Geophysical Journal International

Geophys. J. Int. (2016) **207,** 29–46
Advance Access publication 2016 July 13
GJI Seismology

doi: 10.1093/gji/ggw258

Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression

S. Mostafa Mousavi,[1] Stephen P. Horton,[1] Charles A. Langston[1] and Borhan Samei[2]

[1] *Center for Earthquake Research and Information (CERI), University of Memphis, Memphis, TN* 38152, *USA. E-mail:* smousavi@memphis.edu
[2] *Institute for Intelligent Systems, University of Memphis, Memphis, TN* 38152, *USA*

Accepted 2016 July 11. Received 2016 July 10; in original form 2016 February 8

SUMMARY

We develop an automated strategy for discriminating deep microseismic events from shallow ones on the basis of the waveforms recorded on a limited number of surface receivers. Machine-learning techniques are employed to explore the relationship between event hypocentres and seismic features of the recorded signals in time, frequency and time–frequency domains. We applied the technique to 440 microearthquakes $-1.7 < M_w < 1.29$, induced by an underground cavern collapse in the Napoleonville Salt Dome in Bayou Corne, Louisiana. Forty different seismic attributes of whole seismograms including degree of polarization and spectral attributes were measured. A selected set of features was then used to train the system to discriminate between deep and shallow events based on the knowledge gained from existing patterns. The cross-validation test showed that events with depth shallower than 250 m can be discriminated from events with hypocentral depth between 1000 and 2000 m with 88 per cent and 90.7 per cent accuracy using logistic regression and artificial neural network models, respectively. Similar results were obtained using single station seismograms. The results show that the spectral features have the highest correlation to source depth. Spectral centroids and 2-D cross-correlations in the time–frequency domain are two new seismic features used in this study that showed to be promising measures for seismic event classification. The used machine-learning techniques have application for efficient automatic classification of low energy signals recorded at one or more seismic stations.

Key words: Neural networks, fuzzy logic; Wavelet transform; Earthquake source observations; Seismic monitoring and test-ban treaty verification; Volcano seismology.

1 INTRODUCTION

Microseismic tremors are low-amplitude events, attributed to reduction in effective stress. These typically negative-magnitude events provide useful information for understanding slow-slip earthquakes, incipient volcanic activity, fluid transport properties in geothermal reservoirs, potential ground control safety hazards in mining operations and induced seismicity from hydraulic fracturing for extraction of unconventional oil and gas resources.

Hypocentre information is the core product of microseismic monitoring. Locations of microearthquakes are inverted from seismic signals recorded by sensors either distributed at the surface or in monitoring boreholes. While surface monitoring usually suffers from low signal-to-noise ratio (SNR), the ability to place receivers in multiple azimuths and offsets allows for precise horizontal event location. On the other hand, downhole monitoring provides robust detection due to a higher SNR if an event is sufficiently close to the monitoring borehole. Another advantage of downhole monitoring over surface monitoring is the ability to estimate the depth based on arrival time moveout and wave polarization observed by the array. This allows a rough estimate of an event's depth without inversion and detailed velocity models.

The primary goal of this study is to develop a high-performance strategy for automatic clustering of microearthquakes recorded on a limited number of surface receivers based on their source depths.

Template matching is a common practice in observational seismology which has the advantage of combining the detection and classification. However, in some cases—such as the case study of this paper—waveforms may be highly incoherent due to complex wave propagation reducing the efficiency of the waveform cross-correlation for some sensitive classification tasks. Hence, in this study we try to understand signal characteristics of deep and shallow microearthquakes and cluster them based on these characteristics using machine-learning techniques.

In data mining using machine-learning systems, an algorithm can learn patterns from a sample data set and then determine the class of new data based on this previous knowledge. The advantage of machine-learning techniques is that they adopt data-driven learning schemes to find the solution to the problem. These techniques are capable of learning the input/output relationship directly from the data being modelled. Therefore, no prior knowledge of the statistical distribution of features is necessary to obtain a solution, even if these features are redundant or noisy. Machine-learning techniques, such as artificial neural networks (ANNs) have been extensively applied to seismic data primarily for the automatic classification of seismic events (e.g. Cercone & Martin 1994; Falsaperla *et al*. 1996; Scarpetta *et al*. 2005; Esposito *et al*. 2006; Langer *et al*. 2006; Hammer *et al*. 2012, 2013; Ait Laasri *et al*. 2013; Esposito *et al*. 2013; Vallejos & McKinnon 2013; Riggelsen & Ohrnberger 2014), the discrimination of artificial explosions from natural earthquakes (e.g. Dowla *et al*. 1990; Dowla 1995; Shimshoni & Intrator 1996; Amidan & Hagedom 1998; Fedorenko *et al*. 1999; Tarvainen 1999), discrimination of earthquakes from chemical explosions (e.g. Dysart & Pulli 1990; Benbrahim *et al*. 2005), discrimination of quarry blasts from microearthquakes (e.g. Musil & Plesinger 1996; Ursino *et al*. 2001; Kuyuk *et al*. 2011), discrimination of earthquakes from oil prospecting explosions (e.g. Abu-Elsoud *et al*. 2004), discrimination of earthquakes and underwater explosions (e.g. Del Pezzo *et al*. 2003), seismic event detection and automatic onset-time determination (e.g. Dai & MacBeth 1995, 1997; Wang & Teng 1995; Gravirov *et al*. 1996; Mousset *et al*. 1996; Tiira 1999; Zhao & Takano 1999; Glinsky *et al*. 2001; Gentili & Michelini 2006; Beyreuther *et al*. 2012; Kong *et al*. 2016), model-driven seismic interpretative processing (e.g. Maurer *et al*. 1992; Enescu 1996), automated seismic facies mapping (e.g. Baaske *et al*. 2007; Yuan *et al*. 2010), classification of seismic windows for full wave inversion (Diersen *et al*. 2011) and earthquake early warning (Böse *et al*. 2008; Zazzaro *et al*. 2012).

Most of the applications above belong to a class of problems referred to as pattern matching. In most of these cases, careful analysis of waveforms by an experienced seismologist can reveal the pattern and provide enough information for a robust detection/classification. However, the goal is to automate these time-consuming processes by training an algorithm to search for these patterns in large data sets. However, machine learning can be used for seismic event characterization problems beyond just pattern-matching (Perry & Baurngardt 1991). Some examples of these types of applications are: earthquake prediction (Katz & Aki 1992; Sharma & Arora 2005), imaging and interpretation of temporal patterns in seismic array data (Köhler *et al*. 2009, 2010), denoising of seismic signals (Essenreiter 1999; Djarfour *et al*. 2008), estimation of peak ground accelerations (García *et al*. 2006) and velocity model inversion (Moya & Irikura 2010).

Machine-learning techniques have been little utilized for characterization of seismic-source information such as event depth. Dowla (1995) used wavelet decomposition of regional events followed by a radial basis network and reported success in the depth estimation. However, details about the method and results of his study are not available. Perry & Baumgardt (1991) used ANNs to characterize event depth for regional earthquakes. They implemented a technique, called matched field processing (Bucker 1976; Baggeroer *et al*. 1988), to compute the spectral matrix of the Lg wave, and compare it to the Lg wave spectral matrix for master events at different depths. Using this method on synthetic data, they were able to distinguish deep regional events from shallow ones with 69.4 per cent precision.

In this study, we use machine-learning techniques such as correlation-based feature selection (CFS), ANNs, logistic regression (LR) and X-mean as research tools to explore the relationship between different seismic features of microearthquakes and their source depth and categorize events based on these features. Here, the proposed method is used for classifying pre-detected events. We successfully applied the method on real data for a sequence of microearthquakes observed at very close distance. Results of this study can have applications in the automatic classification of induced seismicity especially when signals are recorded by local networks with a limited number of sensors. In these cases, low energy signals recorded by single stations may be important manifestations of ongoing induced seismic activity, and their classification on the sole basis of the seismogram characteristics might be a way to discriminate between different types of induced microearthquakes.

2 DATA

In June and again in early July 2012, two widely felt events strongly shook the residents of a small community next to the Napoleonville salt dome in south Louisiana (Fig. 1). Around June 14, the USGS and CERI installed six stations in the area and observed ∼14 shallow microearthquakes per day. The rate increased to several hundred events/per day. On August 3rd, a sinkhole was found to have opened up near the earthquakes through August 2nd at which time they ceased. It later turned out to be due to an underground collapse of a cavern that fractured to surface and formed the sinkhole. Seismic monitoring of the sinkhole continues at this time although the surface network was reformed in a denser shape in January 2013 and a downhole string array of geophones was installed down to the depth of 915 m on top of the collapsed cavern in October 2013. Higher quality of recorded waveforms by the downhole array revealed that many deeper events are occurring in the salt body in addition to the shallow microearthquakes previously observed. Some of these deep events are observable on the surface network, but since they are usually noisy and recorded on just one or two stations, they cannot be located using conventional methods.

In this study, we develop a model for deep and shallow events based on the high-quality data recorded after October 2013. This model can be later applied on the data set recorded prior to the sinkhole formation for automatic discrimination of deep events (microearthquakes in salt body) from shallow ones (microseismic events in the cap rock). To develop the model, we used the catalogue of events located by downhole array and selected 4712 events located at depths between 1.0 and 2 km (deep), and 4498 events with hypocentre depths ranging between 40 and 400 m (shallow) for the initial processing. Out of these numbers, only a small set could be identified on the surface data and pass the criteria of having a minimum number of three stations (nine seismograms) and SNR of at least 2.0. More shallow events meet these criteria for the surface sensors while only the bigger deep events meet these criteria (Fig. 2). Therefore, 143 events, located at depths between 1.0 and 2 km (deep), and 297 events, with hypocentre depths ranging between 40 and 400 m (shallow) were ultimately selected for this study. Hypocentres for these events are shown in Fig. 3. MW magnitudes range from −1.7 to 1.29.

The surface network consists of eight broad-band three-component instruments at the surface and three short-period (2 Hz) geophones in shallow boreholes at distances up to 3 km from the collapsed cavern. Data were continuously recorded with a sampling rate of 200 Hz. For the selected events, all traces were highpass filtered above 2 Hz, and they were cut from 2 s before to 10 s after the event origin time.

A simple and common method for seismic event classification is to use cross-correlation between each event and template

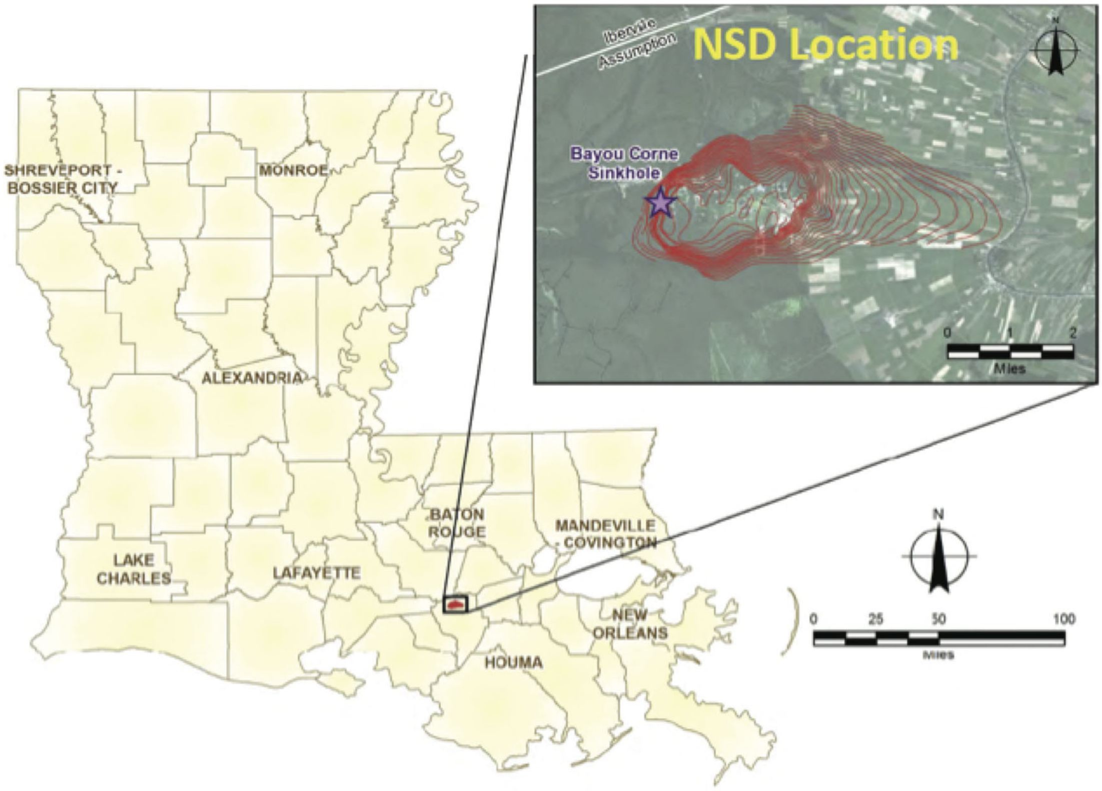

Figure 1. Location of Napoleonville Salt Dome (NSD) in Louisiana, USA. Contours show the top of the salt dome in feet below sea level.

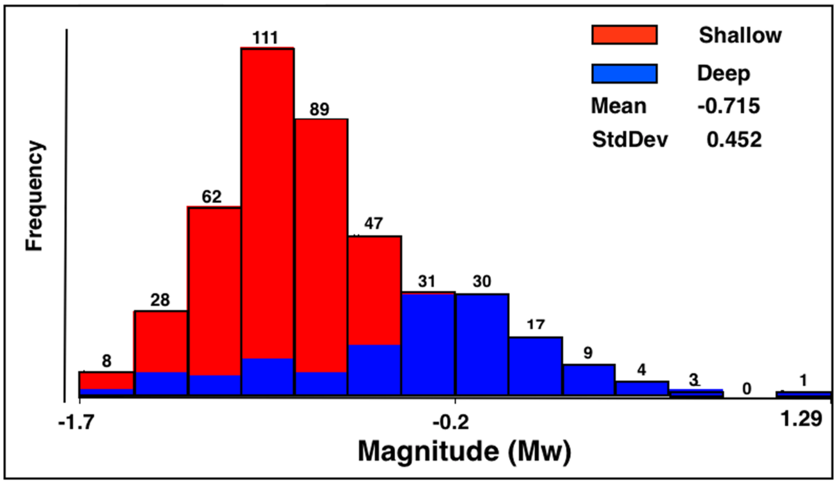

Figure 2. Magnitude (M_{w}) distribution of deep and shallow events.

events. To test the feasibility of this method we used the Hierarchical clustering method and clustered all the selected events based on their pair cross-correlation values (Fig. 4). In Hierarchical clustering, each event starts out as its own cluster with similar clusters being iteratively combined until only a single cluster remains. The cluster dendrogram for our selected events is shown in Fig. 4, with rectangles corresponding to cluster membership. Deep and shallow events spread out across different clusters. This is because waveforms associated with each type of event are not very coherent due to complex wave propagation at the region. Relatively low coherency of the waveforms can be seen in the plot of stacked waveforms for events with highest correlation (Fig. 5)

3 METHODOLOGY

The aim of pattern recognition is the classification of objects into a finite number of categories. In a pattern recognition system an object and a set of categories are given as input and the system decides to which category the object belongs. In general, it works in two stages. In the first stage, feature extraction (also known as the pre-processing or parameterization stage), a set of measures is extracted from the input object (seismogram). In the second stage, classification, the object is associated with one of the categories based on these features.

In common pattern recognition problems associated with seismic studies usually the parameters associated with each group (class)

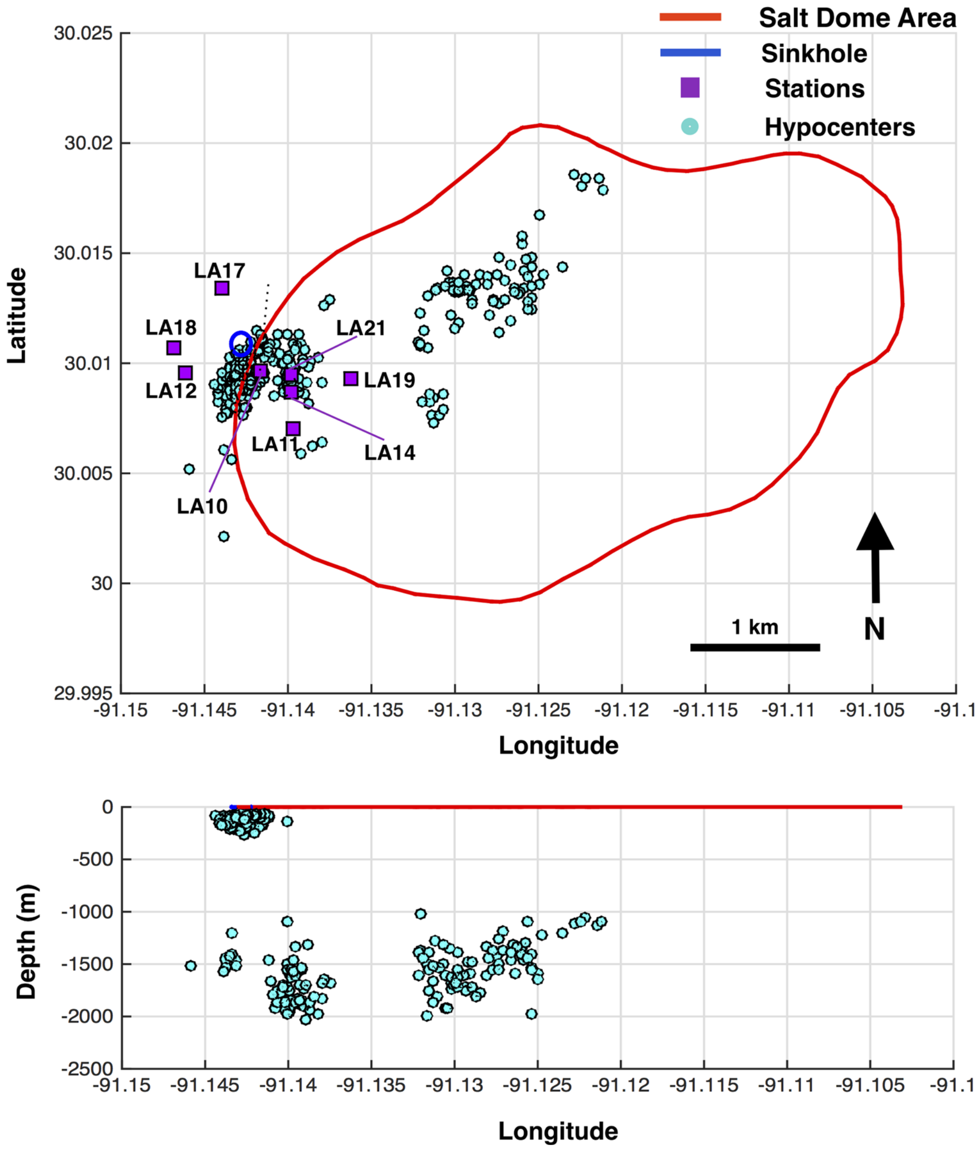

Figure 3. Location of microearthquake events and seismic stations used in this study.

of data are known beforehand. But for the classification purpose of this study, waveform features associated with event source depths are not completely known in advance. However, one advantage of using machine-learning algorithms is that the machine is able to discern patterns in the solution space that are difficult even for human experts.

3.1 Feature selection

Feature extraction is basically a transformation stage from data space into a feature space to extract robust information from the waveform in a compressed form. This step is critical for the success of the classification task.

Each datum, here three-component seismograms associated with one event, is represented by a feature vector which is used to train and test machine-learning models. It is important to select features that are informative and predictive of the individual datum properties. Furthermore, the size of the feature set and types of features (e.g. nominal, numeric, etc.) define the size of the learning problem. In other words, the larger the feature space, the more possible combinations of features need to be examined and learned by the machine-learning algorithm. In this study many different attributes of the seismic signals have been extracted from the data in time, frequency and time–frequency domains. We then test which features best represent the characteristics of the signal and limit the classification step to those features.

Fig. 6 shows a selected set of seismograms and their continuous wavelet transform (CWT) spectrograms. Shallow earthquakes are characterized by surface waves with the dominant energy concentrated around scale 4 and longer periods. Deep events radiate relatively higher frequency energy (Fig. 6) with no prominent surface waves.

Based partly on these observations, we implemented a broad range of parametrization methods including spectral analysis and polarization analysis. Table 1 gives an overview of the 40 features initially used in this study. Details of the mathematical definitions of features can be found in Appendix A. There are 12 frequency-based measurements, 11 time-based measurements and 17 time–frequency-based measurements. For every earthquake, the

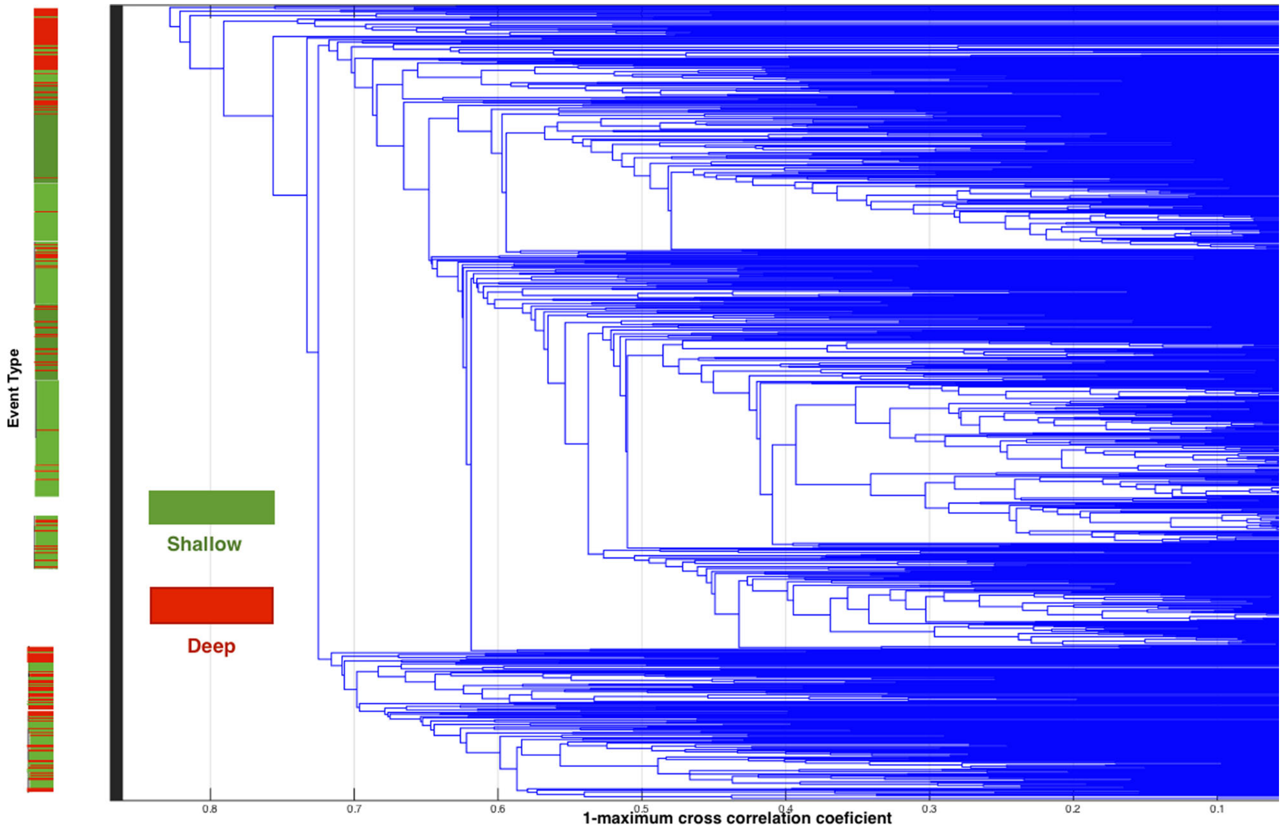

Figure 4. Hierarchical cluster tree generated based on intercluster correlation for vertical components of selected events. Showing different clusters of events based on their pair cross-correlations.

three-component waveforms are parametrized into a 40-element feature vector automatically by averaging measured attributes from all stations. Spectral features are measured on all three components and averaged for each station. In this study, we have incorporated the cross-correlation between waveforms and one shallow and one deep events (template matching) into the method as two features. In addition to the cross-correlating waveforms in the time domain, we have also measured cross-correlations in the time–frequency domain.

The 40-element feature vector provides a broad characterization of the waveforms. However, the exact relationship between each feature and event depth is not known yet. Moreover, some features may be irrelevant or redundant. It has been shown that whenever superfluous features are detected and removed using a feature selection technique before the classification step, the accuracy of the model will be improved and also time and effort will be saved (e.g. Karegowda *et al*. 2010; Samei *et al*. 2014). Removing redundant and/or irrelevant attributes can prevent the overfitting problem. Hence, we next assess the extracted features to find the best subset of features relative to the classification problem of our study. This was done by applying a CFS method (Hall 1998). This algorithm evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the classes while having low intercorrelation are preferred (Hall 1998).

The merit of a feature subset Z containing k features is defined by (Ghiselli 1964)

$$\mathrm{Merit}_Z = \frac{k\overline{r_{\mathrm{cf}}}}{\sqrt{k + k\,(k-1)\,\overline{r_{\mathrm{ff}}}}}, \qquad (1)$$

where, $\overline{r_{\mathrm{cf}}}$ is the average feature–class correlation, and $\overline{r_{\mathrm{ff}}}$ is the average feature–feature intercorrelation. Correlations, r, in this formula are standard Pearson's correlations. The numerator in this equation can be thought of as an indicator that how predictive a group of features are and the denominator shows how much redundant they are. The Genetic algorithm is used as a search method with CFS as the subset evaluation mechanism. The Genetic algorithm is a stochastic, general search method, capable of effectively exploring large search spaces, which is usually required in the case of attribute selection (Goldberg 1989).

In the Generic search algorithm (Goldberg 1989), CFS values (merits in eq. 1) are calculated for 20 different combinations of features (subsets) and this process is repeated 40 times and at the end the combination with highest merit is selected. We ran a cross-validation for this procedure to evaluate features. The average merit is calculated for each feature based on the number of times that the feature is selected over the cross-validation.

Based on initial feature evaluation one new feature was designed to combine the power of two other worthy features. Maximum power of frequency amplitude and the dominant frequency were combined to produce a new feature defined as

$$\mathrm{maxPF\_FA} = \frac{\text{Maximum power of frequency amplitude}}{\text{Dominant frequency}}. \qquad (2)$$

At the end, 30 features (a subset with highest merit) were selected for the classification step.

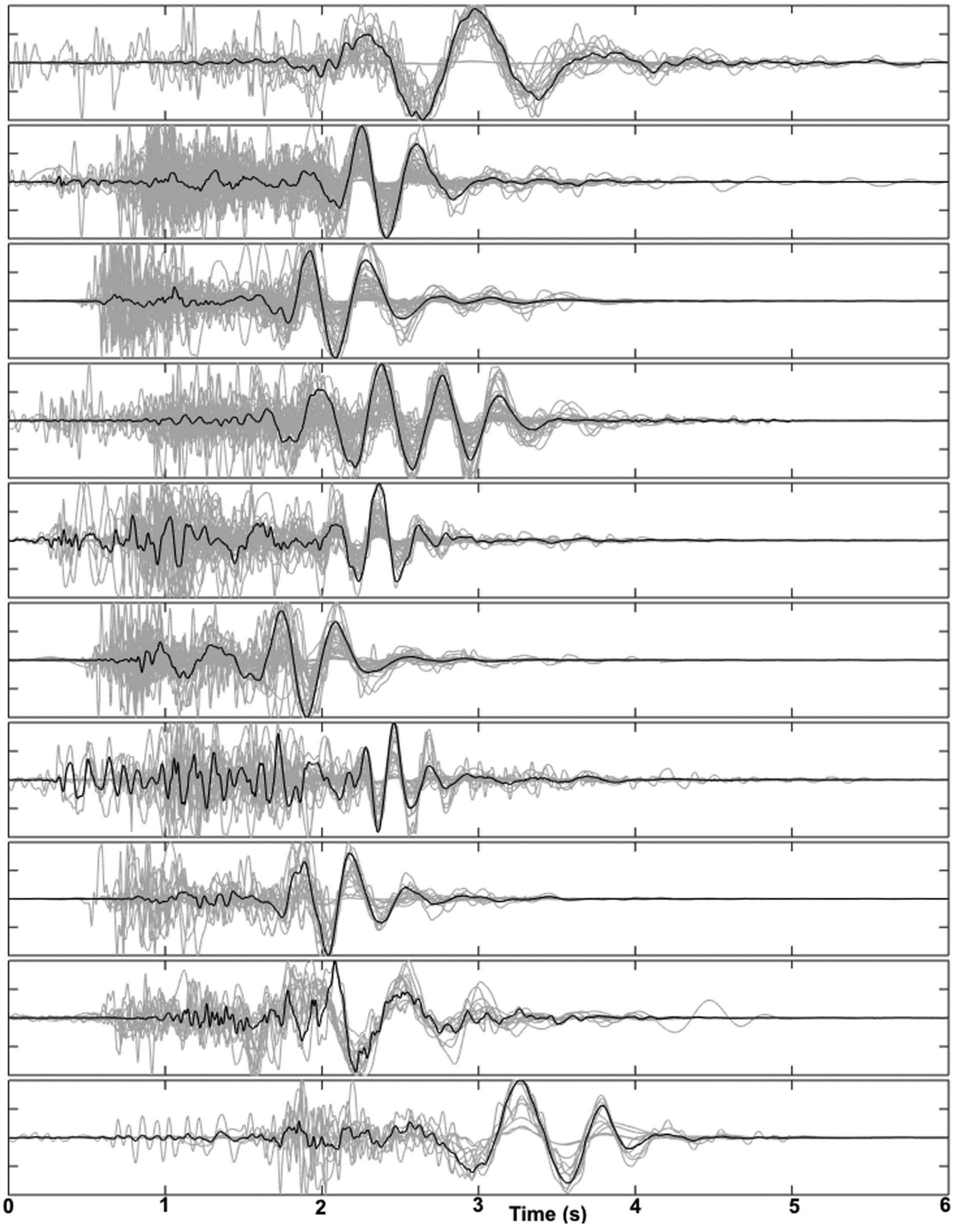

Figure 5. The stack of highly correlated events in each cluster showing a wide dispersion in their waveform shapes.

3.2 Classification

In machine-learning research, models that predict an outcome from a set of categories (e.g. deep and shallow) are often referred to as classifiers and the task of predicting an outcome from a datum (i.e. feature vector) is called classification. Several learning algorithms can be applied to deduce the classification models. Algorithms are designed to learn either from a set of labelled data (i.e. supervised) or unlabelled or partially labelled data (unsupervised and semi-supervised, respectively). We applied the supervised learning approach. Labelled training data with known class-membership (i.e. deep or shallow events) are used to introduce the patterns to be recognized by the algorithm. This training data set is then employed to predict class-memberships of the unseen data. The goal is to minimize the misclassification rate and the amount of false alarms.

LR and ANNs learning algorithms were applied to train a model that predicts shallow versus deep hypocentres based on selected features. Each of these algorithms has certain properties that take into account different characteristics of data. The reason for the comparative application of both techniques was to increase the reliability of discrimination.

3.2.1 Logistic regression

LR (Cox 1958) is a popular, powerful and easily understood statistical method to model and analyse multivariate problems (Press & Wilson 1978; Kleinbaum & Klein 2010). In LR the probability that a category is related to a set of explanatory variables (i.e. features) and the relationship between variables and a response variable is explored (Hosmer & Lemeshow 1989).

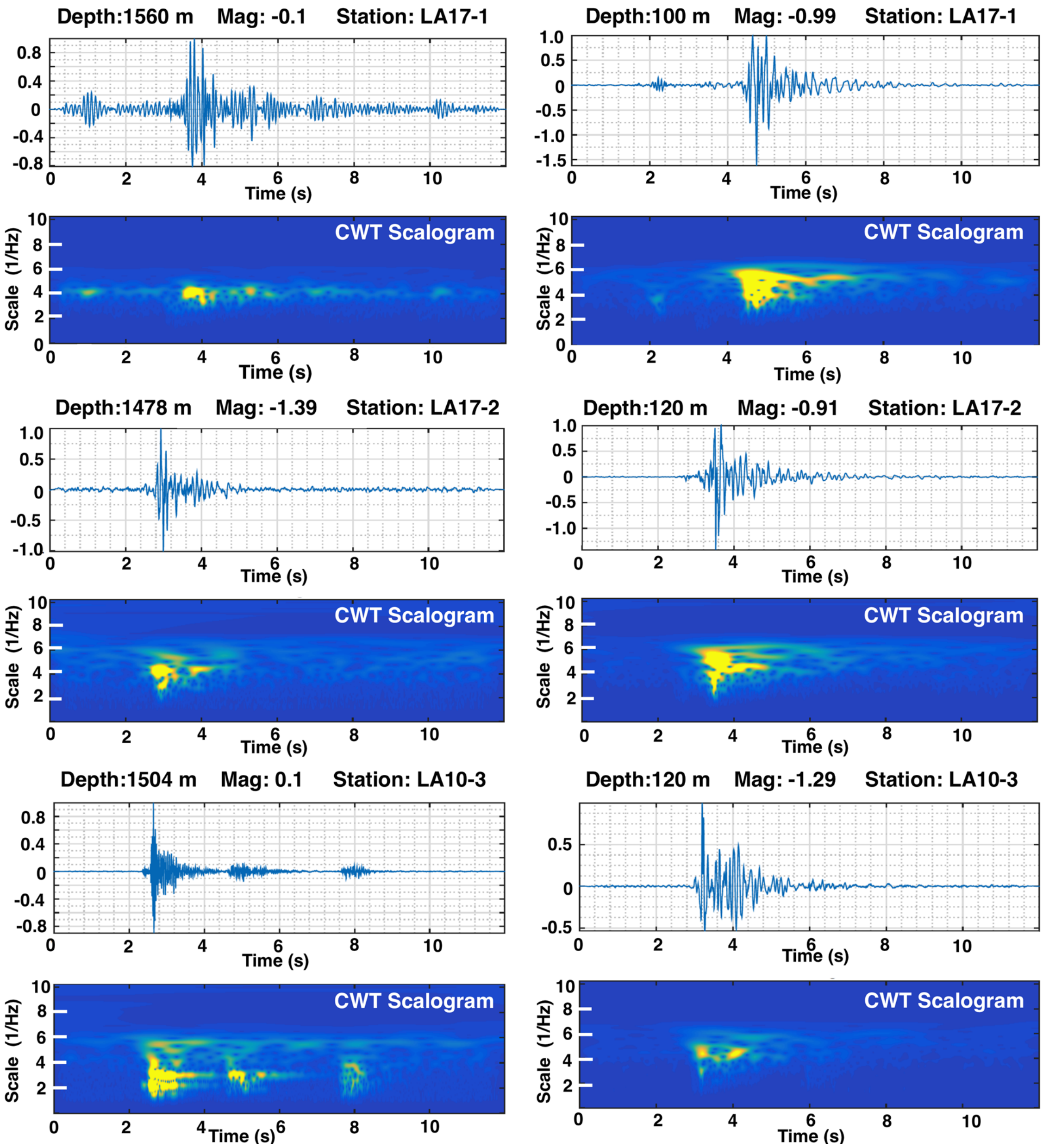

Figure 6. Seismograms for six events presented both in the time and time–frequency (continuous wavelet transform, CWT) domains. Left and right columns show deep and shallow events, respectively. Signals associated with deep events exhibit a limited frequency content mostly concentrated around scale 4 and spread out at the higher frequencies. Shallow events have a relatively broader spectrum with the concentration of the energy in lower frequencies (scales above 4).

Suppose that there are M classes with N measured seismic attributes. The probability for a particular class m with the exception of the last class is (LeCessie & van Houwelingen 1992)

$$P\left(m|Z\right) = \frac{e^Z}{1 + \sum_{m=1}^{M-1} e^Z}. \tag{3}$$

The last class has probability of

$$P\left(M\right) = 1 - \left(\sum_{m=1}^{M-1} P\left(m|Z\right)\right) = \frac{1}{1 + \sum_{m=1}^{M-1} e^Z}, \tag{4}$$

where $P(m|Z)$ is the categorical response of variables x_{i}, which represents the probability of a particular outcome, m. Z is a measure of the contribution of observed predictor variables x_{i}, in the category m, computed from a logistic model. The logistic model is a weighted summation of a set of explanatory variables, which is defined as

$$Z = \sum_{i=1}^{N} \beta_i x_i \; + \beta_0, \tag{5}$$

where β_0 is a constant (intercept), β_i are the predictor variable coefficients (the regression coefficients) which are estimated by maximum likelihood procedure.

In our case, the outcome variables are the event depth categories, deep or shallow, and $P(m|Z)$ is the probability of having a deep event based on the contribution of the observed features x_{i}. Coefficients are estimated using an iterative computation procedure. LR can be viewed as tossing a coin with ‘deep’ and ‘shallow’ sides,

Table 1. Seismic parameters derived from waveforms used in this study.

Feature	Description	Domain
Maximum amplitude (1 feature)	The maximum amplitude of the waveform, which characterizes the size of energy released.	Time
Spectral centroid (1 feature)	Indicates the 'centre of mass' of the spectrum.[a]	Frequency
Spectral attributes (4 features)	Includes the RMS of frequency amplitude, the maximum power, the dominant frequency and maximum power of frequency amplitude divided by dominant frequency.	Frequency
Energy density (1 feature)	The energy density of the signal.[a]	Time
Polarization parameters (5 features)	Include the measurement of the degree of rectilinearity, the apparent vertical incidence angle of rectilinear motion, azimuth range and dip angle, and dip times rectilinearity determined from eigenvalues of 3C-covariance matrix.[a]	Time
Waveform cross-correlation (2 features)	The estimation of the degree of similarity between waveforms and waveforms of one shallow and one deep template events.[a]	Time
Dominant period in CWT (1 feature)	The maximum scale continues wavelet representation of the data associated with the dominant energy in the signal.[a]	Time – Frequency
2-D wavelet cross-correlation (14 features)	2-D cross-correlation of CWT and DWT pictures of the signal and template events.[a]	Time – Frequency
The spectral coherency (2 features)	The measure of temporal (and spatial) variability in the spectral character of signals and template events.[a]	Time – Frequency
The spectral semblance (2 features)	Compares an event and a master event based on the correlation between their phase angles, as a function of frequency.[a]	Frequency
The envelop similarity (2 features)	A measure of the similarity between the signal shapes.[a]	Time
Spectral distance (4 features)	Measuring the spectral distance as a measure of correlation of event's spectra.[a]	Frequency
Spectral skewness (1 feature)	The skewness is used here to characterize the degree of symmetry or asymmetry of spectral contents of an event around its dominant frequency.[a]	Frequency

[a] The full description is provided in Appendix A.

where the probability of having a deep event is a function of $\beta_i x_i$. An event with seismic attribute measurements x_{i} could be classified as a deep event if $P(deep\,|\,Z) > P(shallow\,|\,Z)$ (Amidan & Hagedom 1998).

3.2.2 *Artificial neural networks*

ANNs are powerful mathematical models for organizing information relevant to the phenomena under study and representing complex relationships between inputs and outputs.

A neural network is made up of large numbers of simple, highly interconnected processing elements called neurons (nodes). Each node takes one or more inputs from other nodes and produces an output by applying an activation function over the weighted sum of its inputs. Nodes interact using weighted connections and are arranged in layers. A common type of ANN is a multilayer perceptron (MLP) which consists of an input layer, hidden layer(s), and an output layer (Duda *et al.* 2000). MLPs are capable of modelling complex nonlinear functions. Assuming an MLP with one hidden layer, the sigmoid activation function is given by

$$S\left(x\right) = \frac{1}{1 + e^{-x}}. \tag{6}$$

Therefore, given N features in the input layer, the output for the hidden node j, is calculated as

$$H_j = S\left(\sum_{i=1}^{N} w_{ji} x_i + \theta_{j0}\right), \tag{7}$$

where w_{ji} is the weight of the connection from the ith node in the input layer to the node j, the x_{i} are outputs of the input layer (features) and θ_{j} is a variable bias for the node j. In the output layer, the contribution Z_{N} from the N_{H} hidden nodes on category m is given by

$$Z_N = \sum_{j=1}^{N_H} w_{mj} H_j + \theta_{m0}, \tag{8}$$

where w_{mj} is the weight from hidden node j to the output category m. The probability for each category is calculated using eqs (3) and (4). The data are entered into the input layer. The neurons then process the input data, with the values resulting from each sigmoid progressing through the network towards the output layer. Once the values reach the output layer, the output computed by the network is compared with the desired output, and any error is employed as the basis for adjustment of all the connection weights, w using backward propagation. The weights associated with each connection are adjusted to strengthen connections that produce correct answers and weaken those that produce incorrect answers. This is done by iterative minimization of the errors using steepest descents in the backward propagation process. The direction of steepest descent is determined by the partial derivatives of the error with respect to the weights and bias in the network. This back propagation process is repeated until the network has learned the relationship between inputs and desired outputs. With no hidden layer, the ANN is equivalent to the LR.

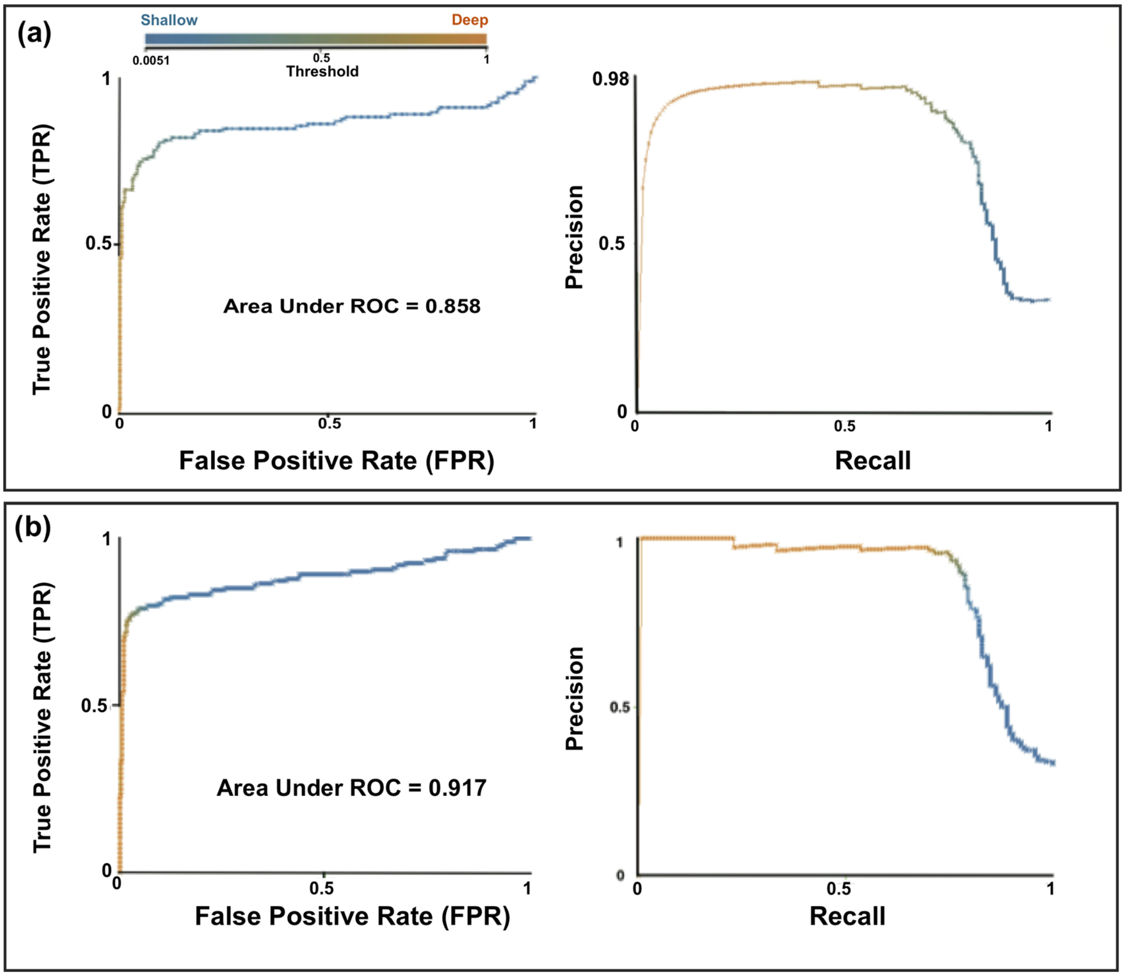

Figure 7. ROC curves (left) and precision-recall curves (right) for the LR (a) and ANN (b). The closer the curve follows the left-hand border and then the top border (bending in the curve towards the upper left corner of the chart), the more accurate the test.

Table 2. Performance of classifiers.

Model	Accuracy	AUC	Precision	Recall	RMS error	F-measure	Optimum threshold
LR	88.2%	0.86	0.88	0.88	0.32	0.88	0.53
ANN	90.7%	0.92	0.91	0.91	0.28	0.90	0.55

Table 3. Confusion matrix for LR.

Classified as	Deep	Shallow
Deep	103	40
Shallow	12	285

Table 4. Confusion matrix for ANN.

Classified as	Deep	Shallow
Deep	111	32
Shallow	9	288

3.3 Model evaluation

To assess the quality of the classification model, the whole training data set is divided into k unique subsets (folds) with roughly equal size. Then $k - 1$ folds are used for training the network, and the remaining fold is used for testing the learned model. This process is repeated so that each fold is used for testing exactly once. Therefore a *k-fold* cross-validation process builds k models and the results are averages over all k test sets.

In binary classification problems, like the subject of this study, an attempt is made to categorize the outcome of an event into one of two categories, either true (1) or false (0). This process can result in one of four possible outcomes that are defined as follows:

True Positive (TP): Evaluated and actual results are 1 (Valid Detection)

False Positive (FP): Evaluated result is 1, but actual result is 0 (False Alarm)

False Negative (FN): Evaluated result is 0, but actual result is 1 (Missed Detection)

True Negative (TN): Evaluated and actual results are 0 (Valid Non-detection)

This information is displayed in a two-by-two 'confusion' matrix describing the performance of the resulting model on the test data. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class. For identifying deep and shallow events with LR and ANN there is only one case which corresponds with the probability of having

Table 5. Ranking of selected features based on the CFS evaluation.

Feature	Short name	Average merit + standard deviation	Domain
Spectral centroid	spCen	0.445 ± 0.016	Frequency
Degree of rectiliniarity (polarization)	rect	0.358 ± 0.019	Time
RMS of frequency amplitude	rmsA	0.286 ± 0.021	Frequency
Maximum power of frequency amplitude	maxPFA	0.278 ± 0.015	Frequency
Dominant frequency	maxFA	0.27 ± 0.014	Frequency
2-D-CWT cross-correlation for deep template	ccnAb2D	0.262 ± 0.018	Time–frequency
Dip (polarization)	Dip	0.261 ± 0.015	Time
Waveform cross-correlation for deep template	ccD	0.254 ± 0.023	Time
The envelop similarity for deep template	envelopD	0.245 ± 0.05	Time
The envelop similarity for shallow template	envelopS	0.234 ± 0.02	Time
2-D-DWT cross-correlation for deep template	xp2S	0.214 ± 0.039	Time–frequency
Spectral coherency for deep template	semD	0.21 ± 0.017	Time–frequency
Dip angle times rectilinearity	DipRec	0.205 ± 0.071	Time
Dominant period in CWT	xotsu	0.167 ± 0.05	Time–frequency
Spectral semblance for deep template	semblanceD	0.132 ± 0.083	Frequency
Average power spectral skewness around dominant frequency	skwnss	0.13 ± 0.018	Frequency
Spectral distance for deep template	udD	0.127 ± 0.01	Frequency
2-D-CWT cross-correlation for shallow template	ccnAb2S	0.122 ± 0.095	Time–frequency
Spectral distance for shallow template	udS	0.111 ± 0.055	Frequency
Maximum coefficient of normalized 2-D-CWT cross-correlation along scale axis for shallow template	ypicD	0.101 ± 0.002	Time–frequency
Maximum coefficient of normalized 2-D-CWT cross-correlation along scale axis for deep template	ypicS	0.094 ± 0.012	Time–frequency
Maximum power of frequency amplitude/dominant frequency	maxPF-FA	0.092 ± 0.005	Frequency
Azimuth	azmth	0.0901 ± 0.061	Time
2-D-DWT cross-correlation for shallow template	xp2D	0.088 ± 0.086	Time–frequency
Mean coefficient of normalized 2-D-CWT cross-correlation between real parts for shallow template	ccnRel2S	0.081 ± 0.088	Time–frequency
Maximum amplitude	maxAmp	0.080 ± 0.081	Time
Waveform cross-correlation for shallow template	ccS	0.076 ± 0.08	Time
Apparent vertical incident angle	indAngle	0.052 ± 0.05	Time
Energy density	enrg	0.0501 ± 0.085	Time
Spectral semblance for shallow template	semblanceS	0.050 ± 0.09	Frequency

a deep event $P(deep \mid Z) > t$. After selecting a decision threshold value (t) several parameters for evaluating the performance of the model can be calculated such as true positive rate (TPR) and the false positive rate (FPR), which are both functions of the decision (detection) threshold, t:

$$\mathrm{TPR}\,(t) = \frac{\mathrm{TP}\,(t)}{\mathrm{TP}\,(t) + \mathrm{FN}\,(t)} = \text{Sensitivity} \tag{9}$$

$$\mathrm{FPR}\,(t) = \frac{\mathrm{FP}\,(t)}{\mathrm{TN}\,(t) + \mathrm{FP}\,(t)} = 1 - \text{Specificity}. \tag{10}$$

Table 6. Confusion matrix for waveform cross-correlation.

Classified as	Deep	Shallow
Deep	73	70
Shallow	70	227

Table 7. Confusion matrix for 2-D-CWT cross-correlation.

Classified as	Deep	Shallow
Deep	88	55
Shallow	59	238

Accuracy of the model is calculated based on the percentage of correctly classified instances. 'Precision' is defined as the fraction of predictions that are accurate. 'Recall' is defined as the fraction of instances that are accurately predicted. 'F-measure' is another measure of a test's accuracy computed as a weighted average of the precision and recall:

$$\text{Accuracy} = \frac{\mathrm{TP}\,(t) + \mathrm{TN}\,(t)}{\mathrm{TP}\,(t) + \mathrm{TN}\,(t) + \mathrm{FP}\,(t) + \mathrm{FN}\,(t)} \tag{11}$$

Table 8. Feature clusters based on X-mean clustering. The centroid of each cluster is shown by the asterisk (*).

Cluster 1	Cluster 2	Cluster 3	Cluster 4
spCen*	rmsA*	xp2D	rect*
maxFA	maxPFA	xotsu	Dip
envelopD	udD	ccAb2S*	ccS
semblanceD		skwnss	semD
DipRec			
envelopS			

$$\text{Precision} = \frac{\text{TP}\,(t)}{\text{TP}\,(t) + \text{FP}\,(t)} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}\,(t)}{\text{TP}\,(t) + \text{FN}\,(t)} \tag{13}$$

$$\text{F} - \text{measure} = \frac{2 \times \text{Precision} \ \times \ \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

The accuracy and precision of the algorithm represents a measure of its ability to consistently estimate the true outcome of an event. For a perfect classifier TPR = Accuracy = 1, and FPR = 0.

The plot of TPR as a function of FPR for a range of decision thresholds is known as a 'receive operating characteristic' (ROC) curve. The area under the ROC curve (AUC), is an absolute measure of the performance of the model (Fawcett 2006). The AUC takes values between 0.5 and 1.0. The predictive power of the classifier increases as the AUC approaches 1.0 and decreases as the AUC approaches 0.5. The objective of any classifier is to maximize the AUC.

4 RESULTS

Selecting an optimum architecture for the neural network is an important task. The topology of the network impacts the network performance, its generalization skills and its training duration.

In a classification problem, the number of neurons in the input layer is equal to the number of features (30 in our case) and the number of output nodes is determined by the number of classes (2 for deep and shallow classes). However, the appropriate number of hidden layers and the number of neurons in each hidden layer needs to be defined in a way that improves the classifiers performance.

We tested 10 topologies with one hidden layer and 8 topologies with two layers. The number of hidden neurons was selected considering the number of features, samples and classes. Most of these topologies had very similar performance. However, the topology with one hidden layer and five nodes had the best performance in terms of the average mean squared error over all folds, AUC, and accuracy.

We applied the above learning algorithms for classification and evaluated their performance using a 10-fold cross-validation. Our training data set consisted of 440 samples so that each fold had 44 samples. Thus the algorithm had 10 unique runs where each run started with a different set of random link weights, a slightly different example set and a unique test set. After calculation of regression coefficients for the model, the remaining fold was used for testing the learned model and making a prediction. The predictions were then compared to the actual outcomes, and the per cent of correct and erroneous predictions were calculated. This process was repeated 10 times until the model was tested on all the 10 folds. The overall performance was reported. We used WEKA (Hall *et al.* 2009), a machine-learning toolkit, to create and evaluate our models.

Next we examined the ROC curves (or threshold curves) for the LR and ANN (Fig. 7). ROC curves are cost-sensitive measures to evaluate the performance of classifiers and obtained by applying the classifier for various threshold levels. They illustrate the trade-off between the sensitivity and specificity. The closer the curve follows the left-hand border and then the top border of the ROC space (bending in the curve towards the upper left corner of the chart), the more accurate the test. Optimum detection thresholds are associated with points on the ROC curve that are closest to the point of perfect classification (0,1), which represents the highest TPR and the lowest FRP.

The area under the ROC curve is a measure of accuracy of the classification. Based on the AUC, the LR and ANN models present a very good discrimination of deep and shallow events within the seismic records for both the training and testing data sets. Another important visualization tool for evaluating a classifier's quality is the precision-recall curve. The goal is to observe whether your precision-recall curve is located towards the upper right corner of the chart. Final results of model evaluations are presented in Table 2.

As seen in Table 2, ANN achieves the best results while LR is relatively close in terms of performance. Both algorithms show a balance in precision and recall. Overall, these results suggest that ANN is a proper classifier in our case, however, LR also performed well. Confusion matrices for ANN and LR are presented in Tables 3 and 4, respectively. Both tables indicate relatively higher miss-classification of deep events. This was expected since our data set consists of surface recording and hence more constrains for shallower events.

We repeated this process using a different subset of the data consisting of 492 single-station three-component seismograms (218 associated with deep and 274 with shallow events). Similar results were obtained using single station data in that ANN achieved 87 per cent accuracy and LR achieved 85 per cent accuracy.

Background noise level is an important factor in performance of a seismic classifier (Riggelsen & Ohrnberger 2014). To check sensitivity of the classifier to the SNR and magnitude we divided the data set into four groups with $M_{\mathrm{w}} \geq -0.5$, $-0.5 > M_{\mathrm{w}} \geq -0.9$, $-0.9 > M_{\mathrm{w}} \geq -1.2$ and $-1.2 > M_{\mathrm{w}} \geq -1.7$, and repeated the classification using the ANN. Obtained accuracies are 94.54 per cent, 87.07 per cent, 88.09 per cent and 69.23 per cent, respectively. This suggests as events get smaller (as a result lower SNR), the accuracy of the classifier deteriorates.

5 DISCUSSION

We have demonstrated that it is possible to separate deep and shallow microearthquakes based on waveform features. However, it is not only important to be able to separate two data sets, but also to determine which variables are most relevant for achieving this separation.

To evaluate this, we refer to the CFS results. In Table 5, the selected features were ranked based on CFS results after 10-fold cross-validation. In general, we can conclude that frequency and polarization attributes, respectively, have the highest sensitivity for determining event source depth.

To test how a classifier based on waveform-cross-correlation or 2-D-wavelet cross-correlation perform in comparison to LR/ANNs, we classified events solely based on ccnAb2-D and

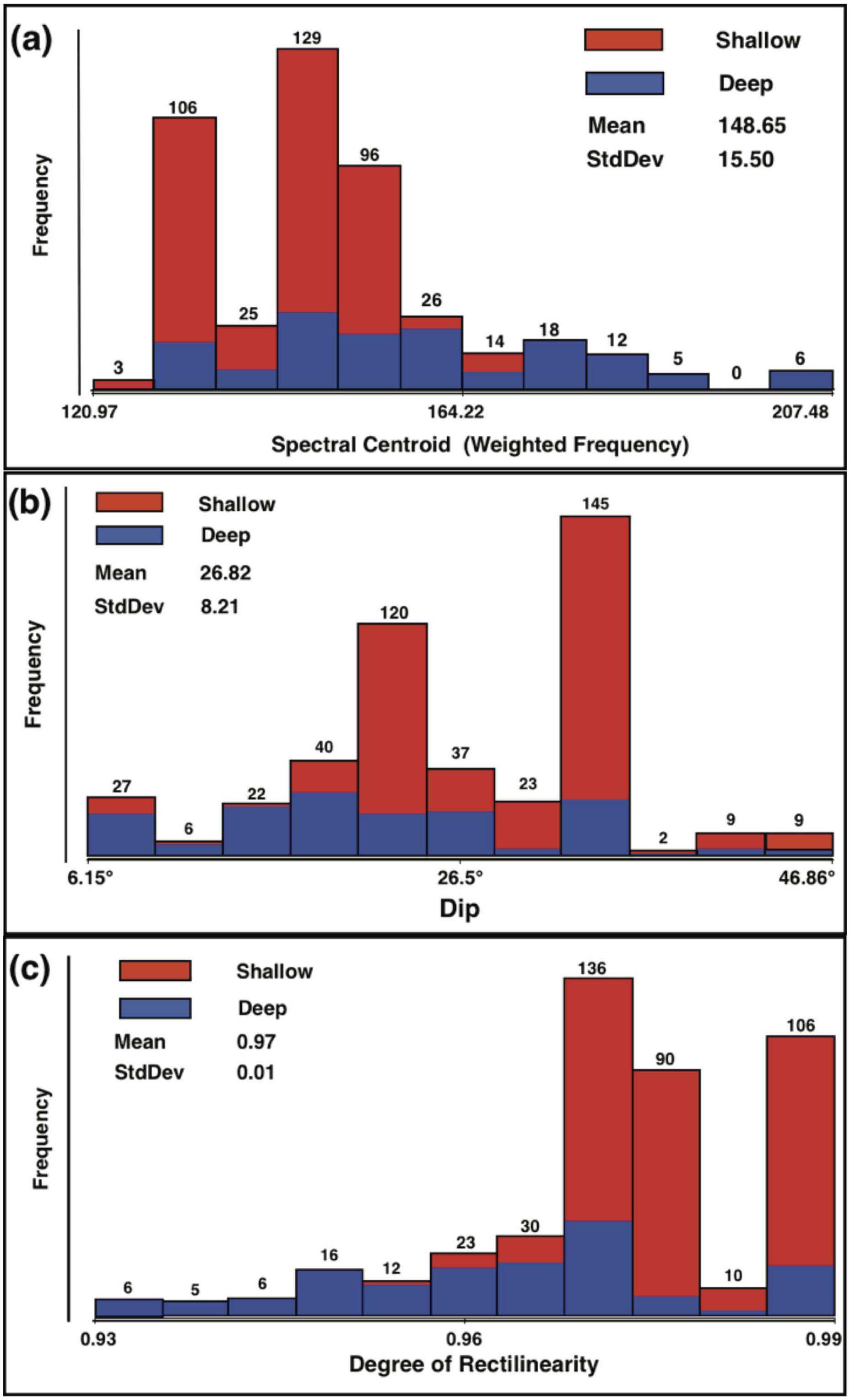

Figure 8. Relative distribution of spectral centroid, dip angle and degree of rectilinearity for deep (blue) and shallow (red) events.

ccD values. Classification accuracies for ccD (waveform) and ccnAb2-D (2-D-CWT) cross-correlations are 0.68 and 0.74, respectively. Confusion matrices are presented in Tables 6 and 7.

To take a closer look at the seismic features and their correlations, we applied the X-means method on the first 17 features with highest merits presented in Table 5. X-mean is an unsupervised clustering method (Duda & Hart 1973; Bishop 1995) similar to the K-means method. In K means clustering a parameter K is preset to the number of clusters and is based on the assumption that we know how many clusters exist. However, in the X-mean approach the number of clusters is unknown and the algorithm starts with assigning each instance to a new cluster and then merges the clusters based on their distance (in our case we used Euclidian distance) until it converges.

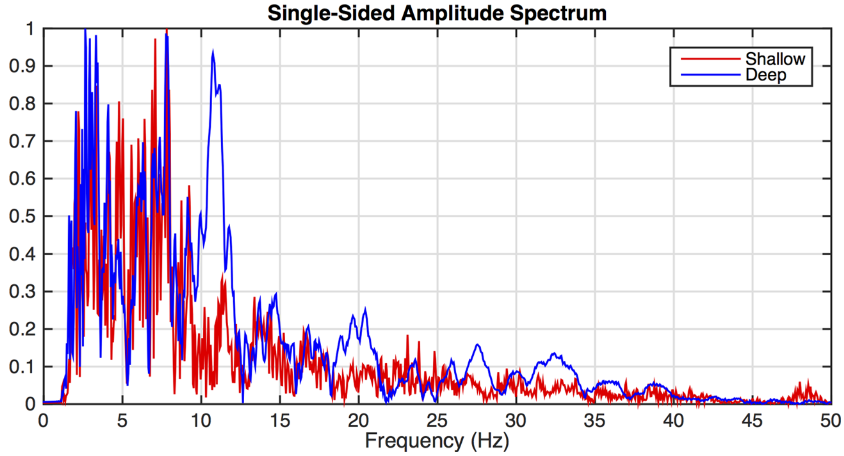

Figure 9. Mean normalized power spectra for the 440 events used in the classification.

X-mean discovered 4 clusters of features based on the correlation of each feature with other features.

Cluster 2 in Table 8 mainly consists of spectral attributes measured in the frequency domain. Cluster 3 mostly consists of time–frequency features with 2-D CWT cross-correlation as the centroid feature. Polarization features dominate in cluster 4 while features in the first cluster are more diverse in types. We have tested feasibility of the classification using just four centroid features of X-mean clusters, however, performance of the classifier is deteriorated. This can be due to high correlation between different features in each cluster with small portion of one class of outputs.

We can see that deep events have relatively higher centroid values from examination of the distribution of spectral centroid values (Fig. 8a). This is in agreement with the observation that deeper events have higher frequency content than shallow events seen in the raw data (Fig. 6).

In Fig. 9, we have plotted the mean normalized power spectra for all 440 events used in the classification. It can be seen that deeper events are slightly richer in the higher frequencies. Fagan *et al.* (2013) also reported a similar observation of higher spectral energy at higher frequencies for events farther from the receiver. We suspect that in our case the phenomenon is caused mainly by large surface waves contributing to the waveforms of shallow events and associated with the structure of low-velocity unconsolidated subsurface sediments. Essentially, low-frequency energy is amplified by surface wave excitation with the near-surface sediments also absorbing higher frequency body wave radiation. However the specific geometry of the salt and underground caverns also can cause complex wave propagation that may contribute to differing frequency content for these two groups of events.

Plots of the spectral centroid and maximum frequency amplitude values for 1033 single seismograms are shown in Fig. 10. Relatively higher frequency values for both measures occur for deep events compared to shallow. However, one can verify that the spectral centroid does a better job representing the higher frequency content of deep events. This is in agreement with the X-mean and CFS results.

Spectral centroid (also known as the barycentre or first order momentum of the magnitudes of the spectrum of frequencies) is a spectral measure usually used in audio signal processing. Although it is not a common measure in seismic practice our results indicate its potential to be used in seismic analyses. These results show that spectral characteristics of waveforms generally have tighter correlations to source parameters of small microearthquakes compared to other measures such as waveform correlation.

The second group of features that show high correlation to source depth are polarization features. We see that shallower events have relatively higher values of dip angle and rectilinearity (Figs 8b and c). This is because deep events in our study area are more horizontally dispersed and they can be located farther from receivers, while shallower events are clustered under the seismic network (Fig. 3). The high correlation between polarization attributes and event depth is due to the special geometry and hypocentre distribution of data used in this study so that it might not be a good idea to draw a general conclusion. However, it is interesting that spectral and polarization features can be ranked above the cross-correlation features, which are currently the most popular measures, used by seismologists for clustering purposes.

Another interesting result of this study is the slightly better performance of 2-D cross-correlation of data in the time–frequency domain compared to the waveform cross-correlation strictly in the time domain. Our results show that 2-D wavelet cross-correlation between time–frequency representations (TFRs) of template signals (one shallow and one deep) has better performance for classification (Tables 5–7). This result can be generalized for applications in other seismic studies. TFRs are powerful representations of the signal that have been shown to contain useful information for study of micro earthquakes induced by hydraulic fracturing (Pettitt *et al.* 2009; Das & Zoback 2011; Tary & van der Baan 2012; Tary *et al.* 2014). They localize information in time and frequency simultaneously and describe the energy distribution in the signal segment as a function of time and frequency. Hence they combine time-domain and frequency-domain analyses to yield a potentially more revealing picture of the temporal localization of a signal's spectral components. Hence, by 2-D cross-correlation of TFRs the similarity of the spectral content of two TFRs and their variation over time can be estimated. This performance can be improved for noisy traces by utilizing time–frequency denoising algorithms (e.g. Mousavi *et al.* 2016; Mousavi & Langston 2016a,b) prior to feature measurement. This can assemble the spectral content of the seismic signal more precisely. Thus, as we saw from higher correlation of spectral features and source depth, it seems reasonable that time–frequency comparisons give us a better measure than just waveform comparison alone.

The better performance of the 2-D cross-correlation in the CWT compared to the DWT, may be due to the higher resolution of the

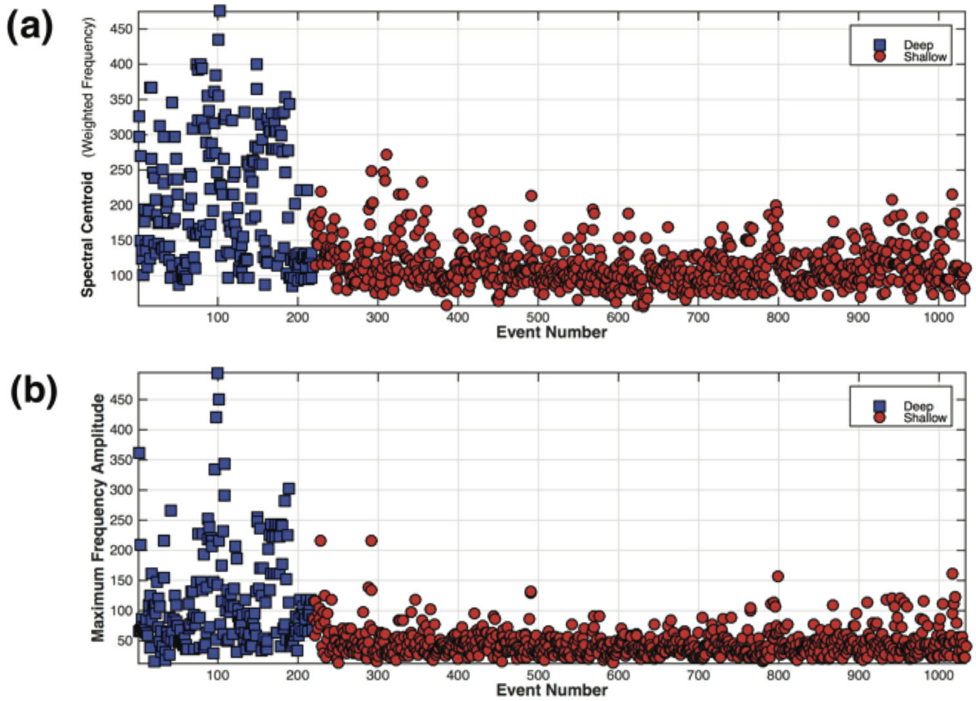

Figure 10. Spectral centroid (a) and the maximum power spectra (b) measured for 1033 single vertical components. The data have been ordered into deep (blue) and shallow (red) events.

CWT. It also can be seen that the Euclidean distance does not seem to be a good measure of correlation for event spectra. Other attributes defined for measuring spectral similarity, such as spectral skewness, and spectral semblance show lower correlation to event depth.

However, we should not forget that the data used in this study have been collected over a very local scale and that differences between event depths, referred to here as deep or shallow, are less than 800 m. Successful discrimination of event depth using waveform information with weak amplitude indicates the potential of the method and definition of seismic attributes proposed in this study for a variety of other seismic studies. One possible application of such methods can be in characterizing regional earthquakes and seismic hazard studies (e.g. Mousavi *et al*. 2011, 2014; McNamara *et al*. 2012). The relationship between these seismic attributes of the signal and other parameters of microseismic events is an interesting topic and can be studied further using other clustering and unsupervised techniques such as self-organizing maps (SOM; e.g. Musil & Plesinger 1996; Bashivan *et al*. 2008; Köhler *et al*. 2009, 2010; Esposito *et al*. 2013).

6 CONCLUSIONS

In this study the possibility of discrimination of an event's source depth was tested using the LR and ANN. The cross-validation test showed that these models were able to correctly predict the depth category of small events in a very local scale with 90.7 per cent of accuracy. ANN had a better performance compared to the LR. The applicability of the method for single-station data was tested with 87 per cent accuracy obtained. Seismic features based on the spectral measurements and polarization analysis had better correlation to the source depth. The spectral centroid had a better performance in representing the spectral contents of signals compared to other spectral parameters. Euclidian distance as a measure of spectral distance did not have good performance compared to other spectral attributes. 2-D cross-correlation of time–frequency representations showed an acceptable performance compared to the common waveform cross-correlation and seems to be a promising tool for analyses of signal similarities.

ACKNOWLEDGEMENTS

This study was partially supported by a contract from the Louisiana Department of Natural Resources. Arkansas and Bayou Corne seismic data used in this study are archived at IRIS/DMC. Authors would like to thank anonymous reviewer for careful review and constructive comments.

REFERENCES

Abu-Elsoud, M.A., Abou-Chadi, F.E.Z., Amin, A.E.M. & Mahana, M., 2004. Classification of seismic events in suez gulf area, Egypt, using artificial neural network, in *Proceedings of the International Conference on Electrical, Electronic and Computer Engineering,* IEEE, doi:10.1109/ICEEC.2004.1374460.

Ait Laasri, E.H., Akhouayri, E.S., Agliz, D. & Atmani, A., 2013. Seismic signal classification using multi-layer perceptron neural network, *Int. J. Comput. Appl.,* **79**(15), 35–43.

Amidan, B.G. & Hagedom, D.N., 1998. Logistic regression applied to seismic discrimination, *Technical Report (No. PNNL-RR-98-12031),* Pacific Northwest National Laboratory (PNNL), the U.S. Department of Energy, Washingtown (US), Richland.

Baaske, U.P., Mutti, M., Baioni, F., Bertozzi, G. & Naini, M.A., 2007. Using multi-attribute neural networks classification for seismic carbonate facies mapping: a workflow example from mid-cretaceous Persian gulf deposits, in: seismic geomorphology: applications to hydrocarbon exploration and production, *Geol. Soc. London,* **277,** 105–120.

Baggeroer, A.M., Kuperman, W.A. & Schmidt, H., 1988. Matched field processing: source localization in correlated noise as optimum parameter estimation, *J. acoust. Soc. Am.,* **83,** 571–587.

Bashivan, P., Fatehi, A. & Peymani, E., 2008. Multiple-model control of pH neutralization plant using the SOM neural networks, in *2008 Annual IEEE India Conference (Volume 1),* Kanpur, pp. 115–119.

Benbrahim, M., Daoudi, A., Benjelloun, K. & Ibenbrahim, A., 2005. Discrimination of seismic signals using artificial neural networks, *Int. J. Comput. Electr. Autom. Control Inform. Eng.,* **1**(4), 984–987.

Beyreuther, M., Hammer, C., Wassermann, J., Ohrnberger, M. & Megies, T., 2012. Constructing a hidden Markov model based earthquake detector: application to induced seismicity, *Geophys. J. Int.,* **189**(1), 602–610.

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition,* Clarendon Press.

Blackledge, J.M., 2003. *Digital Signal Processing,* Horwood Publishing, 757 pp.

Böse, M., Wenzel, F. & Erdik, M., 2008. PreSEIS: a neural network-based approach to earthquake early warning for finite faults, *Bull. seism. Soc. Am.,* **98**(1), 366–382.

Bucker, H.P., 1976. Use of calculated sound fields and matched-field detection to locate sound sources in shallow water, *J. acoust. Soc. Am.,* **59,** 368–373.

Cercone, J.M. & Martin, J.R., 1994. An application of neural networks to seismic signal discrimination, *Phillips Laboratory, Report no. 3,* PLTR-94-2178, Hanscon, AFB, Massachusetts.

Christensen, A.N., 2003. Semblance filtering of airborne potential field data, in *16th ASEG Conference and Exhibition,* Adelaide, doi:10.1071/ASEG2003ab024.

Cooper, G.R.J. & Cowan, D.R., 2008. Wavelet based semblance analysis, *Comput. Geosci.,* **34,** 95–102.

Cox, D.R., 1958. The regression analysis of binary sequences (with discussion), *J. R. Stat. Soc. B.,* **20,** 215–242.

Dai, H. & MacBeth, C., 1995. Automatic picking of seismic arrivals in local earthquake data using an artificial neural network, *Geophys. J. Int.,* **120**(3), 758–774.

Dai, H. & MacBeth, C., 1997. The application of back-propagation neural network to automatic picking seismic arrivals from single-component recordings, *J. geophys. Res.,* **102**(B7), 15 105–15 113.

Das, I. & Zoback, M.D., 2011. Long-period, long-duration seismic events during hydraulic fracture stimulation of a shale gas reservoir, *Leading Edge,* **30,** 778–786.

Daubechies, I. 1992. *Ten Lectures on Wavelets,* Vol. 61, SIAM, 357 pp.

Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M. & Scarpetta, S., 2003. Discrimination of earthquakes and underwater explosions using neural networks, *Bull. seism. Soc. Am.,* **93,** 215–223.

Diersen, S., Lee, E.J., Spears, D., Chenb, P. & Wang, L., 2011. Classification of seismic windows using artificial neural networks, in *Proceedings of the 11th International Conference on Computer Science,* Tsukuba, Japan.

Djarfour, N., Aïfa, T., Baddari, K., Mihoubi, A. & Ferahtia, J., 2008. Application of feedback connection artificial neural network to seismic data filtering, *C. R. Geosci*., **340,** 335–344.

Dowla, F.U., 1995. Neural networks in seismic discrimination, *Tech. Rep.,* Lawrence Livermore National Laboratory.

Dowla, F.U., Taylor, S.R. & Anderson, R.W., 1990. Seismic discrimination with artificial neural networks: preliminary results with regional spectral data, *Bull. seism. Soc. Am.,* **80**(5), 1346–1373.

Dreiseitl, S. & Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review, *J. Biomed. Inform*., **35,** 352–359.

Duda, R.O. & Hart, P.E., 1973. *Pattern Classification and Scene Analysis,* John Francisco.

Duda, R.O., Hart, P.E. & Stork, D.G., 2000. *Pattern Classification* Wiley Interscience.

Dysart, P.S. & Pulli, J.J., 1990. Regional seismic event classification at the NORESS array: seismological measurements and the use of trained neural networks, *Bull. seisem. Soc. Am.,* **80,** 1910–1933.

Enescu, N., 1996. Seismic data processing using nonlinear prediction and neural networks, in *IEEE NORSIG Symposium,* Espoo, Finland.

Esposito, A.M., Giudicepietro, F., Scarpetta, S., D'Auria, L., Marinaro, M. & Martini, M., 2006. Automatic discrimination among landslide, explosion-quake, and microtremor seismic signals at stromboli volcano using neural networks, *Bull. seism. Soc. Am.,* **96**(4A), 1230–1240.

Esposito, A.M., D'Auria, L., Giudicepietro, F., Caputo, T. & Martini, M., 2013. Neural analysis of seismic data: applications to the monitoring of Mt. Vesuvius, *Ann. Geophys.,* **56**(4), S0446, doi:10.4401/ag-6452.

Essenreiter, R., 1999. Identification and attenuation of multiple reflections with neural networks, *PhD thesis*, Universitat Karlsruhe.

Fagan, D., Van Wijk, K. & Rutledge, J., 2013. Clustering revisited: a spectral analysis of microseismic events, *Geophysics,* **78**(2), KS41–KS49.

Falsaperla, S., Graziani, S., Nunnari, G. & Spampinato, S., 1996. Automatic classification of volcanic earthquakes by using multi-layered neural networks, *Nat. Hazards,* **13,** 205–228.

Fawcett, T., 2006. An introduction to ROC analysis, *Pattern Recognit. Lett.,* **27,** 861–74.

Fedorenko, Y., Husebye, E.S. & Ruud, B.O., 1999. Explosion site recognition: neural net discriminator using single three-component stations, *Phys. Earth planet. Inter.,* **113,** 131–142.

García, S.R., Romo, M.P. & Mayoral, J.M., 2006. Estimation of peak ground accelerations for Mexican subduction zone earthquakes using neural networks, *Geofís. Int.,* **46**(1), 51–63.

Gentili, S. & Michelini, A., 2006. Automatic picking of P- and S-phases using a neural tree, *J. Seismol.,* **10**(1), 39–63.

Ghiselli, E.E., 1964. *Theory of Psychological Measurement,* McGraw-Hill.

Gibbons, S.J. & Ringdal, F., 2006. The detection of low magnitude seismic events using array-based waveform correlation, *Geophys. J. Int.,* **165,** 149–166.

Gitterman, Y., Pinsky, V. & Shapira, A., 1998. Spectral classification methods in monitoring small local events by the israel seismic network, *J. Seismol.,* **2,** 237–256.

Glinsky, M.E., Clark, G.A., Cheng, P.K.Z., Devi, K.R.S., Robinson, J.H. & Ford, G.E., 2001. Automatic event picking in prestack migrated gathers using a probabilistic neural network, *Geophysics,* **66**(5), 1488–1496.

Goldberg, D.E., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning,* Addison-Wesley.

Gravirov, V.V., Kislov, K.V. & Ovchinnikova, T.V., 1996. Neural network method for identification of earthquake phases in increased noise level conditions, in *Geophysical Research Abstracts,* EGU2010-2434-1, 12.

Hall, M.A., 1998. Correlation-based feature subset selection for machine learning, *PhD thesis*, Hamilton, New Zealand.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H., 2009. The WEKA data mining software: an update, *ACM SIGKDD Explorations Newsletter,* **11**(1), 10–18.

Hammer, C., Beyreuther, M. & Ohrnberger, M., 2012. A seismic event spotting system for volcano fast response systems, *Bull. seism. Soc. Am.,* **102**(3), 948–960.

Hammer, C., Ohrnberger, M. & Fäh, D., 2013. Classifying seismic waveforms from scratch: a case study in the alpine environment, *Geophys. J. Int.,* **192,** 425–439.

Haralick, R.M. & Shapiro, L.G., 1992. *Computer and Robot Vision,* **Vol. II**, pp. 316–317, Addison-Wesley.

Hosmer, D.W. & Lemeshow, S., 1989. *Applied Logistic Regression,* John Wiley and Sons, Inc.

Jurkevics, A., 1988. Polarization analysis of three-component array data, *Bull. seism. Soc. Am.,* **78,** 1725–1743.

Karegowda, A., Manjunath, A. & Jayaram, M., 2010. Comparative study of attribute selection using gain ratio and correlation based feature selection. *Int. J. Inform. Technol. Knowl. Manage.,* **2**(2), 271–277.

Katz, S. & Aki, K., 1992. Experiments with a neural net based earthquake alarm, *EOS, Trans. Am. geophys. Un.,* **73**(43), 366.

Kleinbaum, D.G. & Klein, M., 2010. *Logistic Regression: A Self Learning Text,* Springer.

Köhler, A., Ohrnberger, M. & Scherbaum, F., 2009. Unsupervised feature selection and general pattern discovery using self-organizing maps for gaining insights into the nature of seismic wavefields, *Comput. Geosci.,* **35**(9), 1757–1767.

Köhler, A., Ohrnberger, M. & Scherbaum, F., 2010. Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps, *Geophys. J. Int.,* **182,** 1619–1630.

Kong, Q., Allen, R.M., Schreier, L. & Kwon, Y.W., 2016. MyShake: a smartphone seismic network for earthquake early warning and beyond, *Sci. Adv.,* **2**(2), e1501055–e1501055.

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, *Nat. Hazards Earth Syst. Sci.,* **11,** 93–100.

Langer, H., Falsaperla, S., Powell, T. & Thompson, G., 2006. Automatic classification and a-posteriori analysis of seismic event identification at Soufriere hills volcano, montserrat, *J. Volcanol. Geotherm. Res.,* **153,** 1–10.

LeCessie, S. & van Houwelingen, J.C., 1992. Ridge estimators in logistic regression, *Appl. Stat.,* **41**(1), 191–201.

Ma, M., Zhao, G., Dong, L., Chen, G. & Zhang, C., 2015. *A Comparison of Mine Seismic Discriminators Based on Features of Source Parameters to Waveform Characteristics,* Vol. 2015, Hindawi Publishing Corporation, 10 pp.

Maurer, W.J., Dowla, F.U. & Jarpe, S.P., 1992. Seismic event interpretation using self organizing neural networks, *Proc. SPIE.,* **1709**, 950–958.

McNamara, D.E. *et al.*, 2012. Frequency dependent seismic attenuation within the Hispaniola Island region of the Caribbean sea, *Bull. seism. Soc. Am.,* **102,** 773–782.

Mousavi, S.M. & Langston, C.A., 2016a. Hybrid seismic denoising using higher-order statistics and improved wavelet block thresholding, *Bull. seism. Soc. Am.,* **106**(4), doi:10.1785/0120150345.

Mousavi, S.M. & Langston, C.A., 2016b. Adaptive noise estimation and suppression for improving microseismic event detection, *J. Appl. Geophys.,* **132,** 116–124.

Mousavi, S.M., Omidvar, B., Ghazban, F. & Feyzi, R., 2011. Quantitative risk analysis for earthquake-induced landslides—Emamzadeh Ali, Iran, *Eng. Geol.,* **122,** 191–203.

Mousavi, S.M., Cramer, C.H. & Langston, C.A., 2014. Average QLg, QSn, and observation of Lg blockage in the continental margin of Nova Scotia, *J. geophys. Res.,* **119,** 7722–7744.

Mousavi, S.M., Langston, C.A. & Horton, S.P., 2016. Automatic microseismic denoising and onset detection using the synchrosqueezed-continuous wavelet transform, *Geophysics,* **81,** V341–V355.

Mousset, E., Cansi, Y., Crusem, R. & Souchet, Y., 1996. A connectionist approach for automatic labeling of regional seismic phases using a single vertical component seismogram, *Geophys. Res. Lett.,* **23**(6), 681–684.

Moya, A. & Irikura, K., 2010. Inversion of a velocity model using artificial neural networks, *Comput. Geosci.,* **36,** 1474–1483

Musil, M. & Plesinger, A., 1996. Discrimination between local microearthquakes and quarry blasts by multi-layer perceptrons and Kohonen maps, *Bull. seism. Soc. Am.,* **86,** 1077–1090.

Perry, J.L. & Baurngardt, D.R., 1991. Lg depth estimation and ripple fire characterization using artificial neural networks, in *Proceedings of 7th IEEE Conference on Artificial Intelligence Applications,* Miami Beach, FL, USA, 24–28 February 1991, pp. 231–234.

Pettitt, W., Reyes-montes, J., Hemmings, B., Hughes, E. & Young, R.P., 2009. Using continuous microseismic records for hydrofracture diagnostics and mechanics, *SEG Expanded Abstracts,* **28,** 1542–1546.

Press, S.J. & Wilson, S., 1978. Choosing between logistic regression and discriminant analysis, *J. Am. Stat. Assoc.,* **73,** 699–705.

Riggelsen, C. & Ohrnberger, M., 2014. A machine learning approach for improving the detection capabilities at ctbto/ims 3c seismic stations, *Pure appl. Geophys.,* **171,** 395–411.

Samei, B., Li, H., Keshtkar, F., Rus, V. & Graesser, A.C., 2014. Context-based speech act classification in intelligent tutoring systems, in *Proceeding of the Intelligent Tutoring Systems,* Honolulu, Hawaii, USA, pp. 236–241.

Scarpetta, S., Giudicepietro, F., Ezin, E.C., Petrocino, S., Pezzo, E.D., Martini, M. & Marinaro, M., 2005. Automatic classification of seismic signals at mt.vesuvius volcano, Italy, using neural networks, *Bull. seism. Soc. Am.,* **95**(1), 185–196.

Sharma, M.L. & Arora, M.K., 2005. Prediction of seismicity cycles in the himalayas using artificial neural networks, *Acta Geophys. Pol.,* **53**(3), 299–309.

Shimshoni, Y. & Intrator, N., 1996. Classification of seismic signals by integrating ensembles of neural networks, *IEEE Trans. Signal Process.,* **46,** 1194–1201.

Tarvainen, M., 1999. Recognizing explosion sites with a self-organizing network for unsupervised learning, *Phys. Earth planet. Inter.,* **113,** 143–154.

Tary, J.B. & van der Baan, M., 2012. Potential use of resonance frequencies in microseismic interpretation, *Leading Edge,* **31,** 1338–1346.

Tary, J.B., Herrera, R.H. & van der Baan, M., 2014. Time-varying autoregressive model for spectral analysis of microseismic experiments and long-period volcanic events, *Geophys. J. Int.,* **196**(1), 600–611.

Tiira, T., 1999. Detecting teleseismic events using artificial neural networks, *Comput. Geosci.,* **25,** 929–939.

Torrence, C. & Compo, G.P., 1998. A practical guide to wavelet analysis, *Bull. Am. Meteorol. Soc.,* **79,** 61–78.

Tzanetakis, G., Essl, G. & Cook, P., 2001. Automatic musical genre classification of audio signals, in *Proc. Int. Symposium on Music Information Retrieval (ISMIR),* Bloomington, Indiana.

Ursino, A., Langer, H., Scarfi, L., Di Grazia, G. & Gresta, S., 2001. Discrimination of quarry blasts from tectonic microearthquakes in the Hyblean Plateau (Southeastern Sicily), *Ann. Geofis.,* **44,** 703–722.

Vallejos, V. & McKinnon, S.D., 2013. Logistic regression and neural network classification of seismic records, *Int. J. Rock Mech. Min. Sci.,* **62,** 86–95.

Vidale, J.E., 1986. Complex polarization analysis of particle motion, *Bull. seism. Soc. Am.,* **76,** 1393–1405.

von Frese, R.R.B., Jones, M.B., Kim, J.W. & Kim, J.H., 1997. Analysis of anomaly correlations, *Geophysics,* **62**(1), 342–351.

Wang, J. & Teng, T.L., 1995. Artificial neural network-based seismic detector, *Bull. seism. Soc. Am.,* **85**(1), 308–319.

Yuan, G., Peimin, Z., Hui, R., Yang, H. & Fanping, Z., 2010. Seismic facies classification using bayesian networks, in *AAPG Search and Discovery Article #90172 © CSPG/CSEG/CWLS GeoConvention 2010,* Calgary, Alberta, Canada, May 10–14, 2010.

Zazzaro, G., Pisano, F.M. & Romano, G., 2012. Bayesian networks for earthquake magnitude classification in a early warning system, *World Acad. Sci. Eng. Technol.,* **6,** 4–27.

Zhao, Y. & Takano, K., 1999. An artificial neural network-based seismic detector, *Bull. seism. Soc. Am.,* **77,** 670–680.

APPENDIX A: DISCUSSION OF FEATURE EXTRACTION

A1 Spectral centroid

The spectral centroid indicates the 'centre of mass' of the spectrum and is measured as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights, divided by the sum of magnitudes (Tzanetakis *et al.* 2001):

$$\text{Spectral Centroid} = \frac{\sum_1^M f_i.m_i}{\sum_1^M m_i}, \tag{A1}$$

where m_i is the magnitude of bin number i, f_i represents the centre frequency of that bin, and M is the number of bins. We measured mean spectral centroid values.

A2 Energy density

The spectral energy density of signal $x(t)$ is

$$E = \frac{1}{N}\int_{-\infty}^{\infty} |x(t)|^2 \mathrm{d}t, \tag{A2}$$

where N is the number of samples.

A3 Polarization analysis

The polarization of a wave depends on wave type, wave propagation and sensor orientation. Here we use the covariance matrix of three-component seismograms for the polarization analysis (Vidale 1986). A covariance matrix, $\sigma,$ is calculated using an M-sample sliding window over three orthogonal ground-motion recordings corresponding to the east, north and vertical components (respectively noted X, Y and Z):

$$\sigma = \begin{pmatrix} Cov\,(X,X) & Cov\,(X,Y) & Cov\,(X,Z) \\ Cov\,(Y,X) & Cov\,(Y,Y) & Cov\,(Y,Z) \\ Cov\,(Z,X) & Cov\,(Z,Y) & Cov\,(Z,Z) \end{pmatrix}. \qquad \text{(A3)}$$

The covariance between any two components X and Y is defined as

$$\mathrm{Cov}(X,Y) = \frac{1}{M}\sum_{i=1}^{M} x_i\ y_i. \qquad \text{(A4)}$$

We compute the covariance matrix of the pre-processed seismograms for a 0.25 s (50 sample) sliding window. Then corresponding eigenvalues ($\lambda 3 \geq \lambda 2 \geq \lambda 1$) and eigenvector matrix $\mathrm{u} = (\mathrm{u}_1, \mathrm{u}_2, \mathrm{u}_3)$ of σ are used to estimate the degree of rectilinearity (i.e. a measure of the strength of polarization in the signal (Jurkevics 1988)), for each window as

$$\mathrm{Rec} = 1 - \left(\frac{\lambda 2 + \lambda 1}{\lambda 3}\right). \qquad \text{(A5)}$$

The degree of linear polarization, Rec is approximately equal to 1.0 when there is only one nonzero eigenvalue, as for pure body waves. Hence, we use the window with the closet value to 1.0 to extract the characteristics of the ground motion based on attributes computed from the principal axes ($\lambda_\mathrm{i}\mathrm{U}_\mathrm{i}$).

The azimuth of propagation can be estimated from the horizontal orientation of rectilinear motion, given by the eigenvector $\mathbf{u}_1$ corresponding to the largest eigenvalue (Jurkevics 1988):

$$P\_\mathrm{azimuth} = \tan^{-1}\left(\frac{u_{13}\mathrm{sign}\,(u_{33})}{u_{23}\mathrm{sign}\,(u_{33})}\right), \qquad \text{(A6)}$$

where $\mathbf{u}_{j3}, j = 1\ldots 3$ are the three direction cosines of eigenvector $\mathbf{u}_3$. The sign function is introduced to resolve 180° ambiguities by taking the positive vertical component of $\mathbf{u}_3$.

The apparent vertical incidence angle of rectilinear motion, ϕ, is obtained from the vertical direction cosine of $\mathbf{u}_3$

$$\phi = \cos^{-1}|U_{33}|. \qquad \text{(A7)}$$

The dip of the direction of maximum polarization is defined as

$$\mathrm{Dip} = \tan^{-1}\left(\frac{U_{33}}{\sqrt{{U_{32}}^2 + {U_{31}}^2}}\right). \qquad \text{(A8)}$$

Possible values for dip angle range from −90° to +90°, where 0° dip represents a vector which points horizontally in the direction back to the epicentre.

A4 Waveform cross-correlation

The waveform cross-correlation is a standard method of estimating the degree to which two time series are similar. Consider two series $x(i)$ and $y(i)$ where $i = 0, 1, 2\ldots N{-}1$. The cross-correlation C at delay d is defined as

$$C\,(d) = \frac{\sum_i \left[(x\,(i) - mx) * (y\,(i-d) - my)\right]}{\sqrt{\sum_i (x\,(i) - mx)^2 \sum_i (y\,(i-d) - my)^2}}, \qquad \text{(A9)}$$

where mx and my are the means of the corresponding series (e.g. Gibbons & Ringdal 2006).

A5 Dominant period in CWT

For a given mother wavelet ψ, the CWT of $x(t)$ at scale a and time shift τ is given by (Daubechies 1992)

$$\mathrm{CWT}_x\,(a,\tau) = \int x\,(t) a^{-\frac{1}{2}} \psi^* \left(\frac{t-\tau}{a}\right) \mathrm{d}t, \qquad \text{(A10)}$$

where the * is the complex conjugate, and CWT_x is the coefficient representing finite energy of the signal $x(t)$ in a concentrated time–frequency picture. The maximum scale corresponds to the minimum frequency.

A6 2-D Wavelet cross-correlation

The 2-D-CWT cross-correlation of an M-by-N matrix X and a P-by-Q matrix Y is a matrix C of size $M + P - 1$ by $N + Q - 1$ given by (Haralick 1992)

$$C\,(k,l) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} X\,(m,n)\,Y^*\,(m-k, n-l) \qquad \text{(A11)}$$

where the * denotes complex conjugation. And

$$\begin{cases} -(P-1) \leq k \leq M-1 \\ -(Q-1) \leq l \leq N-1 \end{cases}. \qquad \text{(A12)}$$

The time–frequency representations of waveforms (CWT coefficients) are used for the 2-D cross-correlation between wavelet representations of each event and master events (one shallow and one deep template).

A7 The spectral coherency

The spectral *(*wavelet) coherency is a measure of the similarity of wavelet representations of two signals and provides the ability to account for temporal (or spatial) variability in spectral character. The cross-wavelet transform (Torrence & Compo 1998), is defined as

$$\mathrm{CWT}_{1,2} = \mathrm{CWT}_1 \times \mathrm{CWT}_2^* \qquad \text{(A13)}$$

and is a complex quantity having an amplitude (the cross-wavelet power) given by

$$A = \left|\mathrm{CWT}_{1,2}\right| \qquad \text{(A14)}$$

and local phase θ,

$$\theta = \tan^{-1}\left(\mathrm{Img}\,(\mathrm{CWT}_{1,2})\,/\mathrm{Rel}\,(\mathrm{CWT}_{1,2})\right). \qquad \text{(A15)}$$

Wavelet coherency is a measure of phase correlation between two wavelet representations and can be calculated by (Cooper & Cowan 2008)

$$\mathrm{Scwt} = \cos^n\,(\theta) \qquad \text{(A16)}$$

where, n is an odd integer greater than zero. Scwt values range from 1 (inversely correlated) through zero (uncorrelated) to +1 (correlated). Combining the phase information of Scwt with the amplitude information of A, another measure of coherency of wavelet representations can be defined as:

$$\mathrm{Dcwt} = \cos^{n}(\theta)\left|\mathrm{CWT}_1 \times \mathrm{CWT}_2^{*}\right| \tag{A17}$$

A8 The spectral semblance

Semblance filtering compares two data sets based on correlations between their phase angles, as a function of frequency. When the Fourier transforms of two time-series x_1 and x_2 are calculated, the difference in their phase angles at each frequency can be computed simply from (von Frese *et al.* 1997; Christensen 2003)

$$S = \cos\theta(f) = \frac{R_1(f)\,R_2(f) + I_1(f)\,I_2(f)}{\sqrt{R_1^2(f)\,I_1^2(f)} \times \sqrt{R_2^2(f)\,I_2^2(f)}} \tag{A18}$$

where, $R_\mathrm{i}(f)$ and $I_\mathrm{i}(f)$ are the real and imaginary components of the Fourier transform of x_i, expressed as a function of frequency f. The semblance S can take on values from 1 to +1. A value of +1 implies perfect phase correlation, 0 implies no correlation, and −1 implies perfect anticorrelation.

A9 The envelope similarity

Envelope similarity is a measure of the similarity between the signal shape of each event e and the reference template (master event) e_m. The envelope similarity Es is measured using the Manhattan distance:

$$E_s = \frac{\sum_{i=1}^{N}\left|e(i) - e_m(i)\right|}{\sum_{i=1}^{N} e(i)} \tag{A19}$$

where,

$$e(i) = \sqrt{X(i)^2 + H[X(i)]^2}, \tag{A20}$$

X is the vertical component seismogram and H indicates the Hilbert Transform.

A10 The spectral distance

We used the squared Euclidean distance between the normalized spectral powers to measure the similarities of events in the frequency domain (Fagan *et al.* 2013). All waveforms start 2 s before the origin time and have 12 s duration and length of 2000 samples. The autocovariance for lag k of event $\{x(t)\colon t = 1, 2, \ldots, n\}$ with zero mean is defined as

$$\mathrm{Ac}(k) = \frac{1}{n}\sum_{t=k+1}^{n} x_t x_{t-k}, \quad k = 0, 1, \ldots, n-1. \tag{A21}$$

The spectrum is the discrete Fourier transform of Ac and has the form of

$$F(\omega) = \mathrm{Ac}(0) + 2\sum_{k=1}^{n-1}\mathrm{Ac}(k)\cos(k\omega), \tag{A22}$$

where Ac(0) is the variance of x and $\omega = 2\pi j/n$ for positive integer $j < n/2$. The first 500 Fourier frequencies are then used for the correlation. The squared spectral distance for event e and master event e_m is

$$d^2(e, e_m) = \left[F_e(\omega) - F_{e_m}(\omega)\right]^T\left[F_e(\omega) - F_{e_m}(\omega)\right]. \tag{A23}$$

A11 The spectral skewness

Skewness is used here to characterize the degree of symmetry or asymmetry of the amplitude spectrum of an event signal around its dominant frequency. Sk for a roughly symmetrical function is near zero and is given by

$$\mathrm{Sk} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(F(\omega) - \bar{\omega}\right)^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F(\omega) - \bar{\omega}\right)^2}\right)^3}, \tag{A24}$$

where N is the Nyquest frequency and $\bar{\omega}$ is mean frequency in the window around the dominant frequency of the signal.