Geophysical Journal International

Geophys. J. Int. (2010) **182,** 1619–1630 doi: 10.1111/j.1365-246X.2010.04709.x

Unsupervised pattern recognition in continuous seismic wavefield records using Self-Organizing Maps

Andreas Köhler,[1] Matthias Ohrnberger[2] and Frank Scherbaum[2]

[1] *Department of Geosciences, University of Oslo,* PO Box 1047, 0316 *Oslo, Norway. E-mail: andreas.kohler@geo.uio.no*
[2] *Institut für Erd- und Umweltwissenschaften, Universität Potsdam, Karl-Liebknecht Str.* 24, 14476 *Potsdam, Germany*

Accepted 2010 June 21. Received 2010 June 16; in original form 2009 June 11

SUMMARY

Modern acquisition of seismic data on receiver networks worldwide produces an increasing amount of continuous wavefield recordings. In addition to manual data inspection, seismogram interpretation requires therefore new processing utilities for event detection, signal classification and data visualization. The use of machine learning techniques automatises decision processes and reveals the statistical properties of data. This approach is becoming more and more important and valuable for large and complex seismic records. Unsupervised learning allows the recognition of wavefield patterns, such as short-term transients and long-term variations, with a minimum of domain knowledge. This study applies an unsupervised pattern recognition approach for the discovery, imaging and interpretation of temporal patterns in seismic array recordings. For this purpose, the data is parameterized by feature vectors, which combine different real-valued wavefield attributes for short time windows. Standard seismic analysis tools are used as feature generation methods, such as frequency–wavenumber, polarization and spectral analysis. We use Self-Organizing Maps (SOMs) for a data-driven feature selection, visualization and clustering procedure. The application to continuous recordings of seismic signals from an active volcano (Mount Merapi, Java, Indonesia) shows that volcano-tectonic and rockfall events can be detected and distinguished by clustering the feature vectors. Similar results are obtained in terms of correctly classifying events compared to a previously implemented supervised classification system. Furthermore, patterns in the background wavefield, that is the 24-hr cycle due to human activity, are intuitively visualized by means of the SOM representation. Finally, we apply our technique to an ambient seismic vibration record, which has been acquired for local site characterization. Disturbing wavefield patterns are identified which affect the quality of Love wave dispersion curve estimates. Particularly at night, when the overall energy of the wavefield is reduced due to the 24-hr cycle, the common assumption of stationary planar surface waves can be violated.

Key words: Neural networks, fuzzy logic; Probability distributions; Site effects; Volcano seismology; Volcano monitoring.

GJI Seismology

1 INTRODUCTION

Almost all research in observational seismology is based on recordings of ground motion caused by propagating seismic waves. Early observatory practice collected data sets that composed of only a limited amount of detectable earthquakes. Furthermore, there was only a small set of available seismic receivers. Therefore, manual seismogram analysis was standard procedure. Today, due to technical advances and the plethora of networks installed worldwide, an increasing amount of continuous data is produced. To find recorded earthquakes or any other temporal patterns of interest, these large seismic data sets can no longer be processed by hand. Manual analysis can be time-consuming, for instance when real-time processing is required for early warning systems. Another problem arising from analysing large data sets is the assessment of data quality. Because detailed, manual inspection of all waveform data may not be possible or be at least very laborious, one may easily miss instrumental failure or other disturbing patterns. Therefore, new analysis tools are required which utilize automatic pattern recognition techniques.

In the field of pattern recognition, two different data learning approaches are available. For supervised learning, labelled training data with known class-memberships are required to introduce the patterns to be recognized by the algorithm. In seismology, this approach primarily involves automatic detection of seismic phases (e.g. Christoffersson *et al.* 1988; Dai & MacBeth 1995; Wang & Teng 1997; Withers *et al.* 1998; Bai & Kennett 2000; Riggelsen *et al.* 2007) or discrimination of different event types (e.g. Joswig 1990; Dowla *et al.* 1990; Ohrnberger 2001). For this purpose, the training data are manually labelled based on expert knowledge (e.g. seismic phase picking). This training data set is then employed to predict class-memberships of unseen data. Hereby, the goal is to minimize the misclassification rate and the amount of false alarms. On the other hand, unsupervised learning uses unlabelled training data for automatic pattern identification. In practice, finding the natural grouping of the data set by clustering is the most common unsupervised technique. As in other disciplines, the benefit of unsupervised analysis in seismology lies in its potential to let the data speak for itself as an initial processing step. This phase should not be biased by preconceptions of the researcher (Bardainne *et al.* 2006). However, after this unsupervised learning phase it is still required that a domain expert (seismologist) evaluates and interprets the results as clustering algorithms do not reveal which cluster corresponds to a particular class of pattern. Moreover, most algorithms suggest more than one meaningful partition of the data set. Hence, human interaction is the last step of any unsupervised pattern recognition approach. The interpreter has to apply his domain knowledge by utilizing a practical visualization of the results. Finding such a visualization is also very crucial and difficult because the data space cannot always be displayed in two or three dimensions.

When we talk about patterns in seismic recordings, we usually mean distinct arrivals of wave phases. They are characterized by suddenly increasing amplitudes and/or a changing frequency content compared to the background wavefield. Besides intensified activity on signal detection, the availability of long, continuous network recordings allows another aspect of seismic data analysis to be addressed. Over the last two decades, there has been an increased focus on the permanently measured background wavefield (seismic noise or ambient seismic vibrations). Ambient noise can be considered as a superposition of waveforms excited by natural and man-made sources and has found to be energetically dominated by surface wave propagation. Therefore, estimating the propagation properties of surface waves from the noise record allows for passive investigations on crustal (e.g. Shapiro *et al.* 2005; Sabra *et al.* 2005) and local scales (e.g. Milana *et al.* 1996; Ohmachi & Umezono 1998; Bard 1998). This is especially relevant in areas of low seismicity and where the use of active geophysical experiments is limited. In this context, temporal patterns in the wavefield develop a more general meaning. In addition to short-term patterns (transients), the changes in wavefield characteristics over longer timescales (long-term patterns) can become important. For example, the spectral content or the directionality may change over hours, days or months and may have an impact on the quality of surface wave velocity estimates (Stehly *et al.* 2006; Pedersen & Krüger 2007). Automatic recognition techniques are required also for those patterns.

In this work, we apply an unsupervised approach for seismic wavefield analysis based on Self-Organizing Maps (SOMs; Kohonen 2001). Köhler *et al.* (2008) introduced an adaptive, unsupervised feature selection approach which automatically finds the seismic wavefield attributes suitable for pattern recognition. Furthermore, Köhler *et al.* (2009) applied SOM-based clustering and fast interpretable data visualization techniques to synthetic data and pre-selected sections of regional earthquake recordings. For the evaluation and validation of the procedure, quantitative performance tests have been carried out in both studies. In Section 2 of this paper, all employed techniques are introduced. The goal of this study is to apply the suggested approach in a further context. We focus on automatic recognition of patterns on different timescales using longer, continuous records. In particular, we consider volcano-seismic signals (Section 3) and temporal patterns in ambient seismic vibrations (Section 3 and 4). Detection and classification of volcano-seismic signals is important and mandatory for eruption forecasting and to assess the activity state of a volcano (Minakami 1960; McNutt 1996; Ohrnberger 2001). On the other hand, ambient vibration patterns reveal further insights into the temporal distribution of natural and anthropogenic sources (Bonnefoy-Claudet *et al.* 2006).

2 METHODS

2.1 Self-organizing maps

The SOM algorithm is a convenient, unsupervised learning method, which is widespread in various scientific fields (see references in Kohonen 2001). SOMs allow for an intuitive visualization of the distribution of data in any dimension. During SOM training, so-called prototype vectors are generated, whose distribution approximates the probability density function of the data set. This approach is well known as vector quantization. It is a very powerful approach to compress large and high-dimensional data sets. Each prototype vector represents a group of similar or close data points. For SOMs, an ordered and topology-preserving mapping into two dimensions is additionally performed. This map (the SOM) consists of a regular grid of usually hexagonal units. Each grid unit corresponds to a prototype vector in the data space. Fig. 1 illustrates this setting by means of a simple example. The data set consists of three clusters in a 3-D space (grey symbols in Fig. 1a). In Fig. 1(a), black symbols indicate the coordinates of all prototype vectors after training the SOM. The Prototype vectors in Fig. 1(a), which correspond to adjacent SOM units (hexagons) in Fig. 1(b), are connected by red lines. Fig. 1(a) shows that those prototype vectors are also neighbours in the data space, where one can imagine the SOM as a warped surface. This property helps to derive statements about similarity (proximity), distribution, and data grouping directly from the SOM.

There are different ways to visualize an SOM and the properties of the underlying data set. In Fig. 2(a), a histogram is shown on top of the SOM. The size of symbols corresponds to the number of data vectors which belong to a particular SOM unit. In other words, these data points have in common that the corresponding prototype vector is their nearest neighbour. New data, which were not used for training, can be easily mapped on the SOM in the same way. Fig. 2(b) shows the so-called unified-distance matrix (U-matrix). The U-matrix illustrates the probability density distribution of the prototype vectors. The space between adjacent SOM unit centres is coloured according to the distance between the corresponding vectors in the data space. The colour scale runs from blue (the lowest distance or highest data density) to red (highest distance). Because each SOM unit has six neighbours (except the outermosts), the hexagonal cells are divided into seven subunits in Fig. 2(b). The inner subunit shows the averaged distance to all neighbours.

The SOM prototype vectors can be clustered using common methods. Here, we employ an average linkage hierarchical clustering algorithm which makes use of the average distance between the members of two clusters to find meaningful groupings (Vesanto & Alhoniemi 2000). Starting with two clusters, the data set is thereby successively split into an increasing number of groups of similar prototype vectors. One can choose the most meaningful clustering manually or by using a cluster validity measure like the Davies–Bouldin index (DB; Davies & Bouldin 1979). The DB index compares the scatter within clusters and the distance between cluster means. High intercluster distances and low scatter within a cluster produces a low DB index relative to other clusterings of the same data set. Hence, the lowest index indicates the best grouping. After clustering, the SOM is coloured to distinguish the cluster memberships of prototypes (Fig. 2c). Comparison with the U-matrix allows

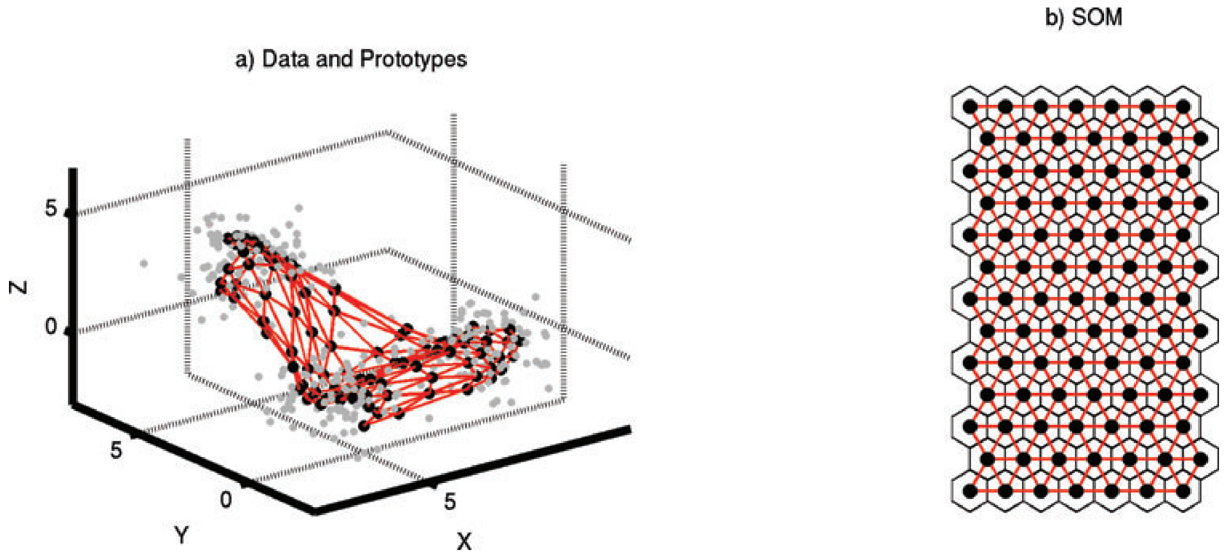

Figure 1. Example for a Self-Organizing Map (SOM) generated from a data set in three dimensions. Data space is shown in (a). Grey symbols indicate data. Black symbols represent SOM prototype vectors which sample the probability density function of data. Red lines connect prototypes whose corresponding SOM units are neighbours. The SOM is shown in (b). Each hexagon represents an SOM unit which corresponds to a prototype vector in (a).

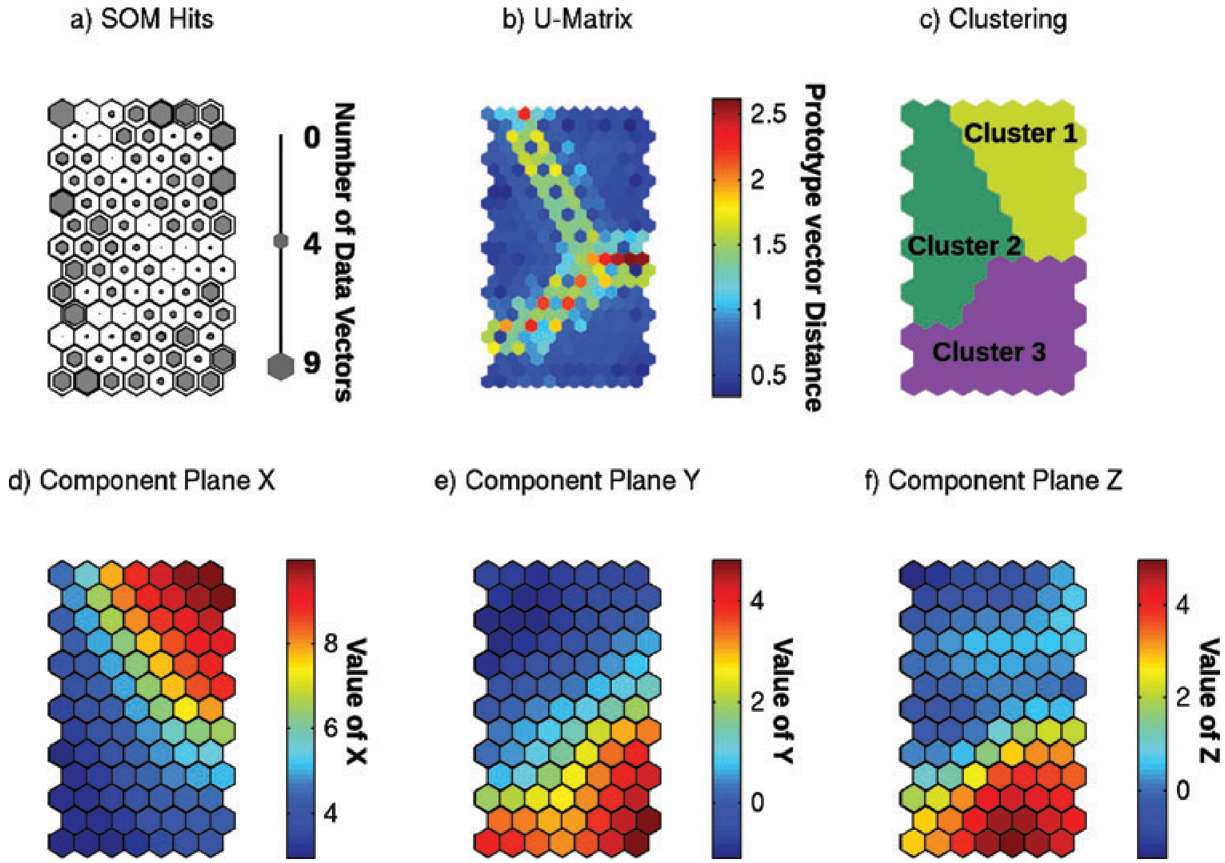

Figure 2. SOM visualizations corresponding to data set in Fig. 1. (a) Histogram of data on SOM. Sizes of black hexagons correspond to number of data vectors represented by an SOM unit. (b) Unified distance matrix (U-matrix) representing distances between prototype vectors in data space on the SOM. Each SOM unit is divided into seven subunit. Each subunit is coloured according to distance to neighbour unit. Areas with low distances (blue) indicate high data density (i.e. clusters). Furthermore, clusters are bounded by red colours (lower density). (c) Clustering of prototype vectors. Cluster membership of each SOM unit is indicated by colour. (d)–(f) Component planes: each SOM unit is coloured according to value of a particular prototype vector component.

Table 1. Features and short names for each method used to parameterise seismic data.

Method	Feature description	Short name[a]
1	*Frequency–wavenumber analysis (Kvaerna & Ringdahl 1986)*	
	Semblance: vertical, radial and tangential comp.	pr
2	*Spatial averaged autocorrelation method (Aki 1957)*	
	Real and imaginary (absolute value), autocorrelation coefficients averaged over station pairs vertical, radial and tangential comp. (Aki 1957; Asten 2006; Köhler *et al.* 2007)	spac, spacim
3	*Eigenvalues of complex 3c-covariance matrix*	
	Degree of polarization (Samson & Olson 1981)	dopII
	Ellipticity, strength of polarization, angle of incidence, planarity (Vidale 1986)	ell, sop, inc, plan
	Linearity, planarity (Jurkevics 1988)	rect, planII
	Linearity (2×), stability of direction cosine, enhanced linearity (Hearn & Hendrick 1999)	linII, linIII, sdc, elin
	Enhanced linear polarization (Bai & Kennett 2000)	elip
	Degree of polarization (Reading *et al.* 2001)	dopIII
4	*Complex seismic trace analysis (Taner et al. 1979)*	
	Instantaneous frequency and variance: vertical and horizontal comp.	if, vif
	Phase difference, ellipticity and tilt between vertical and horizontal components (René *et al.* 1986)	pdiff, ell, tilt
	Variance of azimuth, ellipticity (Morozov & Smithson 1996)	vazi, 3cell
	Component averaged instantaneous frequency (Bai & Kennett 2000)	FQ1
	Degree of polarization, linearity (Schimmel & Gallart 2004)	dop, lin
5	*Spectral attributes (Joswig 1990; Ohrnberger 2001)*	
	Horizontal and vertical power spectrum amplitudes in different frequency bands Normalized with overall energy (sonogram)	sono
	Dominant spectral frequency and bandwidth: vertical and horizontal comp.	domf, bb
	Logarithm of ratio between sum of lower and higher sonogram bands	ratiolf
6	*Spectrum of polarization ellipsoid (Pinnegar 2006)*	
	Normalized semi-mayor minus semi-minor axis and semi-minor axis of polarization Ellipsoid in different frequency bands	ab, b
7	*Amplitude ratios (after Jepsen & Kennett 1990)*	
	Real over imaginary part of complex trace, horizontal over vertical and east component	PQ, HV, HE

[a] Suffixes for short names when stated in the text or in figures: Component: z (vertical), e (east), n (north), h (horizontal), r (radial), t (tangential). Frequency band index: 1, 2, . . ., 3, (. . . , 10).

for the validation of clusterings directly on the SOM. For a perfect grouping, cluster borders should appear as more reddish areas in the U-matrix plot in comparison to the regions inside the clusters. In other words, a cluster is a bounded, blue area on the SOM. To investigate the meaning of clusters, SOM component plane plots are useful for displaying the values of a particular prototype vector component (equivalent to one of the so-called feature components) which is associated with any of the SOM units (Figs 2d–f). Red colours stand for high values of the corresponding feature. In our example, the component plane in Fig. 2(d) shows that Cluster 1 can be described by the distribution of feature X. The corresponding area on the SOM (Fig. 2c) is characterized by the presence of high values of that feature. Moreover, the component plane representation of the SOM allows to identify and group correlated features (i.e. features Y and Z).

SOMs have already been applied, in different contexts, for active seismic data sets (Essenreiter *et al.* 2001; Klose 2006; De Matos *et al.* 2007) and in seismology (Maurer *et al.* 1992; Musil & Plešinger 1996; Tarvainen 1999; Plešinger *et al.* 2000; Esposito *et al.* 2008; Köhler *et al.* 2009). The latter studies employed SOMs for event discrimination (e.g. explosions and earthquakes) using pre-selected seismogram sections. Our intention in this study is a more general pattern discovery using any continuous record.

2.2 Feature generation

The first step of any pattern recognition approach is the generation of useful features from the observables. The seismologist anticipates that seismic signals of the same type (e.g. a seismic phase) should be grouped into the same cluster. Besides direct use of seismogram or spectral amplitudes, also other parameters estimated from seismic data have been found to be useful to discriminate signal classes and improve interpretability. For instance, parameters like signal polarization can be more powerful to characterize a class of signals which are observed with variable maximum amplitudes. This applies for almost all seismic signal classes. Furthermore, information from a seismometer network can be combined in a single feature (e.g. the coherency of the wavefield). We implement several popular parametrization methods, such as frequency–wavenumber, polarization and spectral analysis. Each method generates short-time representatives for various wavefield properties. Table 1 gives an overview over all features. For each time window, features are computed from continuous three-component (array) seismograms. Features of single station methods are averaged over all receivers. All features obtained at one particular time-instance are combined as real-valued components of the so-called feature vector. Hence, the dimensionality of this vector corresponds to the number of features. The time window length for feature generation is specified according to the duration of the temporal patterns of interest. When this information is not available, the length should be at least the longest period which should be resolved. In contrast to previous unsupervised investigations on seismic data (Bardainne *et al.* 2006; Esposito *et al.* 2008), we do not compute a single feature vector for a time window including a complete event. Instead, we successively divide a continuous records into time segments without a previous (supervised) pre-selection of events of interest. Thus, a more general unsupervised learning approach can be performed.

The complete seismic record is analysed instead of clustering only a subset consisting of different types of events.

2.3 Feature selection

When knowledge about existing patterns is available *a priori*, it is appropriate to directly use a manually defined set of features. However, it is often not known which features are most suitable for pattern recognition. Furthermore, one does not want to employ more features than those that are necessary. Therefore, an unsupervised feature selection algorithm is needed. Our method is based on a relevancy and redundancy filter (Köhler *et al.* 2008, 2009).

The relevancy filter performs significance tests for individual features. In particular, we reduce the number of features using the non-parametric Wald–Wolfowitz runs test (Wald & Wolfowitz 1940). As the well-known Kolmogorov–Smirnow (KS) test, the runs test evaluates the hypothesis that data follow a particular distribution. The Wald–Wolfowitz method can be used to check the temporal randomness of a two-valued feature time-series (e.g. a binary sequence of ones and zeros). In other words, it tests whether all samples of a sequence are mutually independent. The runs test takes into account the order in which the samples are presented. A two-valued sequence can be obtained from any time-series by subtracting the mean or median of all values from each sample and keeping the sign. For our feature selection method, we choose a significance level of 5 per cent to reject feature time-series which can be explained as random sequences. Those features would deliver no usable information about temporal patterns. We found that the Wald–Wolfowitz runs test is working reliably on seismic features (Köhler *et al.* 2008). Alternatively, such a test may be performed also by the KS test.

The redundancy filter groups correlated features using the SOM component planes introduced above (Vesanto & Ahola 1999). A representative feature from each group is obtained from the relevance ranking given by the runs test (i.e. the confidence level that the feature sequence is non-random). Subsequently, the final SOM can be trained, clustered and interpreted by employing only the selected, uncorrelated features as input data vectors.

3 APPLICATION TO A VOLCANO-SEISMIC WAVEFIELD

We apply the unsupervised clustering and visualization technique to array recordings of seismic signals at Mount Merapi, which is an active, high-risk volcano on Java (Indonesia). Our aim is to identify events and investigate the behaviour of the background wavefield. We use seismic array data from the beginning of July 1998 during a phase of high volcanic activity. The network (KEN) consists of three broadband three-component stations (aperture ∼200 m). The analysed recordings are spread over five days (July 2–6). We chose 27 1-hour-long records for which the occurrence of volcano-seismic events is known. We use three types of hand-picked events for evaluation after learning: 16 shallow volcano-tectonic events (VT type B or VTB, 5–8 Hz), six multiphase events due to lava dome growth (MP, 3–4 Hz) and 16 rockfall-induced signals (so-called Guguran events, 1–20 Hz). All those events are a subset of the training data set employed by Ohrnberger (2001) for a supervised, hidden Markov model-based classification system. A time window length of 1.7 s is used for feature generation in frequency bands between 0.8 and 16 Hz.

3.1 Feature selection

Fig. 3 shows histograms for all features automatically selected by our procedure. Because the background wavefield dominates the data set, no evidences for clusters of volcano-seismic signals can be derived directly from a linear-scaled histogram (e.g. a mix of two or more normal distributions). Therefore, we plot all histograms using a log-scale to be able to identify more local maxima in the distributions. Fig. 3 shows that features like *sono_z5*, *ab_10* and *pr_r3* have the potential to discriminate between different signals. For all these features, the distribution shows a locally increasing frequency of occurrence at high values (i.e. higher energy or coherency).

We compare the selected features with those manually chosen by Ohrnberger (2001) (*sono_z* in eight frequency bands, *pr_z*, *inc*). In both studies, features derived from the time-frequency spectrum and from array methods (i.e. the semblance) are found to be the most useful wavefield attributes. In addition, Ohrnberger (2001) used the angles of incidence which is not selected by our automatic method. However, in our feature set, information about signal polarization is implicitly available from the ellipticity spectrum (*ab*) and the horizontal to vertical spectral ratio (*HV*). Therefore, the reliability of our feature selection procedure is shown for this example. There is a good correspondence with the features chosen based on seismological expert knowledge.

3.2 Clustering

Fig. 4 presents the visualizations after SOM training and clustering. The U-matrix in Fig. 4(a) shows areas of high data density (blue colours) bounded by low densities (red colours). Two possible clusterings are shown in Figs 4(b) and (c). A number of four clusters (Fig. 4b) is obtained as the best grouping with respect to the Davies–Bouldin cluster validity index (Table 2). However, it was suggested to use the DB index as a guideline rather than to accept only one meaningful clustering (Vesanto & Alhoniemi 2000). Therefore, a second solution is chosen providing a more meaningful fit of all observed U-matrix patterns (10 clusters in Fig. 4c). Nevertheless, note that all these clusters are children of the first solution, since a hierarchical approach is carried out. Hence, no distinct solution, but rather a more detailed view into the data set properties is presented.

In Figs 4(d)–(f), the manually labelled time windows of all volcano-seismic events (VTB, Guguran, MP) are mapped on the SOM as explained in Section 2. The SOM-histogram shows that the well-separated Cluster 2 in Fig. 4(b) contains volcano-tectonic and rockfall events. Moreover, its two children are able to discriminate between both event types (Fig. 4c). Hence, it is possible to find at least two volcano-seismic signal classes, even when the background wavefield dominates the data set. On the other hand, multi-phase events are distributed over the entire map, and no distinct cluster is found.

To interpret the remaining clusters, we present the vertical component seismograms and the cluster memberships of all time windows from Fig. 4(c) within their original temporal context for one array station (Fig. 5). Volcano-tectonic (orange) and rockfall events (yellow) are clearly visible at any time of day. In the daytime, a further class of transients occurs (Cluster 1), which is most likely man-made. The dark green dominated background wavefield in the daytime is more heterogeneous due to various human activity. Furthermore, blue and violet colours highlight the background noise during the night. Except for volcano-seismic events, no other transients occur. The observed transition from green to blue-coloured

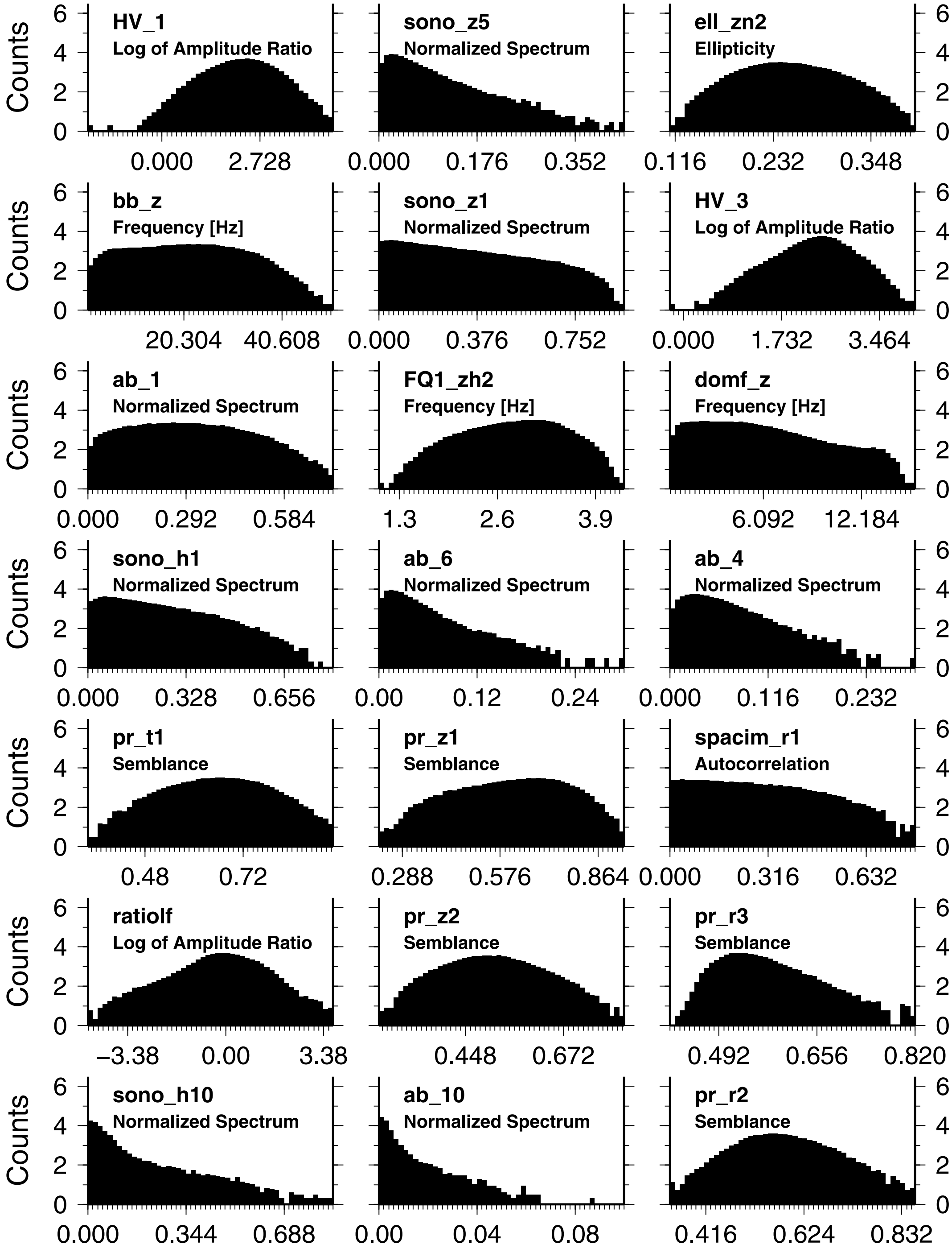

Figure 3. Features automatically selected for the data set recorded at volcano Merapi. Histograms show distributions of feature values. Scale for counts is in $\log_{10}$. Name of feature and x-axis label is given within each panel (see also Table 1).

time windows reflects the successively increasing human activity in the early morning around 6 a.m., shortly before sunrise in tropical regions.

3.3 Cluster-based classification

For a quantitative evaluation, we assign the most frequent label (VTB, Guguran, MP, or noise) to each of all 10 clusters using the hand-picked data. As already suggested by Figs 4(d)–(f), rockfall events are associated with the yellow and volcano-tectonic events with the orange cluster. Subsequently, we map the volcano-tectonic and rockfall events on the SOM again and reclassify those time windows according to their cluster memberships. We obtain a low classification error of 6 per cent for volcano-tectonic events (presented as VTBs but classified as others, false negative). For rockfall events, a higher error of 26 per cent is yielded, due to ambiguities between both event types (Guguran time windows in VTB cluster). By combining both event clusters into one class, we are able to reduce the recognition error significantly to 12 per cent. For a blind

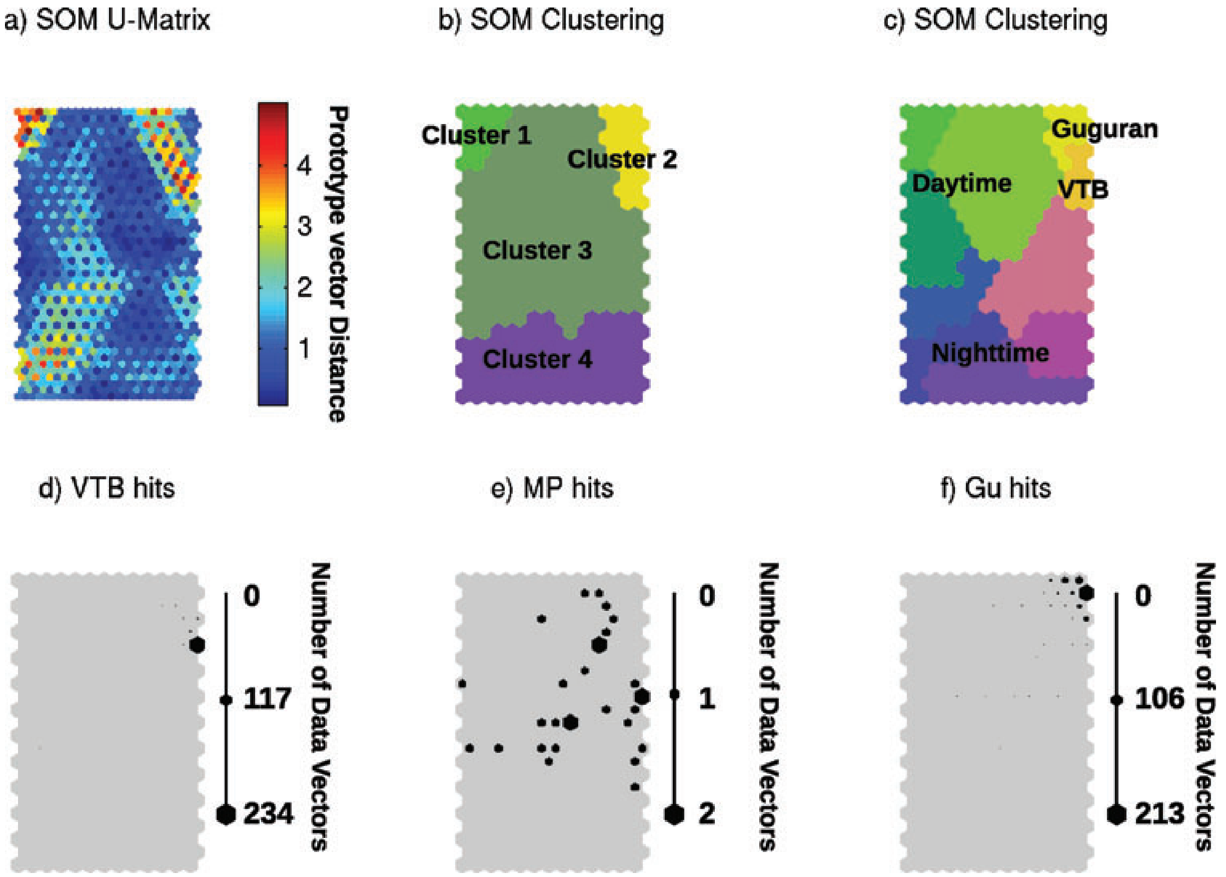

Figure 4. SOM visualizations for Merapi data set. U-matrix in (a) shows distances in data space. Range of data can be obtained from Fig. 3. Colouring in (b) and (c) distinguishes cluster. In (b), number of clusters is four and for (c) 10 groups are chosen manually. In (c), interpretation of clustering is given. Lower panels show SOM hits (SOM-histograms) of volcano-seismic signals observed at Mount Merapi: volcano-tectonic (VTB), multi-phase (MP) and rockfall events (Guguran).

Table 2. Davies–Bouldin cluster validity index for different number of clusters (Merapi data set).

Number of clusters	2	3	4	5	6	7	8	9	10	11	12	13	14	15
DB index	1.18	1.17	1.10	1.36	1.31	1.35	1.44	1.48	1.40	1.43	1.39	1.41	1.36	1.38

Note: The lowest value indicates the best clustering solution.

test, we also determine the false alarm rate from the background wavefield using the same 27 hr of data (wrongly classified as events, false positive). We find that 28 per cent of all time windows classified as volcano-tectonic or rockfall events are not belonging to a visually identifiable event.

Our results show that the volcano-tectonic and rockfall clusters can be used for a simple nearest-neighbour classifier for new incoming data. Higher recognition rates can be achieved when additional expert knowledge is used. Because both events last longer than just 1.7 s, a time window classified as belonging to an event should belong to a sequence of identical class labels. Hence, we may claim that a sequence has to be composed of more than one time window. For our data set, this reduces the false alarm rate from 28 to 20 per cent.

The temporal context of feature vectors was also considered by the approach of Ohrnberger (2001). For a continuous mode application of the trained classifier on five days of data, an averaged misclassification rate of 33 per cent was obtained. The lowest errors have been found for volcano-tectonic (11 per cent), followed by rockfall events (26 per cent). The worst recognition rate has been observed for multi-phase events (36 per cent misclassification) due to weak amplitudes. Our findings are in line with those results, even though we pursue a different strategy. However, in contrast to Ohrnberger (2001), we are not able to identify automatically even a single multiphase event.

4 APPLICATION TO AMBIENT SEISMIC VIBRATIONS

Important target sites for ambient seismic vibration measurements are areas of high seismic hazard for which strong amplification, due to soft-sediments in the subsurface, are expected. Thus, knowledge about the local structure is mandatory for the forecasting of ground motion. As a low-cost alternative to, or in combination with, active geophysical experiments or other geotechnical methods, this information can be inverted from surface wave dispersion curves. These quantities are obtained by stacking the slowness estimates from array recordings over an appropriate time interval (e.g. Herrmann 2002; Scherbaum *et al.* 2003; Wathelet *et al.* 2004; Parolai *et al.* 2005).

We process 17 hr of array data (12 receivers, aperture ~250 m). The record has been acquired at a site close to the village

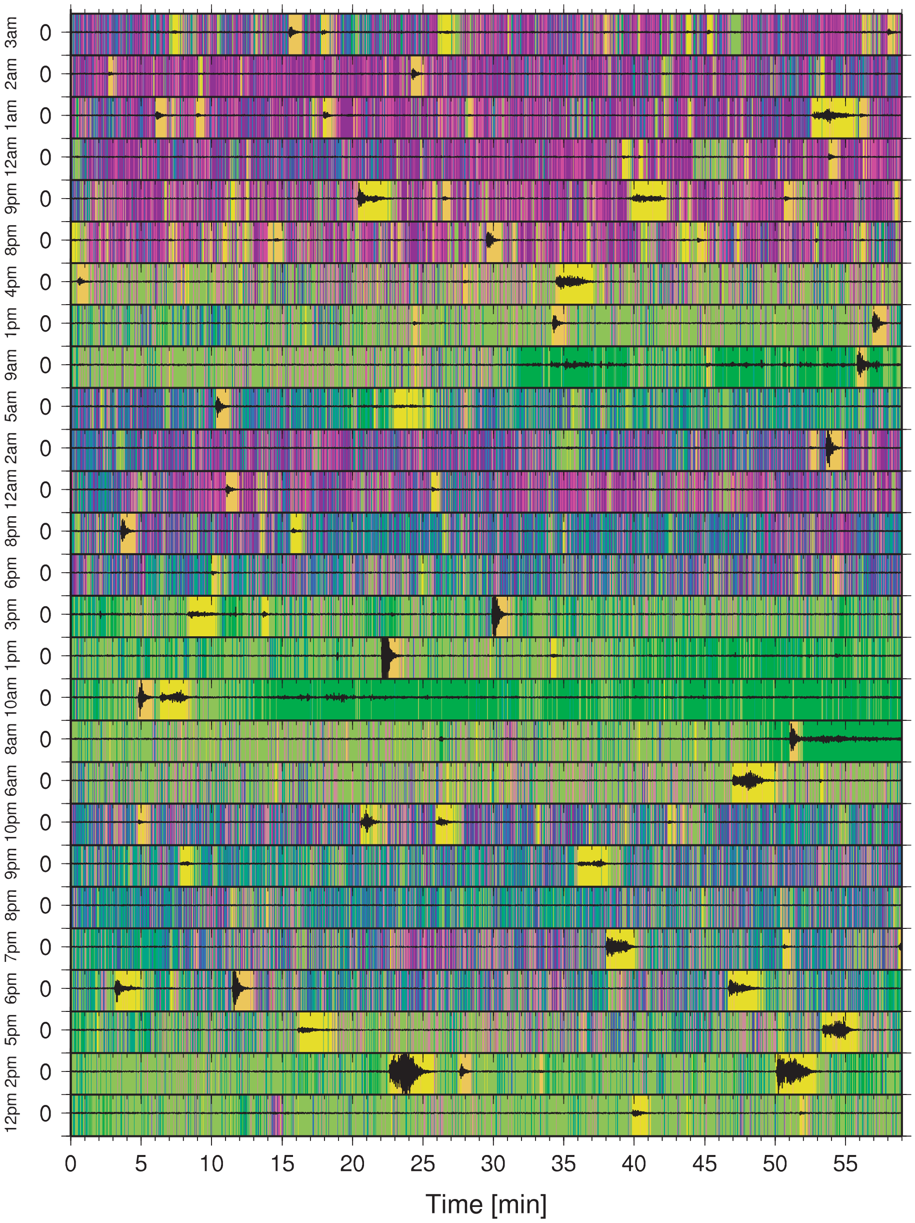

Figure 5. Vertical component seismograms of one array station for Merapi data for all analysed hours. Seismograms are bandpass-filtered between 0.3 and 19 Hz. Background colouring corresponds to cluster membership of each time window. SOM clustering and colours from Fig. 4(c) are used. Yellow time windows correspond to rockfall events and orange windows to volcano-tectonic events. Transition from greenish to bluish time windows reflects change in background wavefield due to human activity.

of Colfiorito in Central Italy. The first 3 hr (1–4 p.m.) have been recorded 1 day before the rest of the data. Rayleigh and Love wave dispersion curves are estimated for each hour using the modified spatial autocorrelation method on three components (3c-MSPAC; Köhler *et al.* 2007). Whereas Rayleigh wave dispersion curves remain stable, we observed significant variation in the Love dispersion curves below a frequency of 1 Hz. Therefore, we decided to apply our unsupervised pattern recognition approach to investigate this phenomenon.

Because we know that we are looking for a pattern on the horizontal components in a particular frequency band, we choose a manually defined feature set for our unsupervised recognition method (Section 2). We use a set of six features (not including the slowness estimates) generated from the horizontal components in a frequency band between 0.4 and 1.2 Hz. We compute four features which reflect the coherency and plane wave character of the tangential and radial-polarized wavefield (*pr_r*, *pr_t*, *spacim_r*, *spacim_t*; Table 1). Furthermore, variation in the horizontal frequency spectrum is considered by the instantaneous frequency (*if_h*) and the power spectrum amplitude normalized with the overall energy between 0.4 and 2.6 Hz (*sono_h*). The time window length for all features is 100 s. We choose this length to be in a similar range compared to the typical durations of transients in the record. The reason is that we are now more interested in long-term patterns and do not aim to resolve patterns on a very small scale (i.e. the changing character of a transient over duration of the signal).

Figs 6 and 7 show the SOM visualizations and the component planes of all features. We choose the clustering in Fig. 6(b) because

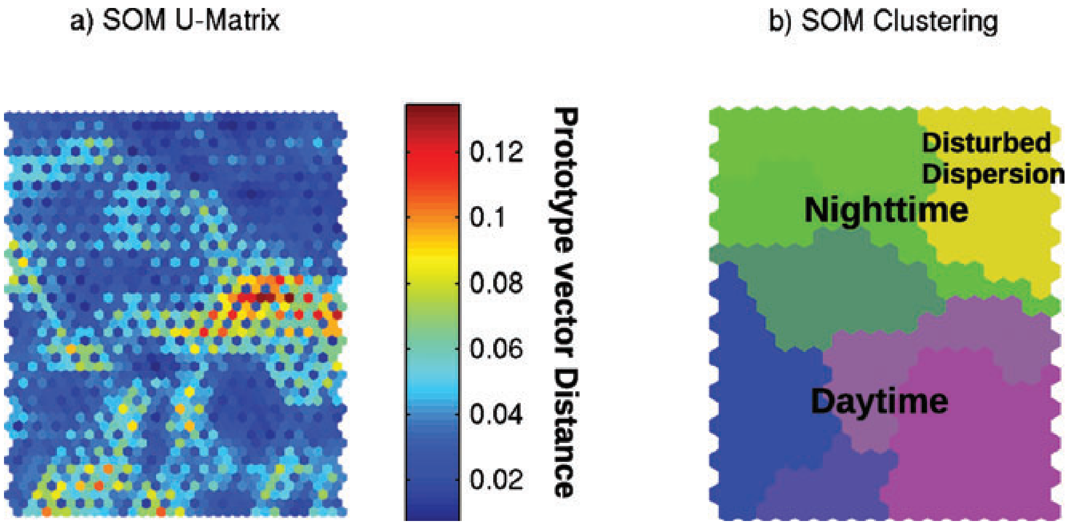

Figure 6. SOM visualizations for ambient vibration wavefield recorded close to Colfiorito (Umbria-Marche region, Central-Italy). U-matrix in (a) shows distances in data space. Colouring in (b) distinguishes clusters. Interpretation of clustering is given (see also text).

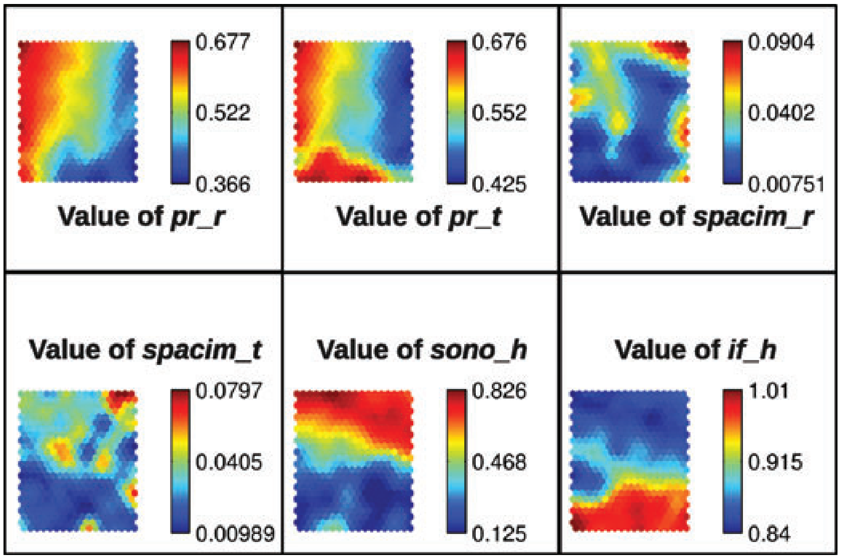

Figure 7. SOM component planes for ambient vibration data set. Features are semblance on the radial (*pr_r*) and tangential component (*pr_t*), imaginary part of the averaged autocorrelation coefficient on the radial (*spacim_r*) and tangential component (*spacim_t*), normalized amplitude of power spectrum on horizontal component (*sono_h*) and instantaneous frequency (*if_h*). More information about features is given in Table 1.

it is in good agreement with the identifiable regions shown by the U-matrix in Fig. 6(a). Fig. 8 presents the seismograms, the cluster membership of each time window, and the Love wave dispersion curve for each hour of the record. As for the previous data set, the daily cycle due to human activity is highlighted. The green and yellow clusters (nighttime) are mainly defined by increased energy contribution at lower frequencies compared to the blue and violet clusters (daytime) (compare Figs 6b and 7). Furthermore, high semblances are observed on the radial and tangential components for the green and blue clusters. On the other hand, the radial is clearly lower than the tangential semblance for the violet clusters, which include mainly high-amplitude signals. It is known that anthropogenic seismic noise at daytime mainly consists of ground motion above 1 Hz and is dominated by Love waves (Bonnefoy-Claudet *et al.* 2006). Furthermore, transients exist due to close sources like local traffic. On the other hand, Rayleigh-wave-dominated oceanic microseisms have a higher contribution below 1 Hz and dominate at nighttime (Bonnefoy-Claudet *et al.* 2006). This behaviour is well reflected by the clustering of all time windows and by their properties. There is a main road and agricultural activity close to the measurement site, which probably generates the violet signal clusters. The continuous high noise level during daytime (blue cluster) can be associated with human activity in the village of Colfiorito. At night, oceanic microseisms become dominant compared to the anthropogenic noise level.

The right-hand panel of Fig. 8 shows that the estimated Love wave dispersion curves become unstable below 1 Hz between 9 p.m. and 1 a.m.. Unrealistic high slowness values and increased uncertainties are obtained. Within the same time interval, most feature vectors belong to the yellow cluster. Low semblance and high imaginary SPAC coefficients indicate that the assumption of planar and coherent surface waves is not fulfilled within this cluster

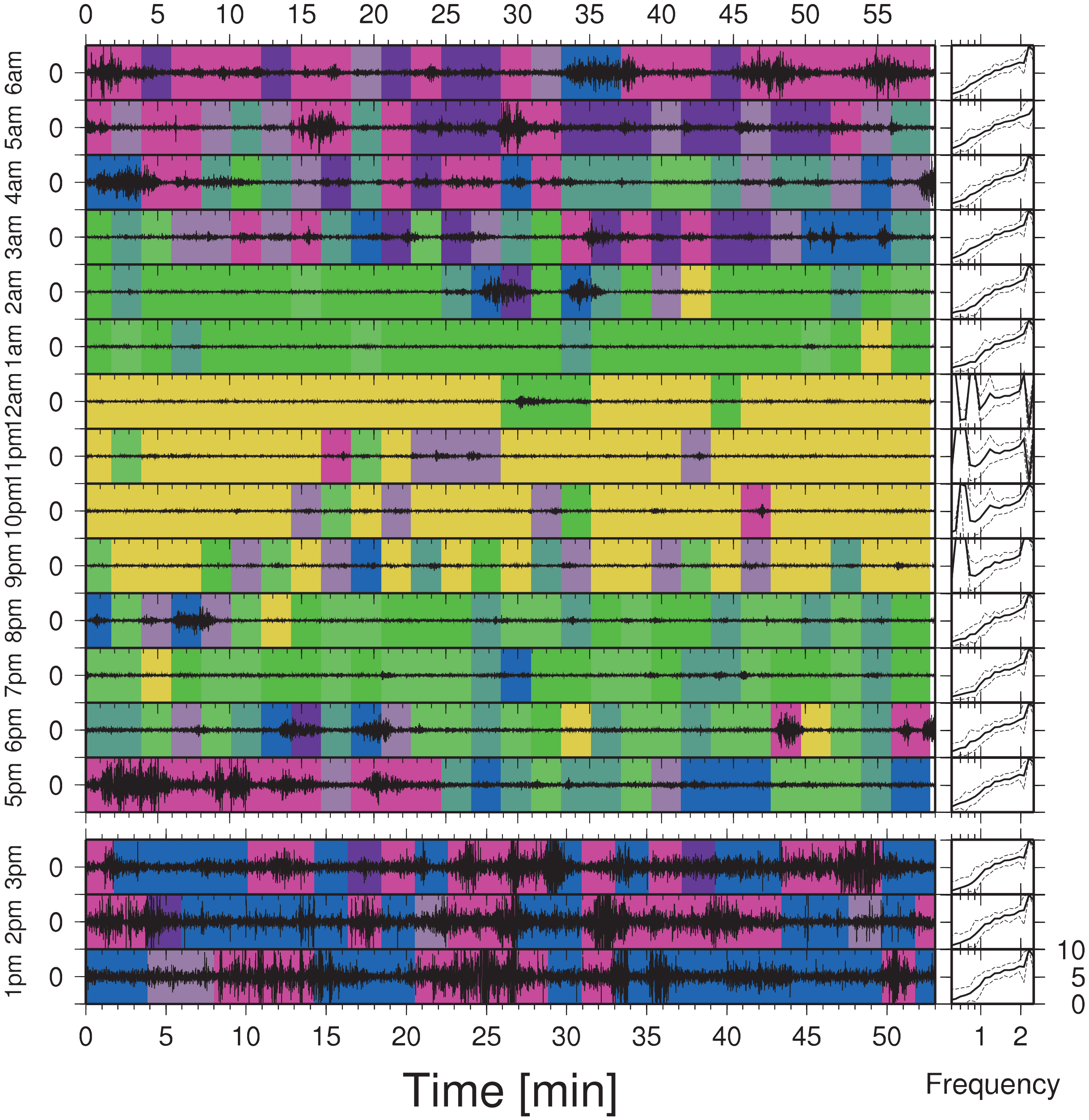

Figure 8. Vertical component seismograms of one array station for ambient vibration data set for all analysed hours. Background colouring corresponds to cluster membership of each time window. SOM clustering and colours from Fig. 6b are used. Blue and violet coloured time windows correspond to seismic noise at daytime due to human activity close to the measurement site. Greenish and yellow time windows show background wavefield at night. Yellow time windows are associated with unstable dispersion curve estimates. Right panel shows Love wave dispersion curves averaged over each hour. The Y-axis labels are given in slowness ($\mathrm{s\,km^{-1}}$).

(Fig. 7). An explanation could be the lack of Love waves in the noise wavefield due to the minimal human activity.

5 CONCLUSIONS

In this study, we have applied an unsupervised processing scheme for the general discovery of temporal patterns in continuous seismic wavefield records based on SOMs. The technique has shown to be a very useful tool to have a first and unbiased look on seismic data. It has the potential to identify long-term variations and short seismic events. Manual inspection of suggested clustering solutions and interpretation based on expert knowledge is an integral part of this approach. Selecting automatically the number of clusters based on a validity measure for example, may not tap the full discrimination potential of the approach.

The method has shown its potential to monitor variations in wavefields close to active volcanos. We have analysed recordings of seismic signals from the volcano Mount Merapi (Indonesia). We have been able to detect and discriminate two characteristic volcano-seismic signal classes which are crucial for eruption forecasting. We have found that they form distinct clusters over the entire day independent of the event amplitudes, which allows to classify those events automatically. Compared to a previously implemented supervised classifier, similar recognition rates have been obtained.

Moreover, we have confirmed the 24-hr cycle, which is known to be related to human activity.

Furthermore, we have applied our method to an ambient seismic vibration data set. We have shown the potential to recognize long-term variations which may potentially affect estimates of surface wave dispersion curves. The analysed data set could be decomposed into clusters of waveforms recorded at day- and nighttime. Furthermore, we have identified a cluster of time windows at night which have impaired the estimate of a local Love wave dispersion curve. By omitting this cluster before stacking the dispersion curves of all time windows, it is possible to decrease the measurement uncertainty. Therefore, this approach can be applied as a pre-processing method for local site characterization based on ambient seismic vibrations.

ACKNOWLEDGMENTS

The study was funded by a PhD-scholarship of the University of Potsdam. Parts of this work have been made possible under the EU-Projects NERIES (contract no. 026130) and SESAME (Colfiorito data set). We used the SOM toolbox implemented in MATLAB® of Vesanto *et al.* (2000). We would like to express our thanks to both anonymous reviewers.

REFERENCES

Aki, K., 1957. Space and time spectra of stationary stochastic waves, with special reference to microtremors, *Bull. Earthq. Res. Inst., Univ. Tokyo,* **35,** 415–456.

Asten, M., 2006. On bias and noise in passive seismic data from finite circular array data processed using SPAC methods, *Geophysics,* **71**(6), 153–162.

Bai, C. & Kennett, B., 2000. Automatic phase-detection and identification by full use of a single three-component broadband seismogram, *Bull. seism. Soc. Am.,* **90**(1), 187–198.

Bard, P., 1998. Microtremor measurements: a tool for site effect estimation?, in *Proceedings of the Second International Symposium on the Effects of Surface Geology on Seismic Motion,* Vol. 3, pp. 1251–1279, eds Irikura, K., Kudo, K., Okada, H. & Sasatani, T., Balkema, Rotterdam.

Bardainne, T., Gaillot, P., Dubos-Sallée, N., Blanco, J. & Sén échal, G., 2006. Characterization of seismic waveforms and classification of seismic events using chirplet atomic decomposition. Example from the Lacq gas field (Western Pyrenees, France), *Geophys. J. Int.,* **166**(47), 699–718.

Bonnefoy-Claudet, S., Cotton, F. & Bard, P., 2006. The nature of noise wavefield and its applications for site effects studies A literature review, *Earth Sci. Rev.,*(3–4), 205–227.

Christoffersson, A., Husebye, E. & Ingate, S., 1988. Wavefield decomposition using ML-probabilities in modelling single-site 3-component records, *Geophys. J. Int.,* **93**(2), 197–213.

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.

Davies, D. & Bouldin, D., 1979. A cluster separation measure, *IEEE Trans. Pattern Anal. Machine Intell.,* **1**(2), 224–227.

De Matos, M., Osorio, P. & Johann, P., 2007. Unsupervised seismic facies analysis using wavelet transform and self-organizing maps, *Geophysics,* **72,** 9–21.

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

Esposito, A., Giudicepietro, F., D'Auria, L., Scarpetta, S., Martini, M., Col telli, M. & Marinaro, M., 2008. Unsupervised neural analysis of very-long-period events at Stromboli volcano using the self-organizing maps, *Bull. seism. Soc. Am.,* **98**(5), 2449–2459.

Essenreiter, R., Karrenbach, M. & Treitel, S., 2001. Identification and classification of multiple reflections with self-organizing maps, *Geophys. Prospect.,* **49**(3), 341–352.

Hearn, S. & Hendrick, N., 1999. A review of single-station time-domain polarisation analysis techniques, *J. Seismic Explor.,* **8,** 181–202.

Herrmann, R., 2002. *Computer Programs in Seismology: An Overview of Synthetic Seismogram Computation, Version 3.20,* 183 pp. Department of Earth and Atmospheric Sciences, Saint Louis University.

Jepsen, D. & Kennett, B., 1990. Three-component analysis of regional seismograms, *Bull. seism. Soc. Am.,* **80**(6 B), 2032–2052.

Joswig, M., 1990. Pattern recognition for earthquake detection, *Bull. seism. Soc. Am.,* **80**(1), 170–186.

Jurkevics, A., 1988. Polarization analysis of three-component array data, *Bull. seism. Soc. Am.,* **78**(5), 1725–1743.

Klose, C., 2006. Self-organizing maps for geoscientific data analysis: geological interpretation of multidimensional geophysical data, *Comput. Geosci.,* **10**(3), 265–277.

Köhler, A., Ohrnberger, M., Scherbaum, F., Wathelet, M. & Cornou, C., 2007. Assessing the reliability of the modified three-component spatial autocorrelation technique, *Geophys. J. Int.,* **168**(2), 779–796.

Köhler, A., Ohrnberger, M., Riggelsen, C. & Scherbaum, F., 2008. Unsupervised feature selection for pattern search in seismic time series, *J. Mach. Learn. Res., Workshop and Conference Proceedings: New challenges for feature selection in data mining and knowledge discovery,* **4,** 106–121.

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.

Kohonen, T., 2001. *Self-Organizing Maps,* Springer Series in Information Sciences, Vol. 30, Third Extended Edition, 501 pp, Springer Berlin, Heidelberg, New York, 1995, 1997, 2001.

Kvaerna, T. & Ringdahl, F., 1986. Stability of various fk estimation techniques, Technical Report, Semianual Technical Summary 1-86/87, 1 October 1985 to 31 March 1986, NORSAR Scientific Report, Kjeller, Norway, 20 pp.

Maurer, W., Dowla, F. & Jarpe, S., 1992. Seismic event interpretation using self-organizing neural networks, in *Proceedings of the International Society for Optical Engineering (SPIE),* Vol. 1709, pp. 950–958, doi:10.1117/12.139971.

McNutt, S., 1996. Seismic monitoring and eruption forecasting of volcanoes: a review of the state-of-the-art and case histories, in *Monitoring and Mitigation of Volcano Hazards,* pp. 99–146.

Milana, G., Barba, S., Del Pezzo, E. & Zambonelli, E., 1996. Site response from ambient noise measurements: new perspectives from an array study in Central Italy, *Bull. seism. Soc. Am.,* **86**(2), 320–328.

Minakami, T., 1960. Fundamental research for predicting volcanic eruptions (Part 1). Earthquakes and crustal deformations originating from volcanic activities, *Bull. Earthq. Res. Inst.,* **38,** 497–544.

Morozov, I. & Smithson, S., 1996. Instantaneous polarization attributes and directional filtering, *Geophysics,* **61,** 872–881.

Musil, M. & Plešinger, A., 1996. Discrimination between local microearthquakes and quarry blasts by multi-layer perceptrons and Kohonen maps, *Bull. seism. Soc. Am.,* **86**(4), 1077–1090.

Ohmachi, T. & Umezono, T., 1998. Rate of Rayleigh waves in microtremors, in *Proceeding of the Second International Symposium on the Effects of Surface Geology on Seismic Motion,* pp. 587–592, eds Irikura, K., Kudo, K., Okada, H. & Sasatani, T., Balkana, Rotterdam.

Ohrnberger, M., 2001. Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia, *PhD thesis*, University of Potsdam, http://opus.kobv.de/ubp/volltexte/2005/31/pdf/ohrnberg.pdf, 158 pp (accessed on 31 March 2009).

Parolai, S., Picozzi, M., Richwalski, S. & Milkereit, C., 2005. Joint inversion of phase velocity dispersion and H/V ratio curves from seismic noise recordings using a genetic algorithm, considering higher modes, *Geophys. Res. Lett.,* **32,** L01303, doi:10.1029/2004GL021115.

Pedersen, H. & Krüger, F., 2007. Influence of the seismic noise characteristics on noise correlations in the Baltic shield, *Geophys. J. Int.,* **168**(1), 197–210.

Pinnegar, C., 2006. Polarization analysis and polarization filtering of three-component signals with the time-frequency S transform, *Geophys. J. Int.,* **165**(2), 596–606.

Plešinger, A., Růžek, B. & Boušková, A., 2000. Statistical interpretation of WEBNET seismograms by artificial neural nets, *Studia Geophysica et Geodaetica,* **44**(2), 251–271.

Reading, A., Mao, W. & Gubbins, D., 2001. Polarization filtering for automatic picking of seismic data and improved converted phase detection, *Geophys. J. Int.,* **147**(1), 227–234.

René, R., Fitter, J., Forsyth, P., Kim, K., Murray, D., Walters, J. & Westerman, J., 1986. Multicomponent seismic studies using complex trace analysis, *Geophysics,* **51,** 1235–1251.

Riggelsen, C., Ohrnberger, M. & Scherbaum, F., 2007. Dynamic bayesian networks for real-time classification of seismic signals, *Lecture Notes Comput. Sci.,* **4702,** 565–572.

Sabra, K., Gerstoft, P., Roux, P., Kuperman, W. & Fehler, M., 2005. Extracting time-domain Green(tm)s function estimates from ambient seismic noise, *Geophys. Res. Lett.,* **32,** L03310, doi:10.1029/2004GL021862.

Samson, J. & Olson, J., 1981. Data-adaptive polarization filters for multichannel geophysical data, *Geophysics,* **46,** 1423–1431.

Scherbaum, F., Hinzen, K. & Ohrnberger, M., 2003. Determination of shallow shear wave velocity profiles in the Cologne, Germany area using ambient vibrations, *Geophys. J. Int.,* **152**(3), 597–612.

Schimmel, M. & Gallart, J., 2004. Degree of polarization filter for frequency-dependent signal enhancement through noise suppression, *Bull. seism. Soc. Am.,* **94**(3), 1016–1035.

Shapiro, N., Campillo, M., Stehly, L. & Ritzwoller, M., 2005. High-resolution surface-wave tomography from ambient seismic noise, *Science,* **307**(5715), 1615–1618.

Stehly, L., Campillo, M. & Shapiro, N., 2006. A study of the seismic noise from its long-range correlation properties, *J. geophys. Res,* **111,** 1–12.

Taner, M., Koehler, F. & Sheriff, R., 1979. Complex seismic trace analysis, *Geophysics,* **44,** 1041–1063.

Tarvainen, M., 1999. Recognizing explosion sites with a self-organizing network for unsupervised learning, *Phys. Earth planet. Int.,* **113**(1–4), 143–154.

Vesanto, J. & Ahola, J., 1999. Hunting for correlations in data using the self-organizing map, in *Proceedings of the International Congress on Computational Intelligence Methods and Applications (CIMA 99), International Computing Sciences Conferences (ICSC),* Academic Press, pp. 279–285.

Vesanto, J. & Alhoniemi, E., 2000. Clustering of the self-organizing map, *IEEE Trans. Neural Network,* **11**(3), 586–600.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., Team, S. & Oy, L., 2000. Som toolbox for matlab, *Techn. Ber., Helsinki University of Technology*.

Vidale, J., 1986. Complex polarization analysis of particle motion, *Bull. seism. Soc. Am.,* **76**(5), 1393–1405.

Wald, A. & Wolfowitz, J., 1940. On a test whether two samples are from the same population, *Ann. Math. Stat.,* **11**(2), 147–162.

Wang, J. & Teng, T., 1997. Identification and picking of S phase using an artificial neural network, *Bull. seism. Soc. Am.,* **87**(5), 1140–1149.

Wathelet, M., Jongmans, D. & Ohrnberger, M., 2004. Surface wave inversion using a direct search algorithm and its application to ambient vibration measurements, *Near Surface Geophysics,* **2,** 211–221.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S. & Trujillo, J., 1998. A comparison of select trigger algorithms for automated global seismic phase and event detection, *Bull. seism. Soc. Am.,* **88**(1), 95–106.