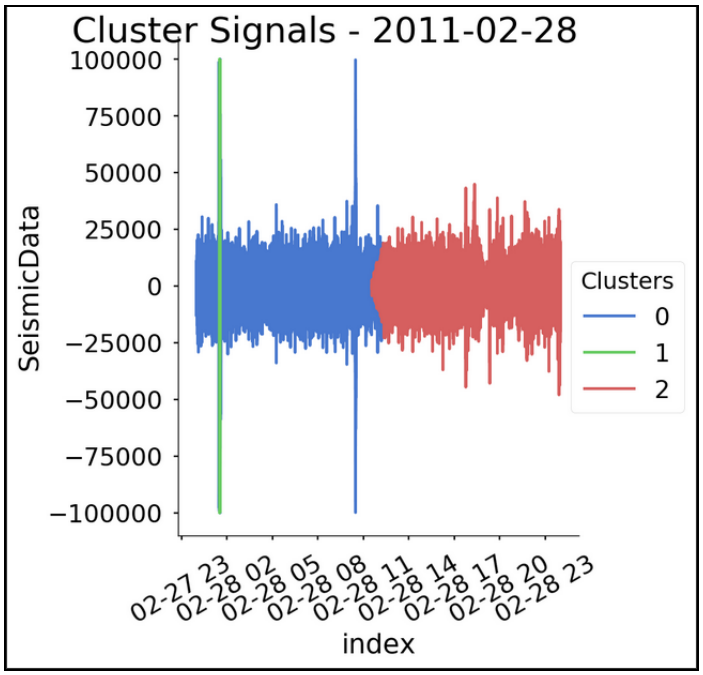

Figure 12. Filtered Raw Data Clustering

To rule out any unwanted noise or excessive extraneous signals from the data in Figure 11, we employed a filter to look at the raw seismic data from 2/28/2011. Figure 12 shows the results. Clustering still shows the seismic activity.

4.3 DBSCAN Clustering

In DBSCAN, clusters are defined as region of higher density than the rest. Advantages of DBSCAN are the following:

1. It does not need the number of clusters to be specified as an input to algorithm.
2. It is robust to outliers.
3. It only requires two parameters. [6]

We ran DBSCAN clustering on the Villarrica data. At this time, the results are inconclusive as to the effectiveness of a signature model from DBSCAN algorithm.

5 Future Work

The future work of this project will concentrate on taking the verified signals of Strombolian events from various Villarrica stations to create a signature of Villarrica Strombolian events. The Strombolian event signature will then be cross-correlated with the raw signal to predict Strombolian events. Additional future work will be to enrich the model with data from other active volcanos to see if the model captures and predicts Strombolian events with those volcanos.

5 Ethics

The implications from this study disclaim that they should be taken to proclaim predictability in volcano eruptions. The background and data analytics should not be handled as conclusive. There should be an extreme respect for these acts of "Mother Nature". Health and safety is above all the most important concern and should not be jeopardized by any means forward for any knowledge gained or potential aspects learned from this research.

6 Conclusion

Seismo-volcanologists invest much time and energy to accurately capture and verify long-period and Strombolian seismic and acoustic events. This research is to use machine learning to aid this effort and automatically identify these events. We effectively pulled seismic and acoustic data from the IRIS platform using ObsPy Python library. We demonstrated that acoustic waves lag seismic waves and that this lag increases the farther away the station is from the crater. To increase accuracy, we are working with signals from the closest station to the crater. Through K-Means clustering we detected big seismic and acoustic events. We identified a classic Strombolian seismic signal to use in model generation. We used this Strombolian seismic signal to cross correlate to raw data with methods like LOF from sklearn.

We can further enhance our current model to predict Strombolian events for other active volcanos across the globe, which could potentially save lives and properties from getting damaged or lost. This model should aid the government agencies to plan and take appropriate measures before big Strombolian events happen.

Acknowledgments. The authors, Kyle Killion, Rajeev Kumar, and Celia Taylor, would like to give an enormous thanks to Dr. Gabriele Morra and Dr. Daniel Engels for their assistance, guidance, and feedback on this article.

This work was assigned as part of the Southern Methodist University Data Science Program for Capstone Project.

Celia Taylor is with the Southern Methodist University, Master of Science in Data Science Program, Dallas, TX 75205 USA (e-mail: celiat@mail.smu.edu).

Rajeev Kumar is with the Southern Methodist University, Master of Science in Data Science Program, Dallas, TX 75205 USA (e-mail: rajeevk@mail.smu.edu).

Kyle Killion is with the Southern Methodist University, Master of Science in Data Science Program, Dallas, TX 75205 USA (e-mail: kkillion@mail.smu.edu).

References

1. http://ds.iris.edu, "IRIS Incorporated Research Institutions for Seismology," Incorporated Research Institutions for Seismology. [Online].

2. https://github.com/obspy/obspy/wiki, "ObsPy," Open source repository. [Online].

3. Richardson, Joshua P., Waite, Gregory P., "Waveform inversion of shallow repetitive long-period events at Villarrica Volcano, Chile," *American Geophysical Union, Journal of Geophysical Research: Solid Earth,* Vols. 118, doi:10,1002, pp. 4922-4936, 2013.
4. Richardson, Joshua P., Waite, Gregory P., Palma, Jose Luis, "Varying seismic-acoustic properties of the fluctuating lava lake at Villarrica volcano, Chile," *American Geophysical Union, Journal of Geophysical Research: Solid Earth,* Vols. 119, doi:10.1002, no. 2014JB011002, pp. 5560-5573, 2014. 4. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
5. http://ds.iris.edu/ds/nodes/dmc/data/formats, "IRIS Incorporated Research Institutions for Seismology Data Formats," [Online].
6. https://en.wikipedia.org/wiki/DBSCAN, "DBSCAN," WIKIPEDIA The Free Encyclopedia. [Online].
7. GitHub Link for the Project Code : https://github.com/rajeev4k/CapstoneProjectStrombolian

8. Strombolian eruptions are relatively mildly explosive, with a volcanic Explosivity index of about 2 to 3. https://en.wikipedia.org/wiki/Strombolian_eruption