Automatic Reclassification of Volcano-Seismic Signals from Soufrière Hills Volcano, Montserrat, 1996-2008.

Glenn THOMPSON, Alexis FALCIN, Jean-Philippe METAXIAN, Francois BEAUDUCEL

Seismic activity during the eruption of Soufrière Hills volcano comprised various transient signals, which were classified visually by the Montserrat Volcano Observatory (MVO), considering waveforms recorded at several stations. For 217,290 transients detected on the MVO digital seismic network between 1996/10/21 and 2008/10/16, five main classes have been identified: rockfall (ROC: 58%), hybrid (HYB: 19%), long-period (LPE: 11%), lp-rockfall (LP-ROC 5.8%), and volcano-tectonic (VT: 3.1%). Temporal trends in the rate and energy release of these different transients (in addition to swarms and tremor) were key to short-term forecasting of eruptive activity. However, visual classification is highly subjective and non-repeatable, and the inconsistency of the catalog is a barrier to research.

In a pilot study, we automatically removed waveforms with dropouts, and manually verified transient classifications until we had approximately 100 transients of each class (total 522). We found ~21% of these transients were incorrectly classified at MVO. Our re-labelled dataset was then used as a starting point for supervised learning, using code from <http://github.com/malfante/AAA>. This code was used by Malfante et al. (2008) to classify 109,609 transients at Ubinas volcano with a 93.5% accuracy. They transformed each waveform into a set of 102 features: 34 features for each of three domains (time, spectral, cepstral). We added 6 frequency features of our own, including band ratios, peak frequency, median frequency, bandwidth, and frequency change. The resulting 108-point vectors of features were then used for modeling. The dataset is randomly divided 50 times into training and testing datasets, to produce a robust model. One model is produced per channel. We use the Random Forest Classifier algorithm from the scikit-learn library. For each waveform, a probability is computed for each class.

Initial results are promising. Separate models for 3 channels yield accuracies of 76-80%. If the LP-ROC class is omitted (following Langer et al, 2006), accuracy rises to 82-85%. If only VT and LP classes are considered, accuracy is 96-99%. We intend to expand our labelled dataset to 1000 events, build models for each channel, and reclassify the catalog of 217,290 transients by a weighted average of probabilities.

Seismic activity during the eruption of Soufrière Hills volcano comprised various transient signals, which were classified visually by the Montserrat Volcano Observatory (MVO), considering waveforms recorded at several stations. For 217,290 transients detected on the MVO digital seismic network between 1996/10/21 and 2008/10/16, five main classes have been identified: rockfall (ROC: 58%), hybrid (HYB: 19%), long-period (LPE: 11%), lp-rockfall (LP-ROC: 5.8%), and volcano-tectonic (VT: 3.1%). Temporal trends in the rate and energy release of these different transients (in addition to swarms and tremor) were key to short-term forecasting of eruptive activity. However, visual classification is highly subjective and non-repeatable, and the inconsistency of the catalog is a barrier to research.

In a pilot study, we automatically removed waveforms with dropouts, and manually verified transient classifications until we had approximately 100 transients of each class (total 522). We found ~21% of these transients were incorrectly classified at MVO. Our re-labelled dataset was then used as a starting point for supervised learning, using code from <http://github.com/malfante/AAA>. This code was used by Malfante et al. (2008) to classify 109,609 transients at Ubinas volcano with a 93.5% accuracy. They transformed each waveform into a set of 102 features: 34 features for each of three domains (time, spectral, cepstral). We added 6 frequency features of our own, including band ratios, peak frequency, median frequency, bandwidth, and frequency change. The resulting 108-point vectors of features were then used for modeling. The dataset is randomly divided 50 times into training and testing datasets, to produce a robust model. One model is produced per channel. We use the Random Forest Classifier algorithm from the scikit-learn library. For each waveform, a probability is computed for each class.

Initial results are promising. Separate models for 3 channels yield accuracies of 76-80%. If the LP-ROC class is omitted (following Langer et al, 2006), accuracy rises to 82-85%. If only VT and LP classes are considered, accuracy is 96-99%. We intend to expand our labelled dataset to 1000 events, add new features, build models for each channel, and reclassify the catalog of 217,290 transients by a weighted average of probabilities.