Journal of Volcanology and Geothermal Research 405 (2020) 107042

Contents lists available at ScienceDirect

Journal of Volcanology and Geothermal Research

journal homepage: www.elsevier.com/locate/jvolgeores

Using unsupervised machine learning to identify changes in eruptive behavior at Mount Etna, Italy

Leighton M. Watson

Department of Earth Sciences, University of Oregon, Eugene, OR, United States

ARTICLE INFO

Article history:
Received 22 June 2020
Received in revised form 27 August 2020
Accepted 31 August 2020
Available online 12 September 2020

Keywords:
Volcano infrasound
Unsupervised machine learning
K-means
Mount Etna
Volcano monitoring

ABSTRACT

Volcanoes frequently generate infrasound signals that need to be processed before they can be used to monitor and track changes in eruptive activity. Unsupervised machine learning is complementary to existing processing methods and can be used for data exploration to identify features of interest in the data. Here, I examine three days of infrasound data from Mount Etna, Italy, that encompasses the 24 December 2018 fissure eruption. The continuous infrasound data is divided into overlapping windows and for each window I extract seven features in the time and frequency domains that characterize the signal. I apply the *k*-means clustering algorithm to group the data into seven clusters and generate a discrete time series of cluster labels. The cluster labels clearly identify a change in eruptive activity from Strombolian explosions at the summit to lava fountaining at the fissure. Feature distributions and representative waveforms for each cluster are analyzed and source mechanisms are hypothesized. This work illustrates how advances in unsupervised machine learning can be leveraged to explore volcano infrasound data sets and demonstrates the potential of these techniques for monitoring eruptive activity.

1. Introduction

Infrasound signals (low frequency acoustic waves in the atmosphere) are commonly recorded at active volcanoes and are increasingly used in monitoring eruptive activity (Fee and Matoza, 2013; De Angelis et al., 2019). Infrasound signals are coupled to subaerial processes and can be used to detect eruptions (Arnoult et al., 2010; Coombs et al., 2018), estimate volume and height of erupted material (Caplan-Auerbach et al., 2010; Kim et al., 2015; Fee et al., 2017; Iezzi et al., 2019), track changes in eruptive activity (Garcès et al., 1999; Fee et al., 2010b), and provide constraints on magma movements (Richardson et al., 2014; Johnson et al., 2018; Cannavò et al., 2019; Watson et al., 2020).

There is a need to develop automatic processing workflows to analyze continuous volcano infrasound data and identify periods of unrest. Over the past two decades, there have been massive advances in machine learning (ML). ML algorithms are now commonly used to analyze large data sets in a range of disciplines and could be complementary to conventional processing workflows. ML techniques are becoming increasingly common in seismology (see Kong et al. (2019) for a review) and there are multiple volcano seismic studies that use ML techniques (e.g., Malfante et al., 2018; Anzieta et al., 2019; Hajian et al., 2019). There are far fewer ML volcano infrasound studies. Cannata et al. (2011a) associated periods of activity with specific vents at Mount Etna, Italy, using the density-based spatial clustering of applications with noise (DBSCAN) algorithm (Ester et al., 1996). Witsil and Johnson (2020) used *k*-means clustering (MacQueen, 1967) to identify shifts in eruptive activity at Stromboli, Italy.

Here, I apply the *k*-means clustering algorithm to three days of continuous infrasound data from Mount Etna. I generate a discrete time series of cluster labels and investigate how the cluster labels changed during a fissure eruption. This work illustrates the utility of unsupervised ML and *k*-means clustering for identifying changes in eruptive activity and extends the work of Witsil and Johnson (2020) by demonstrating that these techniques can be applied across different volcanoes.

2. Mount Etna and infrasound data

Mount Etna is a frequently-active open-vent volcano located on the eastern coast of Sicily, Italy. Activity at Mount Etna is characterized by strombolian explosions and lava fountaining at the summit along with fissure eruptions on the flanks (Allard et al., 2006; Behncke et al., 2014; De Beni et al., 2015) punctuated by occasional sub-plinian paroxysmal eruptions (Calvari et al., 2018).

Mount Etna is a prodigious source of infrasound with multiple active vents (Fig. 1a) (Marchetti et al., 2009; Cannata et al., 2011b, 2013). It is an open vent volcano and the infrasound signals often display harmonic peaks due to resonance at the summit craters (Sciotto et al., 2013; Spina et al., 2015; Cannavò et al., 2019; Watson et al., 2020). Previous work by

E-mail address: lwatson2@uoregon.edu.

https://doi.org/10.1016/j.jvolgeores.2020.107042
0377-0273/

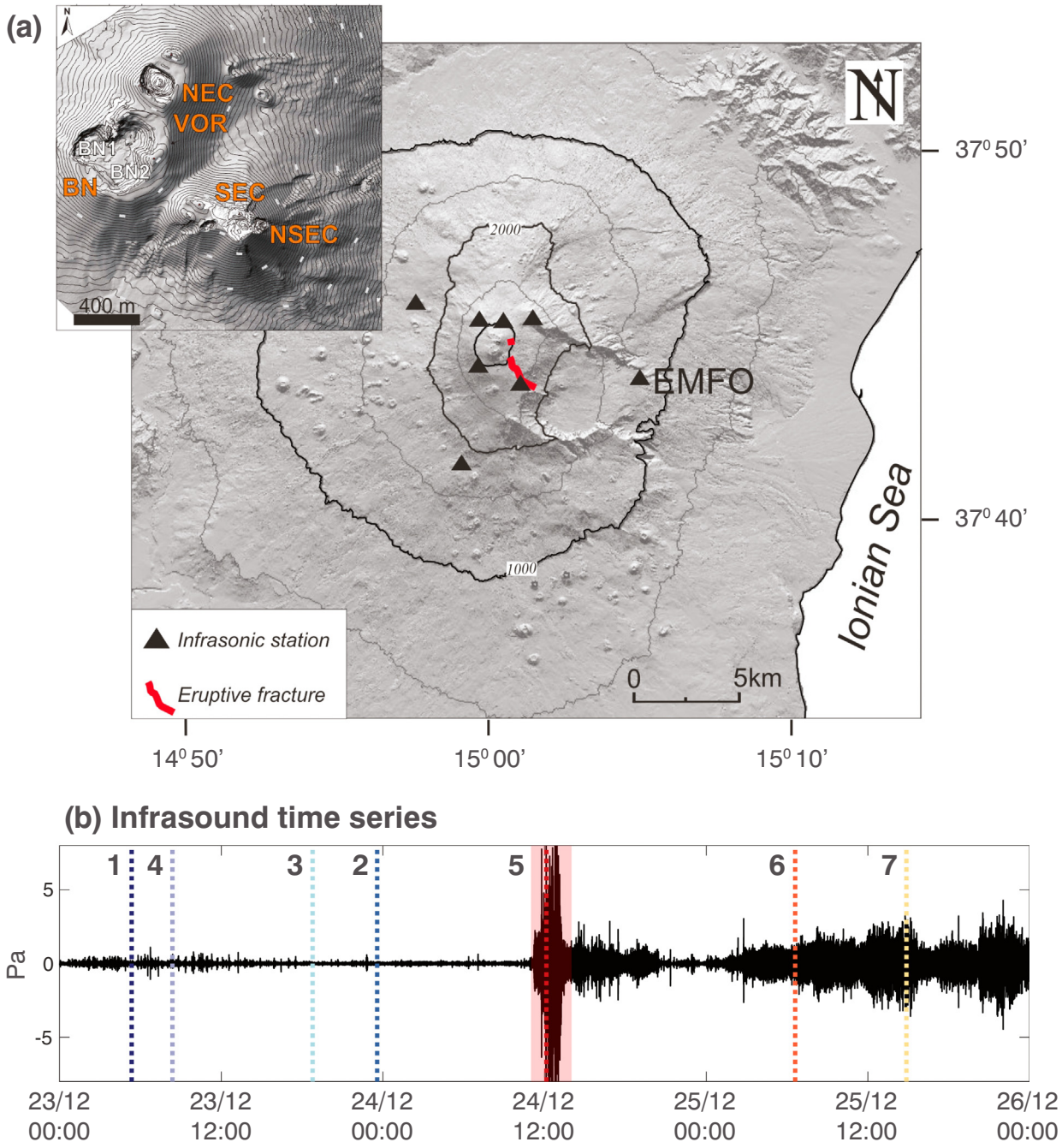

Fig. 1. (a) Map of Mount Etna showing the location of infrasound sensors (black triangles) and eruptive fracture (red). Inset shows named craters (VOR, Voragine; BN, Bocca Nuova; NEC, North-East Crater; SEC, South-East Crater; NSEC, New South-East Crater; from Neri et al. (2017)). Infrasound sources at the summit were predominantly located at BN-1 (Cannavò et al., 2019). (b) Infrasound time series recorded at station EMFO. Red shaded area shows the approximate duration of the explosive portion of the fissure eruption (Cannavò et al., 2019). Vertical dotted lines and cluster numbers indicate the timing of the time series shown in Fig. 7. Figure after Watson et al. (2020).

Ulivieri et al. (2013) related changes in infrasound observations to changes in eruptive behavior while Ripepe et al. (2018) observed that sustained eruption columns at Mount Etna are commonly preceded by violent strombolian activity and used this relationship to develop an infrasound-based early warning system.

In the second half of 2018, Mount Etna experienced a period of elevated activity (Global Volcanism Program, 2018) that culminated on 24 December 2018 with a fissure eruption on the south-east flank and intense strombolian activity at the summit craters (Cannavò et al., 2019; Global Volcanism Program, 2019; Laiolo et al., 2019). The eruptive fissure opened at 11:11 UTC (Cannavò et al., 2019) and erupted explosively for several hours.

The infrasound amplitude drastically increased after the onset of the fissure eruption. The peak frequency of the signal also decreased, which Cannavò et al. (2019) and Watson et al. (2020) interpreted to be caused by a draw-down of magma from the summit craters during the fissure eruption; draw-down of magma increased the two-way travel time for an acoustic wave that propagates from the top of the magma column and is reflected from crater outlet, and hence decreased the peak frequency (Watson et al., 2019).

The INGV maintains a network of eight infrasound stations around Mount Etna (Fig. 1a). Each station is equipped with a GRASS 40AN microphone that has a flat response in the frequency range of 0.3–20,000 Hz and a sensitivity of 50 mV/Pa. The infrasound data were sampled at 50 Hz and then re-sampled at 100 Hz in agreement with previous work on the same data set by Cannavò et al. (2019) who re-sampled the infrasound data to match seismic observations. This work focuses on 72 h of infrasound data from 23/12 00:00 to 26/12 00:00 UTC 2018 (Sciotto et al., 2020), which encompasses the duration of the fissure eruption, recorded at station EMFO (Fig. 1b), 8.1 km from Bocca Nuova. Despite being further from the summit, station EMFO has the highest signal to noise ratio as it is more sheltered whereas the summit stations are more affected by wind noise during bad weather. The summit stations also have issues with signal continuity during the winter months.

Cannavò et al. (2019) considered the same three days of infrasound data examined here and used a grid-search method based on brightness function for the events and semblance function for tremor (Cannata et al., 2011b, 2013) to locate events. Eruptive activity was predominantly centered at Bocca Nuova crater apart from 11:11–13:30 on 24 December when activity was located at the fissure (Cannavò et al., 2019).

3. Unsupervised machine learning and *K*-means clustering

Machine learning (ML) techniques are categorized as supervised or unsupervised. For supervised learning, algorithms are trained on labeled data to "learn" the relationship between inputs and outputs. Once trained, the algorithm can extrapolate to compute output values for new input data. Rather than try to determine the relationship between inputs and outputs, unsupervised ML techniques try to infer the natural structure present within a data set. Unsupervised techniques do not require labeled data or training data sets so are useful for data exploration.

A common unsupervised ML technique is clustering where a data set is grouped into clusters such that objects in the same cluster are more similar to each other than to objects in other clusters. There are many different clustering algorithms (for a review of clustering techniques see Saxena et al., 2017). Here, I use the *k*-means clustering algorithm, which is one of the most popular and widely used clustering algorithms.

k-means is an iterative algorithm that divides the data set into a specified *k* number of clusters (MacQueen, 1967). Fig. 2 shows how *k*-means clustering can be applied to continuous infrasound data. The time series is divided into *m* windows. For each window, *n* features are extracted that characterize the signal to create *m* objects that are *n*-dimensional. The *k*-means algorithm assigns each of the *m* objects in the *n*-dimensional feature space to *k* clusters.

The *k*-means algorithm is initialized by randomly placing *k* cluster centroids in the feature space. Objects are assigned to the closest centroid. The cluster centroids are then updated to the mean location of each cluster and the objects are redistributed to the closest centroid. This process is repeated until convergence or the maximum number of iterations is reached. Cluster size is quantified by the total within-cluster sum of squares (TWCSS), which is the sum of the distance between each object in the cluster and the cluster centroid. The *k*-means algorithm aims to minimize the sum of the TWCSS across all clusters, however, the algorithm can get trapped in local minima. Therefore, *k*-means clustering is often repeated multiple times with different initial cluster centroids to attempt to find the global minimum.

There are two main choices when applying the *k*-means clustering algorithm: (1) the features that are used to characterize the data set and (2) the number of clusters that the signal is categorized into. In the next two sections, I discuss these issues and the choices used in this work.

3.1. Features

I initially consider 11 different features in both the time and frequency domains that characterize the infrasound signal; (1) standard deviation, (2) skewness and (3) kurtosis of the time series amplitude distribution, (4) peak frequency and (5) quality factor, (6) standard deviation, (7) skewness and (8) kurtosis of frequency domain amplitude distribution, and the frequency at the (9) 25th, (10) 50th, and (11) 75th percentile of the cumulative spectral density. With the exception of the quality factor, these features were previously used by Witsil and Johnson (2020). The quality factor provides a measure of the amount of damping in the system and is inversely proportional to the exponential rate of amplitude decay in the time series (Rossing and Fletcher, 2004). I include the quality factor here because it has been shown to be diagnostic of eruptive behavior at open-vent volcanoes such as Mount Etna (Johnson et al., 2018; Watson et al., 2019).

Machine learning algorithms are most efficient without repeated or duplicate information (Li et al., 2008). Following the treatment of Witsil and Johnson (2020), I calculate the correlation between all features and remove the highly correlated features. The standard deviation and kurtosis of frequency domain amplitude distribution and the 25th and 75th percentiles of the cumulative spectra density are highly correlated with other features and are therefore removed from the analysis. The seven remaining features that are used in the clustering analysis are shown in Table 1 with the time series and histograms of each feature displayed in Fig. 3. Before applying the clustering algorithm, features are normalized to have a mean of zero and standard deviation of one so that all features are weighted equally and the clustering algorithm is not dominated by the features with the largest range.

When applying *k*-means to other volcanoes, care should be taken to explore all possible features and to remove redundant information. It is possible to combine two or more features to create new engineered features that can be more instructive for the clustering process (Garla and Brandt, 2012). However, this level of complexity is outside of the scope of this work. Future work could consider using dimension reduction techniques such as principal component analysis to identify the most important features (e.g., Celik, 2009). Infrasound amplitudes are not included because the amplitude depends strongly on the distance from the source to the receiver. During the time period considered, infrasound is generated at both the summit craters and the fissure on the south-east flank (Cannavò et al., 2019). Therefore, it is unclear if changes in infrasound amplitude reflect changes in source location or changes in eruptive activity and hence infrasound amplitude is not included as a feature. Note the infrasound amplitude may be a useful feature in other situations when considering a stationary source and examining changes in eruptive activity or atmospheric conditions.

In order to extract features, the 72 h of continuous data is divided into 5 min (300 s) windows with 90% overlap. This results in 8631 unique values for each of the seven features (Fig. 3). Previous work has used shorter window lengths to examine individual explosions (Witsil and Johnson, 2020). This work focuses on longer-term trends and qualitative changes in eruptive activity and hence uses longer windows. Five minute windows provide a balance between computational cost and efficiency while providing sufficient resolution to accurately capture changes in eruptive behavior on the time scales of interest. In this work, the clustering analysis is performed on the unprocessed infrasound data, which showcases the utility of unsupervised ML for data exploration. It can be performed early on in the analysis workflow and the results can inform other conventional processing techniques.

3.2. Number of clusters

A key parameter when applying the *k*-means clustering algorithm is how many clusters, *k*, to divide the data into. This is a challenging parameter to estimate as it depends on the complexity of the data as well as the desired resolution. Including too many clusters can make

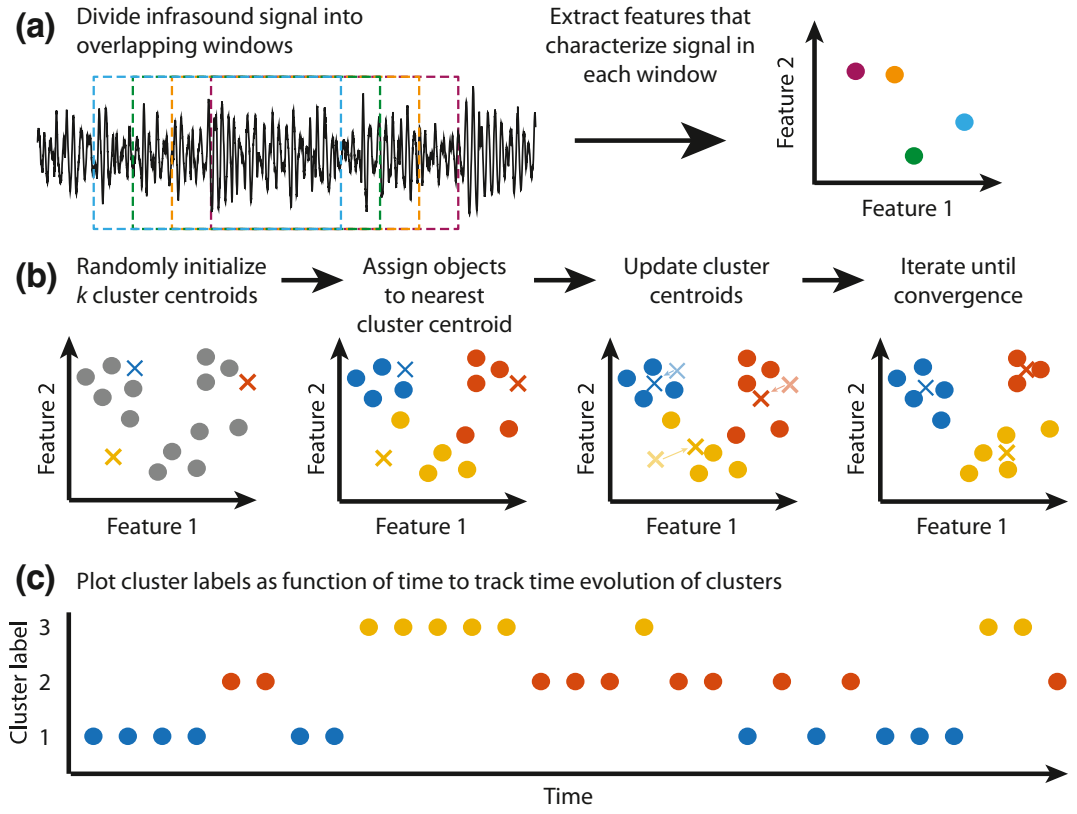

Fig. 2. Diagram showing how *k*-means clustering can be applied to continuous infrasound data. This example shows a two-dimensional feature space with data grouped into three clusters. (a) The infrasound signal is divided into *m* overlapping windows. For each window, *n* features are extracted. The set of *n* features extracted from a specific window is called an object and can be plotted in the *n* dimensional feature space. (b) The *k*-means clustering algorithm is used to categorize each object into one of *k* clusters. (c) Each object is associated with a specific time and therefore cluster labels can be plotted as a discrete function of time to track their time evolution.

the results difficult to interpret but not including enough clusters may cause important trends in the data to be missed.

When the optimal number of clusters is not known from a-priori information, there are three techniques that can be commonly used to determine the optimal value of *k*: (1) the elbow method (Thorndike, 1953; Sugar, 1998; Anzieta et al., 2019), (2) the gap statistic (Tibshirani et al., 2001; Witsil and Johnson, 2020), and (3) the silhouette method (Rousseeuw, 1987; Kaufman and Rousseeuw, 1990; Morales-Esteban et al., 2010; Mato and Toulkeridis, 2017). These methods all involve performing *k*-means clustering for a range of candidate *k* values and computing different metrics (Fig. 4). Here, I consider values of *k* between 2 and 20 as including more than 20 clusters makes interpreting the results excessively difficult while increasing computational expense. The algorithm is run 25 times with different initial cluster centroids for each test and the solution with the smallest sum of TWCSS is chosen.

The first method for determining the optimal number of clusters is the elbow method where the TWCSS is computed for each cluster and the sum of TWCSS is plotted for each candidate *k* value. The sum of TWCSS decreases monotonically with *k* as increasing the number of clusters results in smaller clusters. The optimal value of *k* is where there is an inflection (or elbow) in the curve, which indicates the point where further increasing the number of clusters results in only marginal improvements to the model. Unfortunately, the elbow method is inconclusive for the infrasound data analyzed here (Fig. 4a).

The second method is the gap statistic (Tibshirani et al., 2001). For this method, the clustering algorithm is applied to the observed data and a randomly generated data set that uniformly spans the feature space. The gap statistic is computed as the difference between the sum of the TWCSS for the observed and random data and the maximum of the gap statistic indicates the optimal number of clusters. This technique was used by Witsil and Johnson (2020) who analyzed infrasound data from Stromboli volcano, however, the gap statistic is inconclusive for the infrasound data analyzed here (Fig. 4b).

The third method is the silhouette method (Rousseeuw, 1987). The silhouette width measures how similar an object is to the cluster that it belongs to compared to how similar it is to other clusters. A high average silhouette width indicates a good clustering and the maximum of the average silhouette width indicates the optimal number of clusters. Fig. 4c shows that the silhouette result has a maximum at $k=7$. In the next section, therefore, I use seven clusters when analyzing the infrasound data.

Table 1
Table of features used in clustering analysis.

Feature	Description	Domain
1	Standard deviation	Time
2	Skewness	Time
3	Kurtosis	Time
4	Peak frequency	Frequency
5	Quality factor	Frequency
6	Skewness	Frequency
7	Frequency at 50th percentile	Frequency

4. Clustering analysis

The 72 h of continuous infrasound data recorded at station EMFO are divided into 300 s windows with 90% overlap. The seven features listed in Table 1 are extracted from each window to generate 8631 seven-dimensional objects. The objects are grouped into seven clusters using the *k*-means clustering algorithm. The clustering is repeated 25 times with different initial cluster centroids and the solution with the smallest TWCSS is chosen. Clusters are labeled by order of appearance.

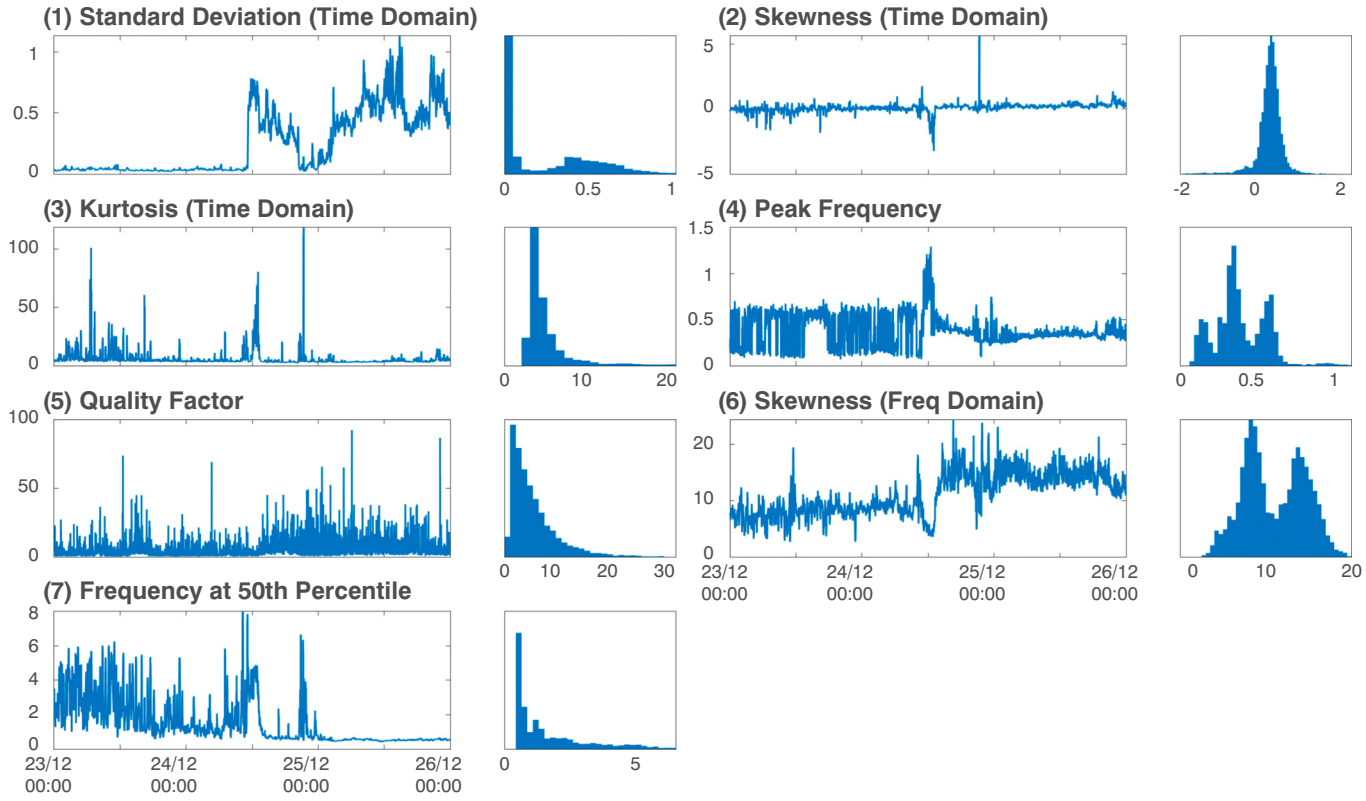

Fig. 3. Time series and histograms of the seven features used to characterize the infrasound signal. Features are calculated from the unfiltered infrasound signal shown in Fig. 1b using 300 s windows with 90% overlap.

The clustering analysis shows a distinct transition in cluster labels around 12:00 on 24 December (Fig. 5). Before this, the activity is predominantly K1, K2, K3 and K4. From 12:06 to 13:16 on 24 December, activity is dominated by K5. Afterward, the activity transitions to K6 and K7.

Properties of each cluster are listed in Table 2. K6 and K7 are the most common clusters and occur 21.1% and 23.2% of the time, respectively, while K5 is the least common and only occurs 1.9% of the time. In order to assess the robustness of a cluster designation, Witsil and Johnson (2020) introduced the cluster quality ratio, which is the percentage of cluster labels that are high-quality. They defined a cluster label as high-quality if it (1) repeated for at least four consecutive windows and (2) was consistent across at least two of the three infrasound channels. The infrasound data examined here was recorded by a single channel receiver and therefore only the first criterion is used in this study. K5 is the most robust cluster label with a quality ratio of 86.8% whereas K6 and K7 have the lowest quality ratios of 57.1% and 61.6%. This indicates that K5 is a stable cluster and activity remains predominantly in K5 during the cluster duration. By contrast, K6 and K7 are less stable. Activity frequently switches between these two clusters. Note that infrasound amplitude is not included in the clustering analysis. The clustering results, however, naturally partition by amplitude; low amplitude before the fissure eruption (K1, K2, K3, and K4), high amplitude during (K5), and moderate amplitude after (K6 and K7).

Feature distributions for the seven clusters are shown in Fig. 6. The mean and standard deviation for each feature are also listed in Table 3. Feature distributions are informative for showing similarities and differences between clusters as well as illustrating which features

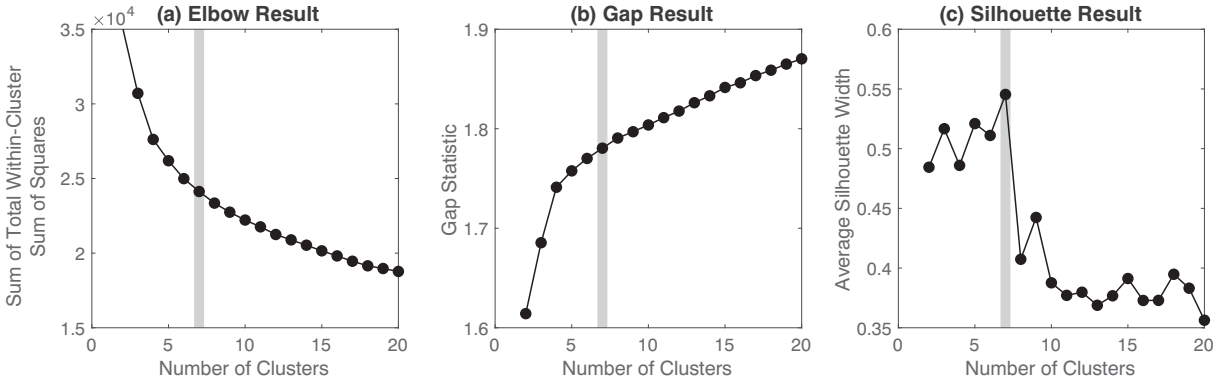

Fig. 4. Results from (a) elbow, (b) gap, and (c) silhouette tests for determining the optimal number of clusters. The elbow and gap results are inconclusive but the silhouette result suggests that seven is the optimal number of clusters (indicated by vertical gray line).

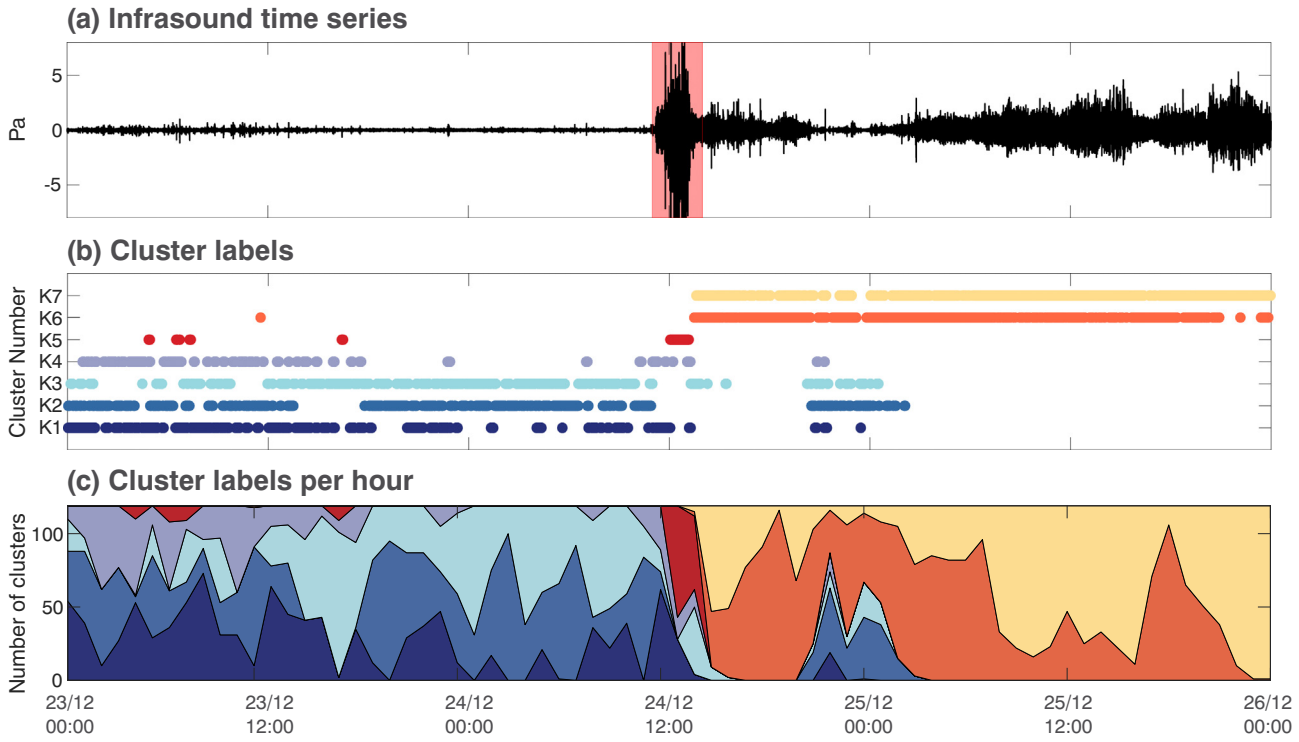

Fig. 5. (a) Time series of infrasound data. Red shaded area shows the approximate duration of the explosive portion of the fissure eruption (Cannavò et al., 2019). (b) Time evolution of clusters. (c) Area plot showing relative contribution of each cluster per hour.

are most diagnostic of a cluster. For example, K5 has a lower skewness (time) and higher kurtosis (time) and peak frequency than the other clusters. K1, K2, K3, and K4 have similar standard deviation (time). However, K4 has a higher quality factor and K2 has a lower peak frequency. Most of the features are similar for K6 and K7 but K6 has a larger quality factor and skewness (frequency) while K7 has a larger standard deviation (time).

Figs. 7 and 8 show example waveforms for each of the seven clusters. Fig. 7 shows 300 s windows where the cluster label, which is assigned for 30 s windows, is constant for the entire 300 s. Three representative 30 s windows for each cluster are shown in Fig. 8 and provide more detail of the waveforms. K1, K2, and K3 are characterized by low amplitude oscillations with intermittent larger amplitude transients. K4 and K5 display numerous impulsive transients, although the amplitude is far larger in K5. K6 and K7 have larger amplitude oscillations without any clear impulsive signals.

5. Discussion

There is a distinct transition in cluster labels on 24 December (Fig. 5). Cluster labels change from predominantly K1, K2, K3 and K4 before 12:06 to K6 and K7 after 13:16 while K5 dominates in between. The timing of K5 coincides with the emergence of a fissure on the southeast flank that erupted explosively for several hours starting at 11:11 (Cannavò et al., 2019; Laiolo et al., 2019), which suggests that K5 represents explosive activity at the fissure.

The distances between cluster centroids are shown in Table 4. The centroid of K5 is located far away from all other cluster centroids. This suggests that the signals in K5 are very different from any other cluster, which may be related to the source mechanism. This further supports the hypothesis that K5 is associated with explosive activity at the fissure while the other clusters are related to eruptive activity at the summit.

The diversity of waveforms (Figs. 7 and 8) suggests different source mechanisms for the different clusters. K1, K2, and K3 do not display any impulsive transients and can be described as infrasonic tremor (Fee and Garcès, 2007). Previous studies have observed infrasound tremor during lava fountaining at Mount Etna (Cannata et al., 2009), resonance at Kilauea (Hawaii) and Villarrica (Chile) (Garcés et al., 2003; Richardson et al., 2014), and gas jetting at Tungurahua (Ecuador) (Fee et al., 2010b). Mount Etna commonly generates infrasound with harmonic peaks, which have been attributed to resonance of the summit craters (Sciotto et al., 2013; Spina et al., 2015). K1 and K3 have similar peak frequency to that hypothesized for crater acoustic resonance (Cannavò et al., 2019) and may be caused by low-level eruptive activity at the summit. K2, however, has a much lower peak frequency and may be due to wind noise or the microbaram peak (Fee et al., 2010a). K4 and K5 display clear impulsive transients, which are representative of discrete explosions. The lower amplitude signals of K4 are likely from explosions at Bocca Nuova as they predominantly occur prior to the onset of the fissure eruption whereas the larger amplitude signals of K5 are likely caused by vigorous lava fountaining and explosions at the fissure. K6 and K7 are qualitatively similar to K1 and K3 with no impulsive transients. K6 and K7, however, are characterized by lower frequency and larger amplitude. Cannavò et al. (2019) observed that the peak frequency of the infrasound data decreased after the onset of the fissure eruption and inferred that this was caused by magma draining

Table 2
Cluster characteristics. Durations are calculated by multiplying the number of cluster labels by the 30 s unique window length. Quality ratio is the percentage of cluster labels that repeat at least four times in a row. Amplitude values are the mean peak amplitude and standard deviation for all 300 s windows in that cluster.

Cluster	Duration (%)	Duration (HH:MM)	Quality ratio (%)	Amplitude (Pa)
K1	12.3	08:52	67.0	0.43 ± 0.83
K2	18.4	13:14	71.5	0.20 ± 0.10
K3	16.3	11:45	73.5	0.21 ± 0.34
K4	6.7	04:51	72.6	0.74 ± 1.19
K5	1.9	01:23	86.8	7.92 ± 4.61
K6	21.1	15:11	57.1	1.23 ± 0.64
K7	23.2	16:44	61.6	2.14 ± 0.89

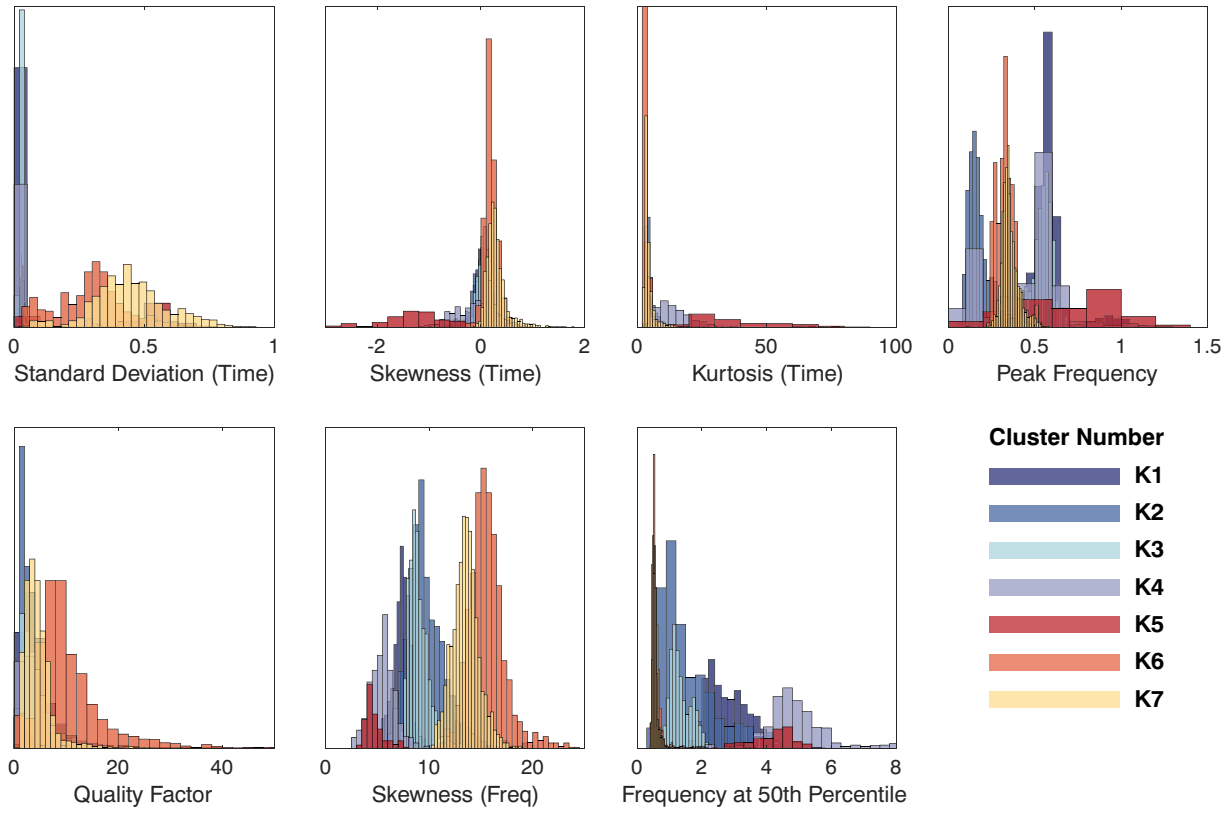

Fig. 6. Feature distributions (in physical units) for the seven clusters identified with the k-means clustering algorithm.

away from the summit, which increased the length of the resonating cavity and decreased the peak frequency. Therefore, K6 and K7 may reflect vigorous eruptive activity at the summit when the magma column was low in the crater. Future work could combine the clustering results presented here with event locations to shed more insight onto source mechanisms.

It is possible that several clusters share the same source mechanism. This does not mean, however, that the *k*-means clustering algorithm should have been run with fewer clusters. Due to the complex nature of the seven-dimensional feature space, it is challenging to predict how changing the number of clusters would change the results or interpretations. When using clustering analysis for data exploration the methods shown in Section 3.2 should be used to determine the optimal number of clusters. Alternatively, a clustering method that does not require the number of clusters to be specified could be used, such as density-based spatial clustering of applications with noise (DBSCAN), which was used to analyze infrasound signals from Mount Etna by Cannata et al. (2011a).

In this work, the average quality ratio is 70.0%, which is much higher than the average quality ratio of 5.9% that was calculated by Witsil and Johnson (2020). There are three reasons for this. First, when defining high-quality cluster labels, Witsil and Johnson (2020) applied an additional criterion that the cluster label had to be consistent across at least two of the three infrasound channels. This criterion could not be applied in this work because the infrasound receivers examined here were single channel. Second, Witsil and Johnson (2020) used much shorter windows (5 s windows with 90% overlap so that each 0.5 s of data is assigned a cluster label) to what is used here (300 s windows with 90% overlap so that each 30 s of data is assigned a cluster label). Using longer time windows averages over short duration variations in eruption amplitude or style and results in more stable cluster labels. Third, Witsil and Johnson (2020) used 16 clusters compared to the seven clusters considered here. Decreasing the number of clusters results in fewer clusters labels to alternate between and hence more stable cluster labels. Future work should examine the sensitivity of the clustering analysis to the window length.

In this work, *k*-means clustering is applied to the unfiltered infrasound data. This is a deliberate choice as the goal of this work is to showcase how unsupervised machine learning can be used as a data exploration tool early on in the processing workflow. Future work could examine how the clustering results depend on the filtering applied.

Table 3
Mean and standard deviation of all features (in physical units) for each cluster.

Cluster	Standard deviation (time)	Skewness (time)	Kurtosis (time)	Peak freq	Q factor	Skewness (frequency)	Freq at 50th percentile
K1	0.06 ± 0.12	−0.02 ± 0.16	5.33 ± 2.09	0.57 ± 0.13	4.73 ± 5.17	7.34 ± 0.82	2.81 ± 0.64
K2	0.03 ± 0.01	0.02 ± 0.24	5.23 ± 2.78	0.17 ± 0.06	3.14 ± 3.54	9.75 ± 1.95	1.59 ± 0.89
K3	0.04 ± 0.05	0.07 ± 0.16	4.17 ± 1.95	0.54 ± 0.07	4.51 ± 4.50	9.00 ± 1.06	1.33 ± 0.34
K4	0.06 ± 0.11	−0.15 ± 0.36	15.34 ± 9.00	0.45 ± 0.21	3.91 ± 4.11	5.52 ± 1.20	4.87 ± 0.92
K5	0.42 ± 0.22	−1.36 ± 0.59	37.01 ± 16.21	0.70 ± 0.28	2.57 ± 1.75	4.72 ± 0.92	4.19 ± 0.62
K6	0.29 ± 0.13	0.22 ± 0.39	4.02 ± 7.79	0.33 ± 0.05	10.61 ± 7.14	15.57 ± 1.76	0.60 ± 0.22
K7	0.45 ± 0.14	0.29 ± 0.22	4.45 ± 3.29	0.36 ± 0.05	4.61 ± 3.39	13.53 ± 1.27	0.56 ± 0.11

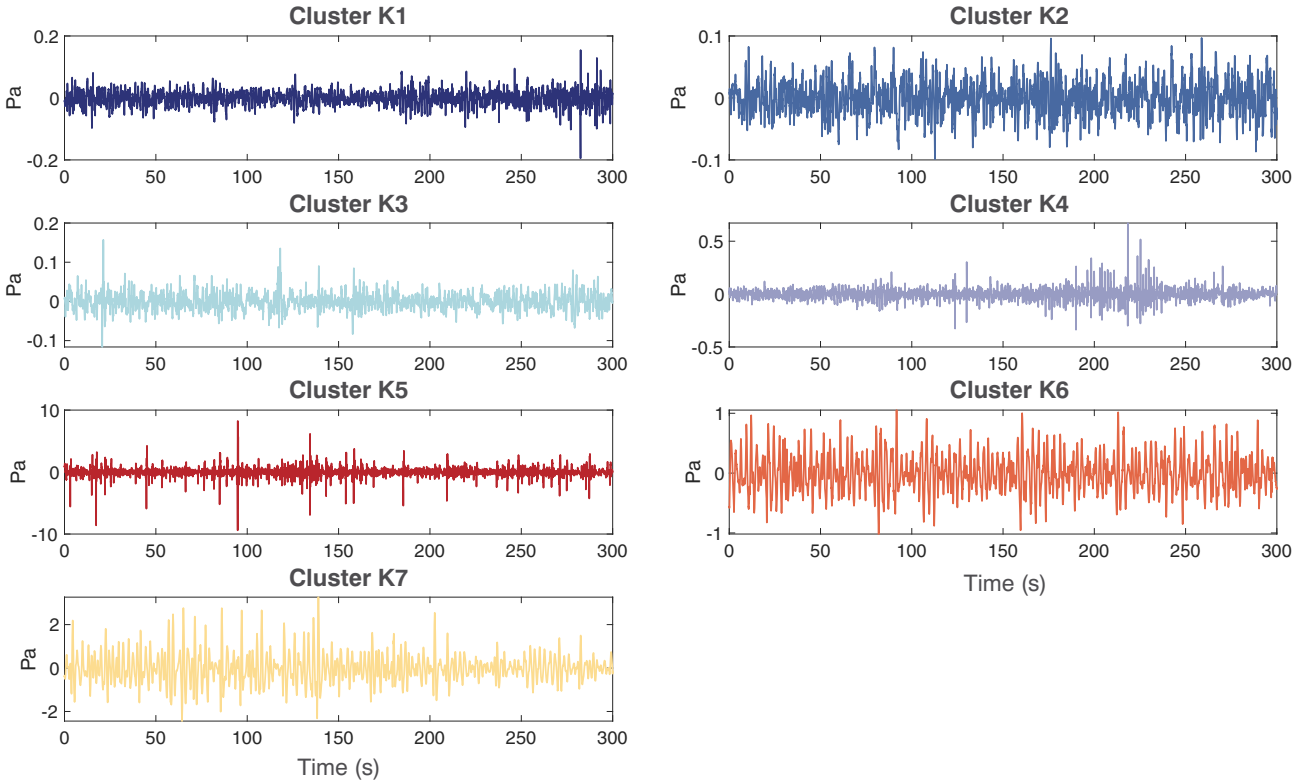

Fig. 7. Example 300 s time series of the seven different clusters. Cluster labels are assigned to 30 s unique windows and are constant for the entire 300 s shown. Cluster 1: Dec 23, 05:18:30 to 05:23:30. Cluster 2: Dec 23, 23:32:00 to 23:37:00. Cluster 3: Dec 23, 18:44:30 to 18:49:30. Cluster 4: Dec 23, 08:20:00 to 08:25:00. Cluster 5: Dec 24, 12:05:30 to 12:10:30. Cluster 6, Dec 25, 06:33:30 to 06:38:30. Cluster 7, Dec 25, 14:49:00 to 14:54:00. The timing of each window is shown graphically in Fig. 1.

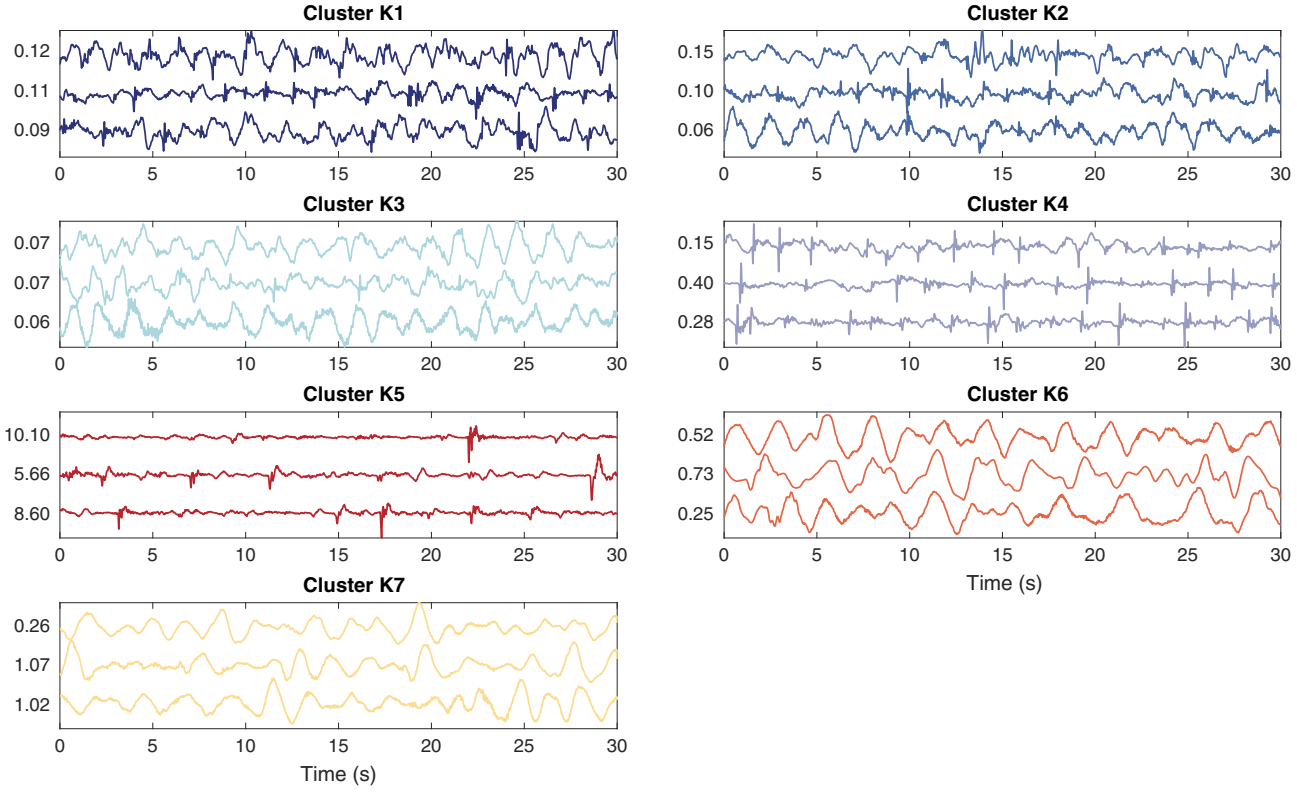

Fig. 8. Example 30 s waveforms for each cluster with trace-normalized scaling. The number to the left of each trace indicates the maximum absolute amplitude (Pa) in the 30 s window.

Table 4
Distance between cluster centroids. The distance is calculated in the normalized feature space that is used for the clustering analysis.

	K1	K2	K3	K4	K5	K6	K7
K1	0	5.82	6.28	4.50	16.58	2.94	7.65
K2		0	2.43	7.58	16.24	5.68	10.95
K3			0	8.27	18.25	6.37	11.64
K4				0	12.66	1.96	3.86
K5					0	14.57	9.95
K6						0	5.53
K7							0

The clustering results presented here show a distinct change in cluster labels associated with a change in eruptive behavior due to the opening of a fissure on the south-east flank of the volcano. This suggests that machine learning techniques could be used for real-time infrasound-based volcano monitoring to identify changes in eruptive activity. A simple clustering-based workflow is to first apply *k*-means to classify a training data set into *k* clusters. Incoming data could then be assigned to existing clusters. If incoming data is sufficiently different to the existing clusters, according to a specified distance metric, it would indicate a departure from the eruptive behavior seen in the training data set. This workflow could be used to highlight changes in eruptive behavior in real-time.

6. Conclusion

Unsupervised machine learning (ML) is used to categorize continuous volcano infrasound data recorded at Mount Etna. The clustering analysis performed here captures a distinct change in eruptive behavior due to a fissure eruption on the south-east flank. Analyzing cluster feature distributions sheds insight on infrasound source mechanisms and provides information about eruption dynamics. This work illustrates how unsupervised ML can be used to explore infrasound data and highlight changes in eruptive activity. Future research should apply clustering algorithms in real-time to identify changes in eruptive behavior and aid in monitoring efforts.

Unsupervised ML algorithms such as *k*-means clustering are complementary to existing processing workflows. These methods can be applied to continuous infrasound data and used to identify trends and gain an intuitive understanding of the data space, which can inform further conventional data processing and analysis efforts. This study builds upon the work by Witsil and Johnson (2020) at Stromboli and demonstrates that unsupervised ML can be applied generally to analyze infrasound signals from any volcano. Here, I focus on volcano infrasound observations but clustering algorithms such as *k*-means are general processing tools that can be used to analyze other geophysical data such as seismic and geodetic data or continuous gas measurements.

CRediT authorship contribution statement

Leighton M. Watson: Conceptualization, Methodology, Software, Investigation, Writing - original draft, Visualization, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I am indebted to the technicians of the INGV, Osservatorio Etneo, for enabling the acquisition of the infrasound data. In accordance with INGV's data policy, infrasound data is available at https://doi.org/10.13127/etna_infra/raw_20181223_25. I would like to thank M. Sciotto and A. Cannata for sharing their extensive knowledge of Mount Etna, A. Witstil for informative discussions about unsupervised machine learning techniques and the inspiration for this work, and the volcanology students at the University of Oregon and Stanford University for their constructive comments. This work was supported by National Science Foundation grant EAR-1949219.

References

Allard, P., Behncke, B., D'Amico, S., Neri, M., Gambino, S., 2006. Mount Etna 1993–2005: Anatomy of an evolving eruptive cycle. Earth Sci. Rev. 78 (1–2), 85–114.
Anzieta, J.C., Ortiz, H.D., Arias, G.L., Ruiz, M.C., 2019. Finding Possible Precursors for the 2015 Cotopaxi Volcano Eruption Using Unsupervised Machine Learning Techniques.
Arnoult, K.M., Olson, J.V., Szuberla, C.A.L., McNutt, S.R., Garcés, M.A., Fee, D., Hedlin, M.A.H., 2010. Infrasound observations of the 2008 explosive eruptions of Okmok and Kasatochi volcanoes, Alaska. J. Geophys. Res.-Atmos. 115, 1–12.
Behncke, B., Branca, S., Corsaro, R. A., De Beni, E., Miraglia, L., Proietti, C., 2014. The 2011–2012 summit activity of Mount Etna: birth, growth and products of the new SE crater. J. Volcanol. Geotherm. Res. 270 (May 1971), 10–21.
Calvari, S., Cannavò, F., Bonaccorso, A., Spampinato, L., Pellegrino, A.G., 2018. Paroxysmal explosions, lava fountains and ash plumes at Etna Volcano: eruptive processes and hazard implications. Front. Earth Sci. 6 (107).
Cannata, A., Montalto, P., Privitera, E., Russo, G., Gresta, S., 2009. Tracking eruptive phenomena by infrasound: May 13, 2008 eruption at Mt. Etna. Geophys. Res. Lett. 36 (5), L05304.
Cannata, A., Montalto, P., Aliotta, M., Cassisi, C., Pulvirenti, A., Privitera, E., Patanè, D., 2011a. Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques. Geophys. J. Int. 185 (1), 253–264.
Cannata, A., Sciotto, M., Spampinato, L., Spina, L., 2011b. Insights into explosive activity at closely-spaced eruptive vents using infrasound signals: example of Mt. Etna 2008 eruption. J. Volcanol. Geotherm. Res. 208 (1–2), 1–11.
Cannata, A., Di Grazia, G., Aliotta, M., Cassisi, C., Montalto, P., Patanè, D., 2013. Monitoring Seismo-volcanic and Infrasonic Signals at Volcanoes: Mt. Etna Case Study. Pure Appl. Geophys. 170 (11), 1751–1771.
Cannavò, F., Sciotto, M., Cannata, A., Di Grazia, G., 2019. An Integrated geophysical approach to track magma intrusion: the 2018 Christmas eve eruption at Mount Etna. Geophys. Res. Lett. 46 (14), 8009–8017.
Caplan-Auerbach, J., Bellesiles, A., Fernandes, J.K., 2010. Estimates of eruption velocity and plume height from infrasonic recordings of the 2006 eruption of Augustine Volcano, Alaska. J. Volcanol. Geotherm. Res. 189 (1–2), 12–18.
Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEE Geoscience and Remote Sensing Letters 6 (4), 772–776.
Coombs, M.L., Wech, A.G., Haney, M.M., Lyons, J.J., Schneider, D.J., Schwaiger, H.F., Wallace, K.L., Fee, D., Freymueller, J.T., Schaefer, J.R., Tepp, G., 2018. Short-term forecasting and detection of explosions during the 2016–2017 eruption of Bogoslof Volcano, Alaska. Front. Earth Sci. 6.
De Angelis, S., Diaz-Moreno, A., Zuccarello, L., De Angelis, S., Diaz-Moreno, A., Zuccarello, L., 2019. Recent developments and applications of acoustic infrasound to monitor volcanic emissions. Remote Sens. 11 (11), 1302.
De Beni, E., Behncke, B., Branca, S., Nicolosi, I., Carluccio, R., D'Ajello Caracciolo, F., Chiappini, M., 2015. The continuing story of Etna's New Southeast Crater (2012–2014): evolution and volume calculations based on field surveys and aerophotogrammetry. J. Volcanol. Geotherm. Res. 303, 175–186 (February 2013).
Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases With Noise. Tech. Rep.
Fee, D., Garcès, M., 2007. Infrasonic tremor in the diffraction zone. Geophys. Res. Lett. 34 (16).
Fee, D., Matoza, R.S., 2013. An overview of volcano infrasound: from hawaiian to plinian, local to global. J. Volcanol. Geotherm. Res. 123–137.
Fee, D., Garcès, M., Patrick, M., Chouet, B., Dawson, P., Swanson, D., 2010a. Infrasonic harmonic tremor and degassing bursts from Halema'uma'u Crater, Kilauea Volcano, Hawaii. J. Geophys. Res. 115 (B11), B11316.
Fee, D., Garces, M., Steffke, A., 2010b. Infrasound from Tungurahua Volcano 2006–2008: Strombolian to Plinian eruptive activity. J. Volcanol. Geotherm. Res. 193 (1–2), 67–81.
Fee, D., Izbekov, P., Kim, K., Yokoo, A., Lopez, T., Prata, F., Kazahaya, R., Nakamichi, H., Iguchi, M., 2017. Eruption mass estimation using infrasound waveform inversion and ash and gas measurements: evaluation at Sakurajima Volcano, Japan. Earth Planet. Sci. Lett. 480, 42–52.
Garcès, M., Iguchi, M., Ishihara, K., Morrissey, M., Sudo, Y., Tsutsui, T., 1999. Infrasonic precursors to a vulcanian eruption at Sakurajima Volcano, Japan. Geophys. Res. Lett. 26 (16), 2537–2540.
Garcés, M., Harris, A., Hetzer, C., Johnson, J., Rowland, S., Marchetti, E., Okubo, P., 2003. Infrasonic tremor observed at Klauea Volcano, Hawai'i. Geophys. Res. Lett. 30 (20).
Garla, V.N., Brandt, C., 2012. Ontology-guided feature engineering for clinical text classification. J. Biomed. Inform. 45 (5), 992–998.
Global Volcanism Program, 2018. Report on Etna (Italy). In: Crafford, A.E., Venzke, E. (Eds.), Tech. Rep. Smithsonian Institution, p. 12.
Global Volcanism Program, 2019. Report on Etna (Italy). In: Crafford, A.E., Venzke, E. (Eds.), Tech. Rep. Smithsonian Institution, p. 4.

Hajian, A., Cannavò, F., Greco, F., Nunnari, G., 2019. Classication of Mount Etna (Italy) volcanic activity by machine learning approaches. Ann. Geophys. 62, 231.
Iezzi, A.M., Fee, D., Kim, K., Jolly, A.D., Matoza, R.S., 2019. Three-dimensional acoustic multipole waveform inversion at Yasur Volcano, Vanuatu. J. Geophys. Res. Solid Earth 124, 1–25 (2018JB017073).
Johnson, J.B., Watson, L.M., Palma, J.L., Dunham, E.M., Anderson, J.F., 2018. Forecasting the eruption of an open-vent volcano using resonant infrasound tones. Geophys. Res. Lett. 1–8.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York.
Kim, K., Fee, D., Yokoo, A., Lees, J.M., 2015. Acoustic source inversion to estimate volume flux from volcanic explosions. Geophys. Res. Lett. 42 (13), 5243–5249.
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. Seismol. Res. Lett. 90 (1), 3–14.
Laiolo, M., Ripepe, M., Cigolini, C., Coppola, D., Della Schiava, M., Genco, R., Innocenti, L., Lacanna, G., Marchetti, E., Massimetti, F., Silengo, M.C., 2019. Space- and ground-based geophysical data tracking of magma migration in shallow feeding system of Mount Etna Volcano. Remote Sens. 11 (10), 1182.
Li, Y., Dong, M., Hua, J., 2008. Localized feature selection for clustering. Pattern Recogn. Lett. 29, 10–18.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. vol 1, pp. 281–297 Oakland, CA, USA.
Malfante, M., Dalla Mura, M., Metaxian, J.P., Mars, J.I., Macedo, O., Inza, A., 2018. Machine learning for volcano-seismic signals: challenges and perspectives. IEEE Signal Process. Mag. 35 (2), 20–30.
Marchetti, E., Ripepe, M., Ulivieri, G., Caffo, S., Privitera, E., 2009. Infrasonic evidences for branched conduit dynamics at Mt. Etna volcano, Italy. Geophys. Res. Lett. 36 (19), L19308.
Mato, F., Toulkeridis, T., 2017. An unsupervised K-means based clustering method for geophysical post-earthquake diagnosis. 2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings. Institute of Electrical and Electronics Engineers Inc, pp. 1–8.
Morales-Esteban, A., Martínez-Álvarez, F., Troncoso, A., Justo, J.L., Rubio-Escudero, C., 2010. Pattern recognition to forecast seismic time series. Expert Syst. Appl. 37 (12), 8333–8342.
Neri, M., De Maio, M., Crepaldi, S., Suozzi, E., Lavy, M., Marchionatti, F., Calvari, S., Buongiorno, M.F., 2017. Topographic maps of Mount Etna's summit craters, updated to December 2015. Journal of Maps 13 (2), 674–683.
Richardson, J.P., Waite, G.P., Palma, J.L., 2014. Varying seismic-acoustic properties of the fluctuating lava lake at Villarrica volcano, Chile. J. Geophys. Res. Solid Earth 1–14.
Ripepe, M., Marchetti, E., Delle Donne, D., Genco, R., Innocenti, L., Lacanna, G., Valade, S., 2018. Infrasonic earlywarning system for explosive eruptions. J. Geophys. Res. Solid Earth 123 (11), 9570–9585.
Rossing, T.D., Fletcher, N.H., 2004. Principles of Vibration and Sound, 2nd Edition. Vol. 1. Springer.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (C), 53–65.
Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T., 2017. A review of clustering techniques and developments. Neurocomputing 267, 664–681.
Sciotto, M., Cannata, A., Gresta, S., Privitera, E., Spina, L., 2013. Seismic and infrasound signals at Mt. Etna: modeling the north-east crater conduit and its relation with the 2008–2009 eruption feeding system. J. Volcanol. Geotherm. Res. 254, 53–68.
Sciotto, M., Cassisi, C., Scuderi, L., Torrisi, O., 2020. Infrasonic Signals Recorded at Mt. Etna by EMFO, EPDN, ECPN Stations During Period 23–25 December 2018 (Version 1). [Data Set]. Istituto Nazionale di Geofisica e Vulcanologia (INGV).
Spina, L., Cannata, A., Privitera, E., Vergniolle, S., Ferlito, C., Gresta, S., Montalto, P., Sciotto, M., 2015. Insights into Mt. Etna's shallow plumbing system from the analysis of infrasound signals, August 2007–December 2009. Pure Appl. Geophys. 172 (2), 473–490.
Sugar, C.A., 1998. Techniques for Clustering and Classification With Applications to Medical Problems. Stanford University, Phd.
Thorndike, R.L., 1953. Who belongs in the family? Psychometrika 18 (4), 267–276.
Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2), 411–423.
Ulivieri, G., Ripepe, M., Marchetti, E., 2013. Infrasound reveals transition to oscillatory discharge regime during lava fountaining: implication for early warning. Geophys. Res. Lett. 40 (12), 3008–3013.
Watson, L.M., Dunham, E.M., Johnson, J.B., 2019. Simulation and inversion of harmonic infrasound from open-vent volcanoes using an efficient quasi-1D crater model. J. Volcanol. Geotherm. Res. 380, 64–79.
Watson, L.M., Johnson, J.B., Sciotto, M., Cannata, A., 2020. Changes in Crater Geometry Revealed by Inversion of Harmonic Infrasound Observations: 24 December 2018 Eruption of Mount Etna. Italy, Geophysical Research Letters.
Witsil, A.J.C., Johnson, J.B., 2020. Analyzing Continuous Infrasound From Stromboli Volcano. Italy using unsupervised machine learning, Computers and Geoscience.