Journal of Volcanology and Geothermal Research 320 (2016) 58–74

Contents lists available at ScienceDirect

Journal of Volcanology and Geothermal Research

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

Principal component analysis vs. self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra

K. Unglert, V. Radić, A.M. Jellinek

Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada

ARTICLE INFO

Article history:
Received 12 December 2015
Received in revised form 11 April 2016
Accepted 11 April 2016
Available online 20 April 2016

Keywords:
Volcano seismicity
Pattern recognition
Hierarchical clustering
Principal component analysis
Self-organizing maps

ABSTRACT

Variations in the spectral content of volcano seismicity related to changes in volcanic activity are commonly identified manually in spectrograms. However, long time series of monitoring data at volcano observatories require tools to facilitate automated and rapid processing. Techniques such as self-organizing maps (SOM) and principal component analysis (PCA) can help to quickly and automatically identify important patterns related to impending eruptions. For the first time, we evaluate the performance of SOM and PCA on synthetic volcano seismic spectra constructed from observations during two well-studied eruptions at Klauea Volcano, Hawai'i, that include features observed in many volcanic settings. In particular, our objective is to test which of the techniques can best retrieve a set of three spectral patterns that we used to compose a synthetic spectrogram. We find that, without a priori knowledge of the given set of patterns, neither SOM nor PCA can directly recover the spectra. We thus test hierarchical clustering, a commonly used method, to investigate whether clustering in the space of the principal components and on the SOM, respectively, can retrieve the known patterns. Our clustering method applied to the SOM fails to detect the correct number and shape of the known input spectra. In contrast, clustering of the data reconstructed by the first three PCA modes reproduces these patterns and their occurrence in time more consistently. This result suggests that PCA in combination with hierarchical clustering is a powerful practical tool for automated identification of characteristic patterns in volcano seismic spectra. Our results indicate that, in contrast to PCA, common clustering algorithms may not be ideal to group patterns on the SOM and that it is crucial to evaluate the performance of these tools on a control dataset prior to their application to real data.

1. Introduction

In volcano monitoring, scientists are faced with the task of correctly identifying patterns of unrest critically indicative of impending of eruptions (e.g. Sparks et al., 2012; Carniel, 2014). A key component of volcano monitoring is seismic activity (Sparks et al., 2012). Seismic signals on volcanoes can be classified in terms of their frequency content: Whereas volcano tectonic earthquakes often have a broadband spectrum, low frequency seismicity including long-period events, very long period events, and volcanic tremor predominantly cover lower frequency ranges of 0.01–5 Hz (Fehler, 1983; Neuberg, 2000; McNutt and Nishimura, 2008; Chouet and Matoza, 2013). Based on its distinct spectral properties, this low frequency seismicity is commonly explained by processes involving fluid movement: Examples include moving bubbles (Ripepe and Gordeev, 1999; Matoza et al., 2010; Jones et al., 2012), gas accumulation (e.g. Johnson et al., 1998; Lesage et al., 2006), resonating fluid pathways (Chouet, 1986; Leet, 1988; Julian, 1994; Benoit and McNutt, 1997; Neuberg et al., 2000; Hellweg, 2000; Balmforth et al., 2005), or bubble/magma flow (Denlinger and Hoblitt, 1999; Jellinek and Bercovici, 2011; Thomas and Neuberg, 2012; Dmitrieva et al., 2013; Lyons et al., 2013). Each of these mechanisms imply a relationship between properties of low frequency seismicity and volcanic activity. Indeed, approximately 80% of a global sample of volcanic tremor episodes have been shown to precede or accompany volcanic eruptions (McNutt, 1992). For a given volcanic setting, knowledge of typical seismicity and the corresponding spectral patterns before, during, and after eruptions (e.g Carniel et al., 1996; Unglert and Jellinek, 2015) is thus crucial for eruption forecasting.

A common approach to analyzing the temporal evolution of volcano seismicity is the visual inspection of spectrograms. For example, Unglert and Jellinek (2015) identify two characteristic phases of seismicity that accompanied two intrusions at Klauea Volcano, Hawai'i. This kind of analysis requires manual identification of characteristic spatio-temporal patterns, which is practically cumbersome, inherently subjective, and informed by the experience of the analyst. For instance, which spectral properties distinguish non-eruptive from eruptive unrest is unclear. Consequently, to be able to objectively identify patterns and extract key information related to imminent or active volcanism, analysts are increasingly reliant on automated algorithms (e.g., Carniel, 2014; Cortes et al., 2015).

Pattern recognition and machine learning methods provide a possible solution and are used in a wide range of disciplines (e.g., Kaski et al., 1998; Oja et al., 2002; Bishop, 2006). In particular, "unsupervised" methods imply that no a-priori knowledge of patterns is necessary, i.e. the algorithm self-learns from the data (e.g. Bishop, 2006; Langer et al., 2009).

http://dx.doi.org/10.1016/j.jvolgeores.2016.04.014
0377-0273/© 2016 Elsevier B.V. All rights reserved.

In volcano monitoring, this feature is essential because the temporal evolution of patterns in monitoring time series is often unknown (e.g. Sparks et al., 2012). A good review of different, unsupervised feature extraction methods and their application to volcano seismicity can be found in Orozco-Alzate et al. (2012), Carniel (2014). Such studies have used self-organizing maps (SOM) and other techniques to detect different types of seismicity (e.g., Carniel, 1996; Langer et al., 2009; Carniel et al., 2013b; Curilem et al., 2014), or link changes in time series from volcano monitoring with different eruptive vents or type of eruptions (e.g., Esposito et al., 2008; Di Salvo et al., 2013). Several studies first use SOM to reduce the amount of data to be analyzed, and subsequently apply clustering algorithms to obtain final groupings (e.g., De Matos et al., 2006; Köhler et al., 2009; Messina and Langer, 2011; Carniel et al., 2013b).

SOM can generate a visual representation of the similarities and differences between patterns in a dataset (e.g., Esposito et al., 2008), require no a-priori knowledge of patterns (e.g., Murtagh and Hernández-Pajares, 1995), and can thus be useful for detecting distinctive spectral characteristics of volcanic tremor. In fields such as oceanography or meteorology, it is common to evaluate pattern recognition techniques against each other, against other methods, and with synthetic data (e.g., Reusch et al., 2005; Liu et al., 2006). In seismology, different methods including SOM have been tested against each other at individual volcanic settings (e.g., Langer et al., 2009; Cortes et al., 2015), and SOM performance has been tested with artificial data consisting of parameters from the time and frequency domains (e.g., Köhler et al., 2009). However, to our knowledge no studies applying SOM combined with cluster analysis to volcanic tremor evaluate the functionality of SOM in spectral space with a synthetic dataset. Thus, the following key knowledge gaps persist:

1. The performance of SOM against more standard techniques such as principal component analysis (PCA) has not been systematically evaluated with synthetic datasets of spectra.
2. Appropriate benchmarking datasets closely aligned with real observations, and with known patterns and their occurrence in time do not exist (Orozco-Alzate et al., 2012).
3. It is not clear that the features of interest (e.g., relative spectral power at different frequencies, occurrence and evolution of various spectral shapes in time) are captured by SOM, or how noise affects the results. The limitations of the method in terms of its application to volcano seismic spectra are thus unclear.

Accordingly, in Section 2, we produce synthetic spectra on the basis of detailed manual extraction of two characteristic spectral signatures during eruptive periods at Klauea Volcano, Hawai'i (Unglert and Jellinek, 2015). Specifically, we address two questions:

1. Can hierarchical clustering, a common approach to identify groupings in data used in previous studies of volcano seismicity, applied to the results from PCA (Section 3) and SOM (Section 4) correctly identify the known spectra and their occurrence/evolution in time in a typical volcanic spectrogram?
2. How do the clustering results differ between the two techniques, and what are the limitations of each of the methods and of our synthetic dataset (Section 5)?

2. Data and preprocessing

Many methods exist for classifying volcano seismicity in both the time and the frequency domains (e.g., Langer and Falsaperla, 2003; Ibs-von Seht, 2008; Curilem et al., 2009). However, Castro-Cabrera et al. (2014) found that classification utilizing entire spectra performs better than classification utilizing sets of other temporal and spectral parameters such as the mean amplitude or mean frequency in a given time window. Furthermore, previous work shows that distinct spectral shapes and transitions between them may relate to the underlying physical processes (e.g., Aki et al., 1977; Benoit and McNutt, 1997; Maryanto et al., 2008; Unglert and Jellinek, 2015).

To evaluate whether SOM or PCA are suitable for automated analysis of time varying spectral signatures, and more accurate and efficient than visual inspection of spectrograms, we create a synthetic dataset of volcano seismic spectra on the basis of the major spectral characteristics of seismic signals from Klauea Volcano, Hawai'i between 2007 and 2011 (Fig. 1(a); Unglert and Jellinek (2015)). During this period, two dike intrusions and accompanying fissure eruptions in the East Rift Zone showed a phase of discrete, seismic events near the intruding dikes (Phase I, Figs. 1(b) and 2), followed by a phase of continuous tremor near the summit (Phase II, Figs. 1(b) and 2) with a stronger decrease of spectral power from low to high frequencies compared to Phase I (Unglert and Jellinek, 2015). A prominent feature of Phase II is gliding of spectral lines (Unglert and Jellinek, 2015). However, because the gliding was expressed at different frequencies during the two intrusions, and because it did not affect the overall character of Phase II, we do not include gliding spectral peaks in Phase II of our synthetic dataset. Such gradual variations in the frequencies of individual spectral peaks are, in principle, similar to transitions over time from one phase to another, which are included in our synthetic dataset. We touch upon this subject again in Section 5.3.

The eruptive phases at Klauea and their temporal variations are not representative of volcano seismicity in general, but they capture some of the main features of pre- and syn-eruptive seismicity observed in other settings such as Redoubt Volcano, (Fig. 1(c)) or Okmok Volcano, (Fig. 1(d)), such as different spectral shapes and impulsive and emergent variations of those shapes over time (e.g., Carniel et al., 1996; Neuberg, 2000; Ruiz et al., 2006; Curilem et al., 2009; Langer et al., 2009; Buurman et al., 2012). The particular value of our dataset is that it enables reliable performance evaluation of SOM and PCA on well understood data that are drawn from well-established observations.

2.1. Synthetic spectra

To create the three spectra, we choose three 5-minute windows of continuous seismic data corresponding to the background state, Phase I, and Phase II at station AHU from Klauea as described above (Figs. 1–2). Station AHU is situated between the inferred locations of Phase I and II (close to the area of dike intrusion and below the summit, respectively) and showed both phases clearly and with relatively similar strength. The data are demeaned, detrended, tapered, and Fourier transformed. The resulting spectra are then smoothed and subsampled with a 50 point moving average to obtain the trends of spectral power (Fig. 1(a)). For the tests in this study, we limit the frequencies to 0.5–10 Hz unless otherwise indicated. The lower frequency limit is dictated by contamination of volcanic signals with low frequency (≤0.3 Hz) seismic noise from the ocean (McNamara and Buland, 2004; Bromirski et al., 2005), and by decreasing sensitivity of short period instruments below 1 Hz (Mark Product L-4 seismometers with a natural frequency of 1 Hz). The upper limit is chosen to include the frequency range where the (non-normalized) spectra differ the most (Figs. 1(b) and 2(a)). Different frequency bounds will be discussed in Section 5.3.

The three spectra in Fig. 2(a) capture the main differences between the three time periods: The background state has a relatively monotonic, steep decay of spectral power between 0.5 and 4 Hz, and flattens slightly at higher frequencies; Phase I and Phase II have increasing spectral power between 0.5 and 1 Hz (stronger for Phase II), and lower, approximately linear slopes from 1 Hz towards higher frequencies (compared to the background; Unglert and Jellinek, 2015). These features suggest differences in the underlying physical processes (Unglert and Jellinek, 2015). The stronger spectral power at low frequencies during Phase II compared to the background is, for example, explained by bubble cloud oscillations in a magma reservoir below Klauea's summit, whereas the relatively even, strong contributions at all frequencies