Contents lists available at ScienceDirect

Computers & Geosciences

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

Unsupervised feature selection and general pattern discovery using Self-Organizing Maps for gaining insights into the nature of seismic wavefields

Andreas Köhler *, Matthias Ohrnberger, Frank Scherbaum

Institut für Geowissenschaften, Universität Potsdam, Karl-Liebknecht-Str. 24, Haus 27, 14476 Golm, Germany

ARTICLE INFO

Article history:
Received 20 February 2008
Received in revised form
13 February 2009
Accepted 18 February 2009

Keywords:
Data mining
Unsupervised learning
Clustering
Feature selection
Seismic features

ABSTRACT

This study presents an unsupervised feature selection and learning approach for the discovery and intuitive imaging of significant temporal patterns in seismic single-station or network recordings. For this purpose, the data are parametrized by real-valued feature vectors for short time windows using standard analysis tools for seismic data, such as frequency-wavenumber, polarization, and spectral analysis. We use Self-Organizing Maps (SOMs) for a data-driven feature selection, visualization and clustering procedure, which is in particular suitable for high-dimensional data sets. Our feature selection method is based on significance testing using the Wald–Wolfowitz runs test for individual features and on correlation hunting with SOMs in feature subsets. Using synthetics composed of Rayleigh and Love waves and real-world data, we show the robustness and the improved discriminative power of that approach compared to feature subsets manually selected from individual wavefield parametrization methods. Furthermore, the capability of the clustering and visualization techniques to investigate the discrimination of wave phases is shown by means of synthetic waveforms and regional earthquake recordings.

1. Introduction

Due to the increasing amount of data available from networks that monitor seismicity worldwide, automatic detection and classification of seismic events is becoming more and more important in seismology. Regarding the literature in this field, great progress has been made since the emergence of simple STA/LTA triggers in the early 1960s (see review Withers et al., 1998). More recent techniques well established in other research fields, such as pattern matching (Joswig, 1990), neural networks (Dai and MacBeth, 1995), hidden Markov models (Ohrnberger, 2001) and dynamic Bayesian networks (Riggelsen et al., 2007), have been applied to seismic recordings. However, automated seismic data processing still has to rely on expert knowledge of seismologists. In fact, selecting a suitable data parametrization (e.g. Bai and Kennett, 2000), manually defining pre-classified training data, and validating reliability of the methods in practice, are crucial and integral parts of supervised learning techniques.

On the other hand, in the context of seismogram analysis and interpretation, less attention has been paid to unsupervised learning. Here, no expert knowledge (class labels for the training data) is required in the learning phase. The probably best known unsupervised technique is clustering, the automatic and meaningful grouping of unlabeled data. Bardainne et al. (2006) pointed out that clustering can be useful for the interpretation of earthquake recordings, because the inherent structure of the seismic data is shown without being clouded by preconceptions of the researcher. Thus, in grouping the data, the seismologist gets an impression about occurrence and characteristics of seismic signals or phases contained in the wavefield. This approach can be particularly useful when size and complexity of the data set hinder a fast, visual inspection. However, interpretation of the discovered patterns and, as for supervised classification, defining a set of suitable features which parametrize the seismic data, is still a task of the seismologist. The latter can be addressed, for example, by unsupervised feature selection, which is a very recent topic of research (e.g. Dy and Brodley, 2004).

It must be emphasized that in seismology, particularly in real-time monitoring, unsupervised learning is no replacement for supervised learning. In fact, clustering and unsupervised feature selection can be regarded as initial learning steps for investigating seismic wavefield recordings in order to understand the nature of wavefields. Even when hard clustering is not possible due to noisy data or continuously changing signal properties, the insights gained will aid the development of automatic classification systems.

While Bardainne et al. (2006) focused on the clustering of preselected sets of seismic events, our objective is to investigate the entire time series of any kind of seismic recording by considering short-time representatives of the seismogram. For

* Corresponding author. Tel.: +49 331 977 5789; fax: +49 331 977 5700.
E-mail address: akoehler@uni-potsdam.de (A. Köhler).

0098-3004/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2009.02.004

this purpose, we use different feature generation methods, which are common in seismology. From this large set of potentially suitable attributes, a smaller, optimal subset, including discriminative and significant features, is automatically selected and combined in a single feature vector.

Since the distribution and grouping of vectors cannot be easily visualized for dimensions higher than three, a visual assessment of the clustering is difficult. Furthermore, shape and density of clusters can differ greatly and the overall number or even the existence of clusters is not known in advance. Thus, quantitative cluster validity measures (Halkidi et al., 2002) may fail to predict the quality and, thus, lead to wrong conclusions. In order to overcome this problem, several methods have been proposed to reduce the dimensionality of the data such as the well known Karhunen–Loève Transform or principal component analysis (PCA). In this study, we use the Self-Organizing Map (SOM) algorithm (Kohonen, 2001), which is a convenient and widespread unsupervised learning method. Especially for large data sets of high dimensions, SOMs allow for an intuitive visualization of the data by vector quantization and allow for ordered mapping into a lower, mostly two-dimensional, space. Furthermore, SOMs can be used to assess intuitively correlations between features (Vesanto and Ahola, 1999). SOMs have already been applied, in different contexts, in seismology (Maurer et al., 1992; Musil and Plešinger, 1996; Tarvainen, 1999; Plešinger et al., 2000; Esposito et al., 2008) and for active seismic data sets (Essenreiter et al., 2001; De Matos et al., 2007).

In this paper, we present a SOM-based feature selection and pattern discovery approach for seismic wavefield recordings. In Section 2 we give a more detailed introduction into SOMs. Section 3 introduces the data-adaptive parametrization of seismic recordings, i.e. the generation and automatic selection of suitable features. We demonstrate and assess the reliability of our approach using synthetic waveforms in Section 4 and real-world data in Section 5.

2. Self-Organizing Maps

In Fig. 1 the SOM technique is demonstrated by means of a simple three-dimensional data set $\vec{x} = (X, Y, Z)$, which consists of three clearly separated clusters (red symbols in Fig. 1a). Usually, SOMs are built on a regular grid (Fig. 1b) whose size and aspect ratio depends on the amount of data N and the corresponding ratio of the two largest eigen-values. Each grid unit n is represented by a prototype vector $\vec{m}_n$ (black symbols in Fig. 1a). For each data sample $\vec{x}_i$ $(i = 1, \ldots, N)$ the closest prototype vector $\vec{m}_c$ can be found, where c is called the best matching unit (BMU). At each learning step t, the prototype vectors in the neighborhood of unit c are moved towards the selected vector $\vec{x}_t$:

$$\vec{m}_n(t+1) = \vec{m}_n(t) + \alpha(t) h_{cn}(t)(\vec{x}_t - \vec{m}_n(t)), \tag{1}$$

where $h_{cn}(t)$ defines the neighborhood around unit c and $\alpha(t)$ is the learning rate, both decreasing with time. The set of prototype vectors obtained after training can be regarded as a vector quantization of the data.

The SOM can be used to visualize high-dimensional data sets since it preserves the topology of the input data, i.e. prototype vectors of neighborhood SOM units are also close neighbors in the data space. One standard visualization method is the distance matrix (U-matrix), where a color scale represents prototype vector distances of directly connected units. In Fig. 1c warm colors (red)

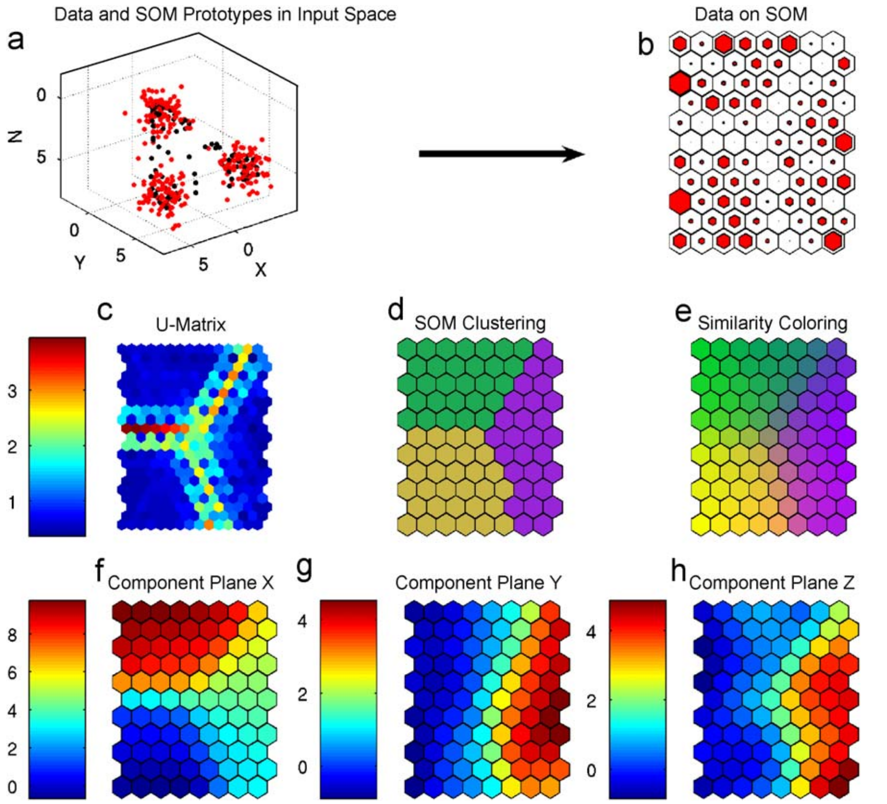


Fig. 1. Illustration of an application of Self-Organizing Maps to a simple three-dimensional data set. (a) Data and SOM prototypes in input space. Red symbols correspond to data. Black symbols correspond to prototypes. (b) Data on SOM. Red hexagons show data. Size symbolizes frequency a SOM unit (prototype) is a best matching unit (BMU) of a data sample. (c)–(e) Several SOM visualizations are shown. (c) U-Matrix, (d) SOM Clustering, and (e) Similarity Coloring. (f)–(h) Component Plane for each feature. (f) Component Plane X, (g) Component Plane Y, and (h) Component Plane Z. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mean large distances (low data density), and cold colors (blue) correspond to small distances (high data density). The U-matrix visualization therefore allows the identification and manual definition of clusters.

However, for many applications, automatic clustering is required since manual grouping is often subjective and inaccurate. As each SOM prototype vector itself can be regarded as a cluster centroid, the clustering algorithms can directly be applied to the set of all prototype vectors (Vesanto and Alhoniemi, 2000). In this study, we use a hierarchical (average linkage) clustering approach (Vesanto et al., 2000). Besides reduced computation time compared to direct clustering, the advantage of SOM clustering is that the obtained grouping can be visualized directly on the SOM by coloring each SOM unit according to its cluster-membership (Fig. 1d). By comparison with the U-matrix, this allows for a fast and simple visual assessment of the clustering validity. In order to find the number of clusters, we run the algorithm for different numbers k and choose the clustering with the lowest Davies–Bouldin clustering validity index, which has low values for high intra-cluster and low inter-cluster similarity (DB index, Davies and Bouldin, 1979; Vesanto and Alhoniemi, 2000). For our data example the number of clusters is correctly found to be three (k-search between 2 and 15).

However, automatic clustering and selection of k is not always reasonable, e.g. when the U-matrix suggests no clear grouping on the SOM. In that case, a similarity coloring can be generated by spreading again a color scale from warm to cold colors on top of the SOM, so that SOM units of close prototype vectors have similar colors (Fig. 1e). In particular, the corresponding prototype vector projection, given by the principal component axes of the two largest eigenvalues of all vectors, is used to generate the color scale (Vesanto et al., 2000). For the SOM clustering color scale (Fig. 1d), we simply average the colors (RGB values) of the SOM similarity coloring within each cluster. Both SOM colorings can be used to label the data and, thus, to characterize a time series or highlight patterns in the data.

Often not all vector components are relevant due to redundancy in the data space. In order to reduce redundancy, which is known as correlation hunting, the component planes (CPs) of a SOM can be used. A CP ($\vec{cp}$) is built on the trained SOM, where each unit n is represented by a particular component i of the corresponding prototype vector $\vec{m}_n$, i.e. for component i the CP is defined as $cp_n = m_{in}$. As proposed by Vesanto and Ahola (1999), the absolute correlation matrix between all CPs can be used as input data for the training of a second SOM on a rectangular grid. This so-called component plane SOM (CP-SOM) intuitively visualizes similarity between components. Correlated features can be grouped by clustering the CP-SOM. In this study, we apply a distance matrix-based clustering to the CP-SOM prototypes (Vesanto and Sulkava, 2002; Barreto and Pérez-Uribe, 2007; Köhler et al., 2008). Fig. 1f–h shows the CPs for X, Y, and Z. Since CPs for Y and Z are very similar, and therefore strongly correlated, CP-SOM clustering implies that only two features are necessary to recognize the clusters in the original data space. Further processing details for the SOM generation, clustering, and coloring are given in the documentation of the MATLAB® SOM toolbox of Vesanto et al. (2000).

3. Seismic wavefield parametrization

3.1. Feature generation

For visual interpretation of seismic wavefield recordings, but also for automatic classification algorithms, considering the three-component amplitudes of the seismogram alone is not sufficient. For earthquake recordings, the onset as well as the type of the event is easily identified considering a few characteristic parameters of the seismogram, such as the time–frequency amplitudes (e.g. Joswig, 1990). However, more complicated situations (noise, multiple signals) may require additional features. For instance, for P wave onset detection, spectral amplitudes are often sufficient, while for S wave detection, additional polarization information may be required. For our unsupervised approach, any information contained in the wavefield should be utilized since each may have the potential to discriminate between a priori unknown signals and to highlight patterns in the data.

Various definitions of seismic features obtained from standard analysis methods can be found in the literature of the last 30 years (see Table 1). The manual combination of features from different parametrization methods has already been proposed in the context of supervised classification of seismic events (Jepsen and Kennett, 1990; Wang and Teng, 1997; Bai and Kennett, 2000; Ohrnberger, 2001). Since our approach should be unsupervised using only a minimum of domain knowledge, we collect features from all proposed method and address automatic selection in the next section. Note that we do not use features which are directly dependent on amplitudes in our study, since they may differ strongly for a particular type of signal or seismic phase. We summarize all parametrization methods and the corresponding features, including information about polarization, coherency and the frequency spectrum, in Tables 1 and 2. All features are calculated for short three-component time windows with length T depending on the considered frequency band:

$$T = WINFAC \cdot 1/f_{cent}. \tag{2}$$

Here, f_{cent} is the center frequency of the overall frequency band and $WINFAC$ a parameter. We use $WINFAC$ values from 4 to 6 in our study, which was found to be optimal for seismic phase classification (Köhler et al., 2008).

Each time window is used to compute the f–k spectrum (Method 1 in Table 1), the averaged, complex autocorrelation coefficients (3c-MSPAC, Method 2), the 3c-correlation or covariance matrix (Method 3), and the frequency spectrum (Methods 5 and 6). Instantaneous seismic attributes (Method 4) and amplitude ratios (Method 7) are defined for each seismogram sample.

Table 1
Name of method, corresponding references, and number of features for each feature generation approach.

1 Frequency-wavenumber analysis: nine features Kvaerna and Ringdahl (1986)
2 Spatial averaged autocorrelation method: 18 features Aki (1957), Asten (2006), Köhler et al. (2007)
3 Eigenvalues of complex 3c-covariance matrix: 39 features Samson and Olson (1981)[1], Vidale (1986)[2], Park et al. (1987), Jurkevics (1988)[3], Hearn and Hendrick (1999)[4], Bai and Kennett (2000)[5] and Reading et al. (2001)[6]
4 Complex seismic trace analysis: 42 features Taner et al. (1979), René et al. (1986)[7] Morozov and Smithson (1996)[8], Bai and Kennett (2000)[9] and Schimmel and Gallart (2004)[10]
5 Spectral attributes: 25 features Joswig (1990) and Ohrnberger (2001)
6 Spectrum of polarization ellipsoid: 20 features Pinnegar (2006)
7 Amplitude ratios: nine features After Jepsen and Kennett (1990)

Table 2
Features and short names for each parametrization method.

Method	Feature description	Short name[a]
1	Semblance: vertical, radial and tangential comp.	pr
2	Real and imaginary (absolute value), frequency band averaged autocorrelation coefficients: vertical, radial and tangential comp.	spac, spacim
3	Degree of polarization[1]	dopII
3	Ellipticity, strength of polarization, angle of incidence, planarity[2]	ell, sop, inc, plan
3	Linearity, planarity[3]	rect, planII
3	Linearity (2×), stability of direction cosine, enhanced linearity[4]	linII, linIII, sdc, elin
3	Enhanced linear polarization[5]	elip
3	Degree of polarization[6]	dopIII
4	Instantaneous frequency and variance: vertical and horizontal comp.	if, vif
4	Phase difference, ellipticity and tilt between vertical and horizontal components[7]	pdiff, ell, tilt
4	Variance of azimuth, ellipticity[8]	vazi, 3cell
4	Component averaged instantaneous frequency[9]	FQ1
4	Degree of polarization, linearity[10]	dop, lin
5	Normalized horizontal and vertical sonogram	sono
5	Dominant spectral frequency and bandwidth: vertical and horizontal comp.	domf, bb
5	Logarithm of ratio between sum of lower and higher sonogram bands	ratiolf
6	Normalized semi-major minus semi-minor axis and semi-minor axis of polarization ellipsoid	a_b, b
7	Real over imaginary part of complex trace, horizontal over vertical and east component	P/Q, H/V, H/E

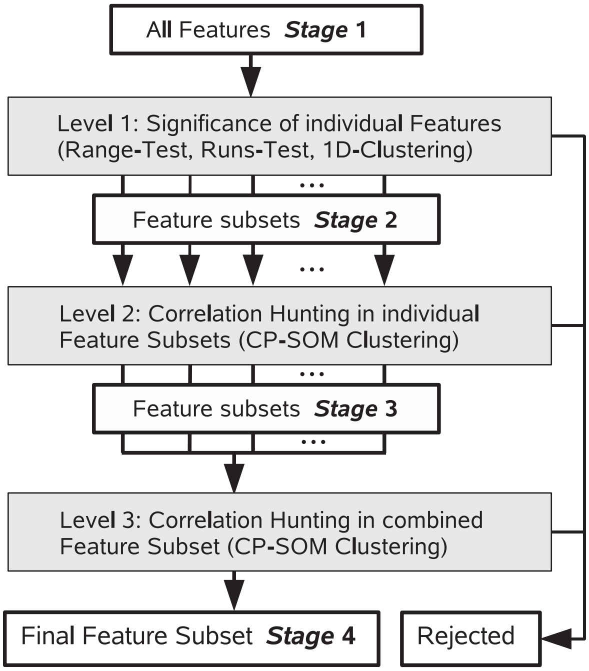
Superscript numbers refer to method list in Table 1.

[a] Suffixes for short names: component: z (vertical), e (east), n (north), h (horizontal), r (radial), t (tangential). Frequency band index: $1, 2, \ldots, 3, (\ldots, 10)$.

For our parametrization they are averaged over the complete time window. If array-network recordings are available, we generate a single feature vector by averaging the single-station attributes over all receivers for the purpose of noise reduction (Jurkevics, 1988; Ohrnberger, 2001). Depending on the network aperture, sub-array smoothing may be required whenever the expected travel time tt between stations is too large compared to the time window length ($tt > 0.5 \cdot T$). Except for Methods 5 and 6, where we choose 10 bands, all features are computed for three different frequency bands. The lowermost and uppermost frequency limits are chosen according to the frequency content of the data. For simplicity, the frequency band index and the component suffix are omitted for the short feature names in Table 2.

3.2. Unsupervised feature selection

Taking into account all parametrization methods and all frequency bands, we calculate a set of about 160 different features. However, relevance of individual features for pattern discovery may vary considerably. Furthermore, we expect strong redundancy between vector components since of course not all features represent different wavefield properties. As a consequence, computation time or occupied disk space may be unnecessarily increased and the quality of the final results may suffer from having useless features. Additionally, interpretation and further analysis is much more intuitive for low number of features.


All Features Stage 1
Level 1: Significance of individual Features (Range-Test, Runs-Test, 1D-Clustering)
Feature subsets Stage 2
Level 2: Correlation Hunting in individual Feature Subsets (CP-SOM Clustering)
Feature subsets Stage 3
Level 3: Correlation Hunting in combined Feature Subset (CP-SOM Clustering)
Final Feature Subset Stage 4
Rejected

Scheme 1. Three-level feature selection procedure (Köhler et al., 2008). Stages 1–4 correspond to different feature subsets obtained at particular processing step. Feature subsets at Stages 2 and 3 correspond to different parametrization methods.

For dimensionality reduction methods such as PCA, characterization of the reduced data space is difficult since the (physical) meaning of the new features, e.g. generated by linear combinations, is unclear. Unsupervised feature selection techniques, which do not transform features, can be divided into wrapper (e.g. Dy and Brodley, 2004), filter (e.g. Li et al., 2007), and embedded methods. The computational cost is very high for wrappers, since a subset search is made for all combination of features and the chosen learning method is applied and evaluated for each subset. Due to the high number of features in our analysis, and because we also want to keep features that might show no clear cluster tendency but significant patterns in their time history, we proposed a new three-level filter approach in Köhler et al. (2008), which iteratively reduces the number of features. The processing flow of our feature selection procedure is illustrated in Scheme 1.

In the first level (relevancy filter), we chose potential feature candidates by assessing the information content of each feature individually. In particular, we compare the observed variability $R_{obs} = \max(F) - \min(F)$ of each feature F with the reasonably expected range R_{exp} derived from physical or data processing parameters. For instance for the ellipticity, the natural limits (0 and 1) imply $R_{exp} = 1$, and for the instantaneous frequency the range of the frequencies f gives $R_{exp} = f_{max} - f_{min}$. We exclude those features providing no significant discrimination between time windows due to small observed ranges ($R_{obs}/R_{exp} < 0.1$). The relevance of a feature, that a temporal context can be represented, is addressed by the Wald–Wolfowitz (1940) runs test, which can be used to assess the randomness of a time series. We evaluate the corresponding test statistic Z_{test} and exclude a feature whenever the hypothesis of randomness is not rejected ($Z_{test} < 1.96$ for a significance level of 5%). For instance, for a completely random time series (Gaussian), we would obtain $Z_{test} = 0.35$.

In the second and third level (redundancy filter), we apply the CP-SOM clustering on all features accepted by Level 1. First, only the features of each parametrization method (see Table 1) are processed together (Level 2 in Scheme 1). The components having the lowest DB index and the highest test statistic Z_{test} are chosen as representative features from each CP-SOM cluster. Finally, in Level 3, again a CP-SOM is generated and clustered for the remaining features. The final set of features is then used as a parametrization vector for all time window. For more algorithm details including parameters and performance tests see Köhler et al. (2008).

4. Application to synthetic data

In order to asses the reliability of the methodology and to show its potentials, we generate 30 min of synthetic data with a sampling rate of 50 Hz, employing the mode-summation technique of Hermann (2002). The wavefield is computed at 12 receivers, forming a station array of 400 m aperture. We use a simple source setting on the earth surface that excites clearly separated wave packages of Rayleigh waves (vertical point force), Love waves (tangential force with respect to the array midpoint), and mixture waves (vertical and horizontal force components) for different azimuths and distances (between 1.2 and 3 km). The seismic velocity model used for wavefield computation is a deep basin structure given in Table 3 (see Köhler et al., 2007, for details). Finally, we add white noise to the waveforms. The time window length for further processing is 3.7 s ($WINFAC = 5$, $f_{cent} = 1.34$ Hz, see Eq. (2)), and the frequency bands are located between 0.26 and 15 Hz.

Table 3
Subsurface velocity model used for generation of synthetic data.

Thickness (m)	Vp (m/s)	Vs (m/s)	Density (g/cm^3)
35	542	209	1.762
178	851–1131[a]	393–522[a]	1.736–1.85[a]
80	3085	1619	2.202
80	3525	1850	2.308
80	3913	2054	2.394
80	4265	2238	2.467
80	4588	2408	2.531
80	4888	2566	2.588
80	5170	2714	2.64
49	5387	2828	2.678
Halfspace	5916	3416	2.782

[a] Gradient.

4.1. Cross-validation of the feature selection procedure

We assess the validity of our feature selection procedure by applying a cross-validation technique (Dy and Brodley, 2004). For this purpose, it is necessary to label the time windows according to the theoretical source setting. We define four classes (see Fig. 5): pure Rayleigh waves (class 1), pure Love waves (class 2), mixture of Rayleigh and Love waves (class 3), and random noise (class 4). The onset and duration of signals after excitation is defined from a previous evaluated seismogram section for different source receiver distances. Note that the labeled time windows are only used to assess the performance of the procedure after the training phase.

The data set is divided into 10 sections of same length, after randomly permuting the order of time windows. While nine sections are used as training data for SOM learning and clustering, including feature selection (Levels 2 and 3), the remaining section serves as a test or validation data set. Level 1 of the feature selection procedure is applied on the complete, not permuted data set, in order to keep the temporal context for the runs test. The selection of the test data set is repeated 10 times, so that each section is used for validation once. Each cluster obtained from the training data is classified with respect to the most frequent theoretical class label within, which is given by projecting the ground-truth class labels of the training data on the SOM. For the testing, we compute the BMUs, and thus the cluster memberships, of the test data set on the training data set SOM. A class error is computed for each fold as the percentage of misclassified data (cluster label does not match with theoretical time window label) with respect to the total number of samples of the test data set. The final classification error is presented as mean and standard deviation of the individual fold errors (CVCE: cross-validated class error, Dy and Brodley, 2004).

In order to show the relationship between number and types of selected features, the CVCEs are estimated for several feature

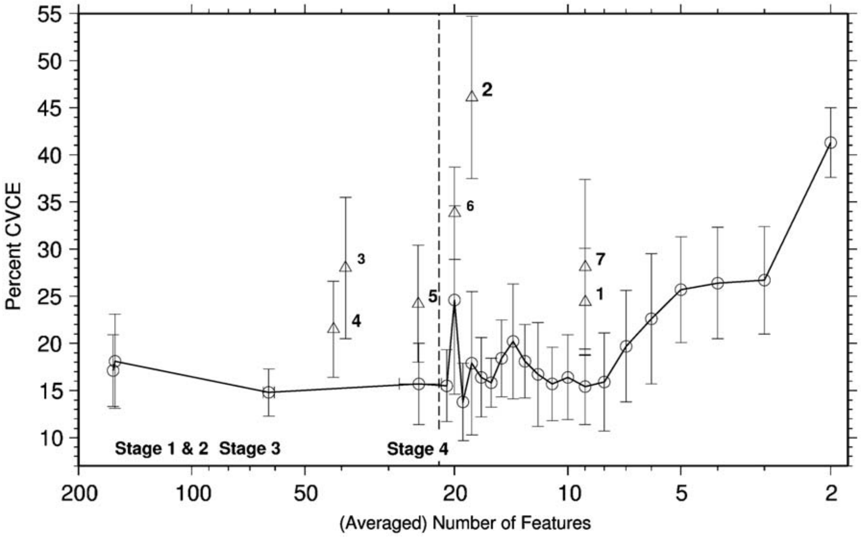


Fig. 2. Results of cross-validation for synthetic data set. Cross-validated class error (CVCE) and standard deviation are plotted over averaged number of selected features for different feature subsets. Curve represents feature selection for features from all generation methods at different stages (see Scheme 1). Furthermore, final feature set at Stage 4 is reduced stepwise (right-hand side of vertical dashed line). Triangles correspond to CVCEs of different parametrization methods. Numbers aside represent method indices used in Table 1.

subsets obtained at different stages of feature selection (see Scheme 1), and for particular feature generation methods (see Table 1). The results for the CVCE, standard deviation, and (averaged) number of features are summarized in Fig. 2. Furthermore, we show results for a stepwise reduction of the final feature set obtained at Stage 4 (right-hand side of vertical dashed line in Fig. 2). For each step the feature with the lowest temporal significance Z_{test} is omitted.

Taking into account the standard deviations, the CVCE curve shows similar classification performance for decreasing number of features down to a number of about 15–10 components. Therefore, this feature set would be the most favorable one for the given data set, since it provides the highest classification accuracy with the lowest possible number of features. Reducing the feature set further, clearly starts to increase the errors. Hence, our automatically selected feature set is slightly larger than necessary (25 features). However, note that we perform an unsupervised procedure. Therefore, this discrepancy is not unexpected. Furthermore, compared to the overall number of features candidates (162), it is acceptable that we still have some useless features within our final set. The most favorable performance for Stages 1–4 is achieved with about 57 features from all methods at Stage 3 (CVCE = 15.8%). However, after assessing correlation between all parametrization methods at Stage 4, the CVCE is still within the range of standard deviations of Stages 1–3.

Furthermore, comparing the results on the curve, where features from all parametrization approaches are combined, and the triangles, which stand for the individual feature generation methods, the CVCEs are significantly lower on the curve for all stages. Considering the individual parametrization methods alone, the spectral features (Methods 4 and 5 from Table 2) seem to provide the best discriminative power for clustering.

From the cross-validation we conclude that it is sufficient to consider only the final feature subset, combining features from different methods (Stage 4). The dimensionality, and therefore computation time and model complexity, is reduced considerably for further analysis of the data set, without significantly losing discriminative power for clustering.

4.2. Application of SOM-based unsupervised analysis

In this section we apply feature selection on the complete synthetic data set, resulting in a subset of 21 variables. Subsequently, a SOM is trained and clustered. The U-matrix visualization, the clustering and the SOM coloring are shown in Fig. 3. The corresponding SOM component planes are plotted in Fig. 4 together with the short name of each feature (see Table 2). The background colors of the seismograms in Fig. 5 represent the labeled time windows. The color of each time segment corresponds to the color of its BMU on the SOM representation in Fig. 3b (SOM clustering). Furthermore, Fig. 5 shows the theoretical occurrence of classes 1–3 on top of the seismograms. The sizes of the symbols correspond to the receiver amplitudes.

Since we present an unsupervised approach, let us first look at the results without preconceptions, pretending that we do not know the source distribution and the existing wave types. The U-matrix visualization (Fig. 3a) suggests a clear natural clustering. Furthermore, the lowermost part of the SOM shows an uniformly high data density without significant pattern. The remaining SOM appears to be more structured, i.e. several areas with high data density are limited by regions of lower density. However, it is difficult to define the exact number of clusters due to the lack of distinct cluster borders. Counting the SOM areas with high data densities, we obtain a number of about 5–10 groups.

The automatic clustering confirms our observations (compare Fig. 3a and b). The lower part of the SOM is assigned to a single cluster (violet), whereas the remaining clusters more or less fit the structure of the U-matrix. Thus, the clustering provides a first order grouping and a meaningful representation of the dominant patterns of the data (see Fig. 5).

An alternative to hard clustering is the SOM coloring technique introduced in Section 2. The colors of SOM units in Fig. 3c allow to identify and assess the transitions between clusters. This is very useful since the boundaries are not necessarily well-defined. See, for example, the more continuous transition between the violet clusters (bottom right SOM).

Let us now introduce some of our expert knowledge for the interpretation, i.e. that we expect dominant surface waves and added noise. Considering the background colors of time windows without signals, random noise is obviously represented by the bottom part of the SOM. The high density indicates a homogeneous and compact class. Focusing on the dependencies between all three spatial components for a single time window, we can derive that the green colors or the top left clusters, respectively, represent dominantly Love waves, since there is no signal amplitude on the vertical component. The other signals have a significant contribution of Rayleigh waves.

The CPs (Fig. 4) allow to investigate the signal properties. Considering e.g. degree of polarization (dopIII), higher values are found for Love waves. Furthermore, values on CPs for ellipticity (ellez, 3cell), coherency on vertical component (prz), and semiminor axis of polarization ellipsoid (b), which are all theoretically zero for Love waves and vary for Rayleigh waves and noise, are consistent with our interpretation.

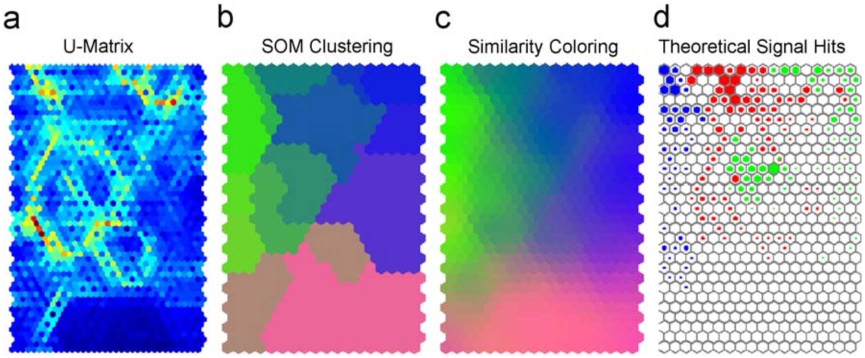


Fig. 3. SOM visualizations for synthetic data set. (a) U-matrix: warm colors mean high and cold colors mean low distances. (b) SOM Clustering and (c) Similarity Coloring: Color scales are generated automatically based on similarity of cluster means (b) or prototype vectors of each SOM unit (c). (d) Theoretical Signal Hits: Best matching SOM units for time windows containing signals. Colors correspond to theoretical class labels: Rayleigh waves (green), Love waves (blue), and mixture of both types (red). Size corresponds to signal amplitude. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

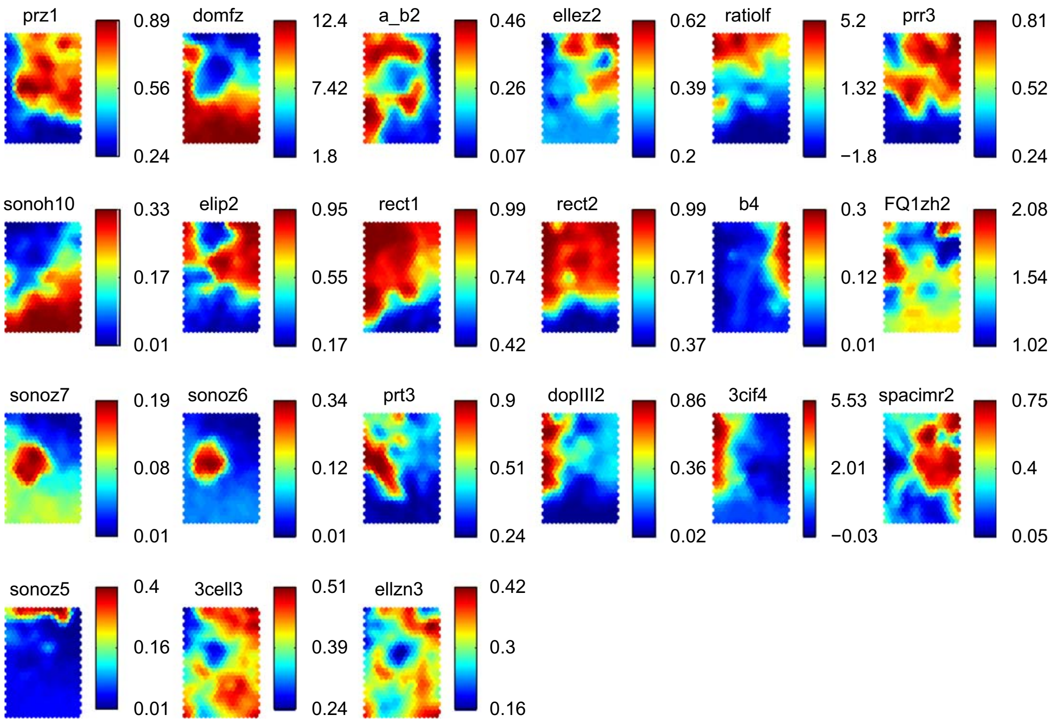


Fig. 4. SOM component planes of 21 automatically selected features for synthetic data. Color scales are plotted on right hand-side of each CP. Short name for each feature is given on top (see Table 2).

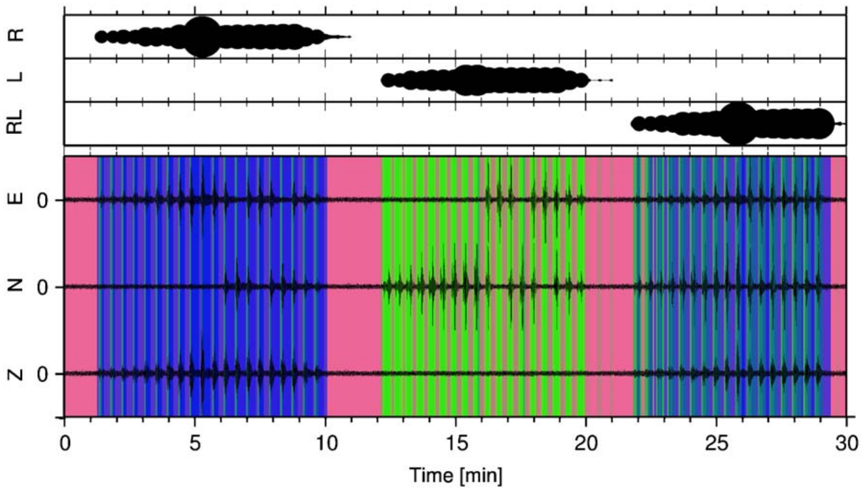


Fig. 5. Synthetic 3c-waveforms at array center station. Background colors of lower panel correspond to time window labels given by cluster membership in Fig. 3b. Uppermost panel shows theoretical occurrence of signal types for each time window. Signal amplitudes (size of circles) are given for Rayleigh waves (R), Love waves (L), and mixture of both types (RL).

In a last step, we consider additionally the employed source setting. The label projection on the SOM, i.e. finding the BMU of the labeled feature vectors, is shown in Fig. 3d. The size of the symbol corresponds to the receiver amplitude and the color to the class label (green: Rayleigh, blue: Love, red: Rayleigh and Love waves). The projection confirms our interpretation since pure Rayleigh waves hit areas on top right and pure Love dominantly on top left of the SOM. The mixture wavefield is distributed between both sides depending on the force orientation. In principle, there should be no dominating amplitude dependency since the amplitude value is not directly used for parametrization. However, a decreasing signal to noise ratio will effect the parametrization, e.g. the estimation of the covariance matrix. Thus, we observe weak signals due to geometrical spreading and small source amplitudes in between the areas of clear transients (top of SOM) and dominant noise on the lowermost part of the SOM, where no hits of signals occur. Note that the observed number of clusters in the data does not correspond to the expected number of wave-type classes (three types and noise).

5. Application to real data

The following application example shows SOM-based analysis of regional seismicity. We consider earthquakes from the European broadband network with magnitudes larger than four

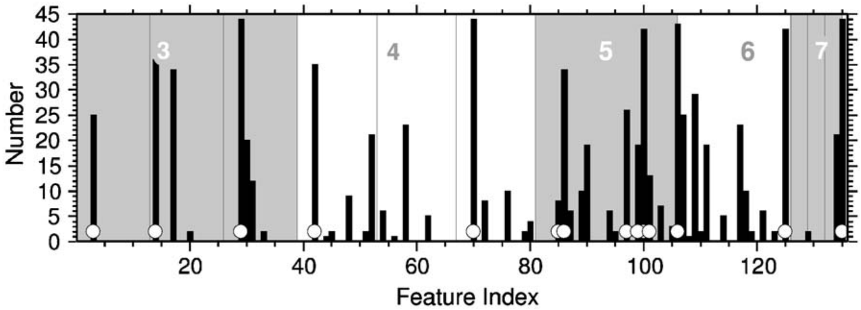


Fig. 6. Cross-validation selection statistic for each feature for earthquake data set. Background coloring distinguishes different feature generation methods (see gray and white numbers). Vertical lines denote frequency bands. White circles correspond to feature set obtained using complete data (no cross-validation).

Table 4
Features automatically selected for earthquake data set.

Feature	Z_{test}
ifh3	37.90
sonoh10	36.52
bbh	34.65
ratiolf	34.16
b9	33.35
sonoh1	27.20
sonoh4	17.45
sonoh3	17.04
H/V3	15.98
sonoh5	15.90
planII3	12.57
ifh1	6.59
planII1	5.57
rect2	5.12

Short name (see Table 2) and runs test statistic Z_{test} are given.

between 2003 and 2006. Due to a priori data selection, the investigation is not unsupervised in a strict meaning. However, even though not using the complete recordings, by making use of known earthquake source times, we still pursue an unsupervised approach. We want explore data inherent similarity properties to allow seismic phase and event discrimination without using onset times directly for learning. This approach is similar to the one of Bardainne et al. (2006) and Esposito et al. (2008). However, these authors used a single parametrization vector for each event (full time structure) and not for short time windows as done in our study.

We employ recordings lasting 6 min and starting 2 min before the P wave onset. We select 44 earthquakes for which we could identity and pick clear P and S onsets at a single station (RDO). The picks will help us to evaluate our observations after feature selection and SOM training. We compute the features for 6.5 s long time slices ($WINFAC = 4$). For each event we obtain 55 time windows. Subsequently, the individual vector time series are merged. Finally, we obtain a single data set of 2420 time slices. Since we do not use a receiver network, only feature generation Methods 3–7 are employed.

Fig. 6 summarizes the automatically selected features for both a cross-validation experiment and using directly the complete data set. The latter ones are also listed in Table 4 in detail. Cross-validation is performed on feature selection and SOM training using 44 folds, i.e. by leaving out time slices of one event each fold. Unlike cross-validation for synthetics, feature selection Level 1 is included, since we do not permute the data. The histogram in Fig. 6 presents the frequency an individual feature was selected during validation. It shows that the selection procedure is stable and robust since similar features are obtained for each fold. Furthermore, the features selected using the complete data set (white circles) are also frequently chosen during cross-validation. The most frequent and most significant features (with respect to Z_{test}) are those generated from the frequency spectrum of the wavefield (e.g. Sonogram and instantaneous frequencies). However, also polarization properties contribute to the final feature set.

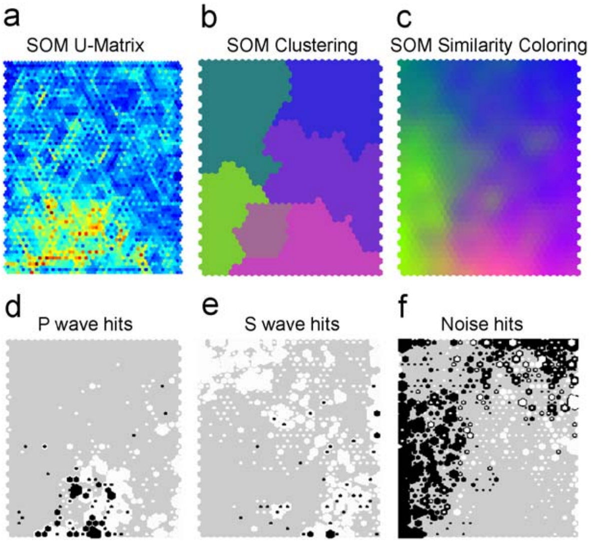


Fig. 7. SOM visualizations for earthquake data set. (a) SOM U-Matrix, (b) SOM Clustering, and (c) SOM Similarity Coloring. (d) P wave hits. Hits of P wave time windows on SOM for all events (white). Black symbols show first onset window for each event. (e) S wave hits. Same meaning as for P wave hits. (f) Noise hits. Hits of noise class (white). Black hits represent noise time windows after event.

Fig. 7 presents the SOM U-matrix (Fig. 7a), automatic SOM clustering (Fig. 7b), continuous SOM coloring (Fig. 7c), and data hits on the SOM. For the latter we consider three manually labeled classes: time slices after P wave onset, time slices after S wave onset including surface waves, and finally the windows visually defined as noise before and after the event. Furthermore, Fig. 8 shows the vertical seismograms of all 44 events using the color scale of Fig. 7b for the background. The U-matrix shows a clustered, more sparse region at the bottom. The observed clustering is less clear for the remaining SOM. Considering only the onset time windows (black hits in Fig. 7d) shows that P wave time slices are located within the clustered area. Most S wave onset windows (black hits in Fig. 7e) are also well separated from P wave and noise. However, the spread for all signal windows after onset is higher for S waves than for the P wave coda (white hits). Furthermore, there is an expected continuous transition from the S wave to the background noise class and no distinct cluster boundaries (Fig. 7f). The noise hits after the event (white) are mainly located on the right-hand area of the SOM. Thus, these

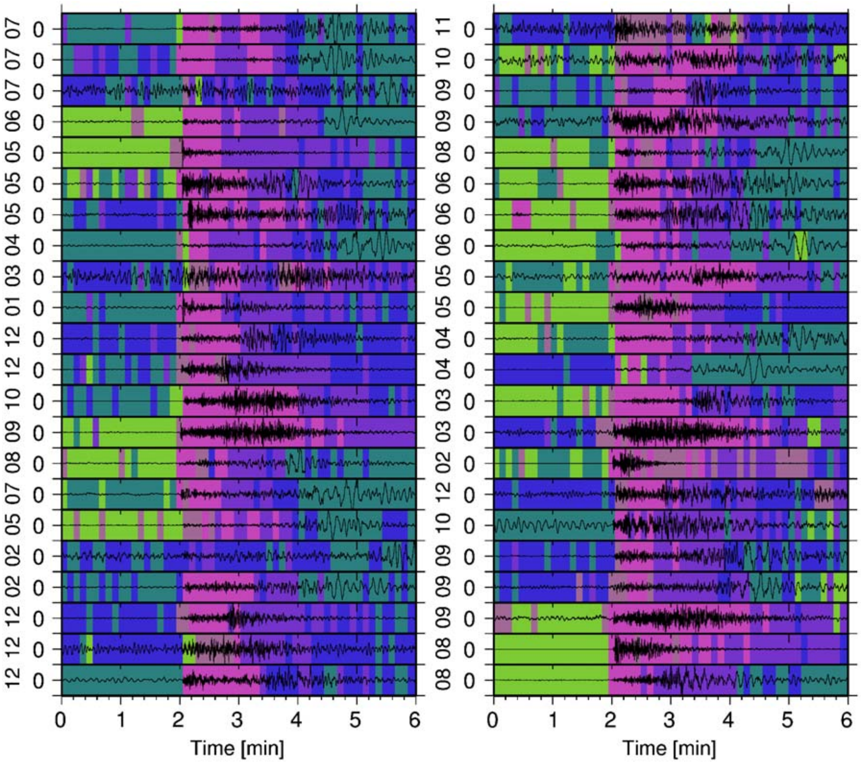


Fig. 8. Vertical component seismograms of 44 events recorded between 2003 and 2006 at station RDO. *Y*-axis label indicates month in which an event occurred. Amplitudes are normalized by maximum of each trace. SOM cluster coloring in Fig. 7b is used for background.

time windows are still affected by the event. In fact, Fig. 8 shows that differences between background noise before and after the event are highlighted for some seismograms, which are not visible from the amplitudes alone. A further observed pattern is the seasonal variation of the background wavefield before P onset. Fig. 8 shows a correlation between the autumn and winter months and the dark green and blue cluster colors. Since also surface waves with longer periods seem to belong to the same cluster (dark green), a likely reasons for that observation is the increasing power of microseismicity caused by increasing storm activity in the Mediterranean Sea during autumn and winter.

In order to quantify seismic phase discrimination on the SOM, a further cross-validation experiment is conducted for two feature sets. We employ the automatically selected feature set, which is composed of attributes from different generation methods (Table 4), and the spectral features alone (Method 5). The latter one is the best performing (see synthetics) and most common feature generation method in seismology (e.g. Joswig, 1990; Riggelsen et al., 2007). Validation is only carried out for SOM training (no clustering). Hence, classification is made based on the best matching SOM prototype vectors, which are labeled after training using the hand-picked onsets and duration of each event. We compute median false positive and false negative classification errors for the noise, P wave, and S wave class.

Although the uncertainties of the results are rather high, due to the simplification of splitting the complete records into three classes, we can derive some qualitative insights from Table 5. The highest misclassification rate is obtained for the S wave time windows, which confirms the observations from Fig. 7 (high spread). Furthermore, for the noise class, the false positive are about 10 percentage points higher than the false negative errors, for both the complete feature selection (FS) and Method 5. Most probably, there are time windows which are presented as S waves, but are classified as noise. Probably, a manual labeling is not reasonable due to the continuous transition from coda to the background wavefield, or because the coda is longer or shorter than suggested by the seismogram amplitudes, respectively. Considering P and S waves as a single class (signal and noise) has a clear tendency to improve classification results. Comparing Method 5 and the automatically selected feature set (FS), only classification errors for the S wave class, and therefore also the class-averaged rates, are higher for the spectral features. Thus, those features alone would be sufficient and suitable for event detection. However, for S wave recognition we obviously need additional polarization information to improve phase discrimination.

Table 5
Cross-validation results for earthquake data set.

Class	FS		Method 5	
	Percent CVCE FP	Percent CVCE FN	Percent CVCE FP	Percent CVCE FN
P	28.6 ± 19.9	32.1 ± 19.3	32.1 ± 23.5	31.2 ± 21.3
S	41.0 ± 20.9	43.2 ± 15.8	55.1 ± 19.4	68.8 ± 14.1
Noise	30.8 ± 16.2	20.1 ± 13.2	34.5 ± 19.3	17.3 ± 10.6
Signal	25.4 ± 16.8	19.2 ± 16.0	23.6 ± 16.2	33.3 ± 18.3

False positive (FP) and negative (FN) classification errors are given for each class and for two feature subsets. A set obtained by feature selection (FS) and features from generation Method 5 (spectral features) are used. Signal class contains P and S wave class.

Employing the same data base, Riggelsen et al. (2007) tested dynamic Bayesian networks, an advanced supervised and context-dependent learning technique, as a signal detection technique (two classes: P and non-P wave). Using 50 s long time windows for each event (25 s before and 25 s after P wave onset), they obtained an accuracy from cross-validation about 0.95 ($CE = 5\%$) on average and 1.0 ($CE = 0\%$) for station RDO. We conduct a similar experiment using the same data length for each event. In spite of a simple vector quantization and nearest neighbor classification without using context-dependent information, we observe similar results for median classification errors. For more than 50% of the events (29), $CE = 0\%$ is obtained. Therefore, the median is zero and mean deviation from median (MD) is 10.7%. Nevertheless, since classification errors and variances are clearly higher for the three-class and complete-event problem (see Table 5), our results do not imply that no context-dependency is required for discrimination of more classes (P and S waves) and longer records.

6. Conclusions

In this work we introduced an unsupervised pattern discovery approach for seismic wavefield recordings. For this purpose, we implemented several common parametrization methods, generating short-time representatives (features) for various wavefield properties. Features have been computed from continuous three-component (array) seismograms for time windows whose lengths are specified according to the temporal patterns of interest. As an unsupervised learning and intuitive two-dimensional visualization method, we have successfully employed the multi-purpose and easily applicable Self-Organizing Map method. An unsupervised three-level feature selection procedure has been suggested. The underlying concept consists in the combination of a relevancy and redundancy filter realized by significance testing for temporal randomness (Wald–Wolfowitz runs test) and correlation hunting using SOMs. Using the automatically selected feature set, training and clustering of the SOM is performed as a last step, in order to group existing patterns. Our approach has been developed for continuous wavefield recordings. However, as a semi-supervised method, it can also be applied to preselected data sections which include, e.g., individual earthquake records. In contrast to previous unsupervised investigations on seismic data, we did not compute a feature vector for the complete event, but also successively divided these records into short time windows.

We have applied our processing scheme to synthetics, composed of Rayleigh and Love waves, and real recordings of regional earthquakes. We showed the robustness of our feature selection method. It is found that a lower model complexity is produced over a decreased feature set whose size (about 20) was comparable with the smallest number required for the best achievable classification accuracy. The final feature set has been a combination of features from different generation approaches. An improvement has been obtained, e.g. for surface wave and S wave discrimination, compared to using only the common and most suitable seismological approach, which we found has been the frequency spectrum.

The final SOM itself facilitated intuitive imaging of patterns in the seismic wavefields. It was found that a meaningful number of clusters does not necessarily have to fit with the expected number of signal classes. This observation confirms the need of unsupervised learning for data inspection, e.g. before more advanced supervised learning can be carried out.

For the earthquake record, using preselected time windows including the background wavefield before and after the onset, we have shown that the natural seismic phase discrimination can be intuitively visualized by projection of the manually picked seismic phase labels on the SOM. Good classification rates or discrimination have been found for P wave onset time windows for arbitrary earthquakes. On the other hand, it turned out to be more difficult to define distinct S wave and background noise classes for all events. However, improvements are expected for supervised S wave classification by including the context-dependency of feature vectors.

Our approach can be easily applied to other problems in seismology. Particular applications would be data quality control or mapping of long-term variations in the background wavefield for large data sets.

Acknowledgments

We want to thank Carsten Riggelsen and the reviewers for their comments that helped to improve this manuscript. The study is funded by a PhD-scholarship of the University of Potsdam. Part of this work has been made possible under the EU-Project NERIES Contract no. 026130.

References

Aki, K., 1957. Space and time spectra of stationary stochastic waves, with special reference to microtremors. Bulletin of the Earthquake Research Institute, University of 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. Bulletin of the Seismological Society of America 90 (1), 187–198.

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

Barreto, M., Pérez-Uribe, A., 2007. Improving the correlation hunting in a large quantity of SOM component planes. In: Lecture Notes in Computer Science, vol. 4669, pp. 379–388.

Dai, H., MacBeth, C., 1995. Automatic picking of seismic arrivals in local earthquake data using an artificial neural network. Geophysical Journal International 120 (3), 758–774.

Davies, D., Bouldin, D., 1979. A cluster separation measure. Institute of Electronics and Electrical Engineers Transactions on Pattern Analysis and Machine Intelligence 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.

Dy, J., Brodley, C., 2004. Feature selection for unsupervised learning. The Journal of Machine Learning Research 5, 845–889.

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. Bulletin of the Seismological Society of America 98 (5), 2449–2459.

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

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2002. Cluster validity methods: part I and II. Special Interest Group on Management Of Data Record 31 (2), 40–45.

Hearn, S., Hendrick, N., 1999. A review of single-station time-domain polarisation analysis techniques. Journal of Seismic Exploration 8, 181–202.

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

Jepsen, D., Kennett, B., 1990. Three-component analysis of regional seismograms. Bulletin of the Seismological Society of America 80 (6 B), 2032–2052.

Joswig, M., 1990. Pattern recognition for earthquake detection. Bulletin of the Seismological Society of America 80 (1), 170–186.

Jurkevics, A., 1988. Polarization analysis of three-component array data. Bulletin of the Seismological Society of America 78 (5), 1725–1743.

Köhler, A., Ohrnberger, M., Riggelsen, C., Scherbaum, F., 2008. Unsupervised feature selection for pattern search in seismic time series. Journal of Machine Learning Research. In: Workshop and Conference Proceedings: New Challenges for Feature Selection in Data Mining and Knowledge Discovery, vol. 4, pp. 106–121.

Köhler, A., Ohrnberger, M., Scherbaum, F., Wathelet, M., Cornou, C., 2007. Assessing the reliability of the modified three-component spatial autocorrelation technique. Geophysical Journal International 168 (2), 779–796.

Kohonen, T., 2001. Self-Organizing Maps. In: Springer Series in Information Sciences, vol. 30, third extended ed. Springer, Berlin, Heidelberg, New York, 501pp. (1995, 1997).

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

Li, Y., Lu, B., Wu, Z., 2007. Hierarchical fuzzy filter method for unsupervised feature selection. Journal of Intelligent and Fuzzy Systems 18 (2), 157–169.

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.

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. Bulletin of the Seismological Society of America 86 (4), 1077–1090.

Ohrnberger, M., 2001. Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia. Ph.D. Dissertation, University of Potsdam, ⟨http://opus.kobv.de/ubp/volltexte/2005/31/pdf/ohrnberg.pdf⟩ [accessed 31 March 2009], 158pp.

Park, J., Vernon III, F., Lindberg, C., 1987. Frequency dependent polarization analysis of high-frequency seismograms. Journal of Geophysical Research 92 (B12), 12664–12674.

Pinnegar, C., 2006. Polarization analysis and polarization filtering of three-component signals with the time-frequency S transform. Geophysical Journal International 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. Geophysical Journal International 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. In: Lecture Notes in Computer Science, vol. 4702, pp. 565–572.

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

Schimmel, M., Gallart, J., 2004. Degree of polarization filter for frequency-dependent signal enhancement through noise suppression. Bulletin of the Seismological Society of America 94 (3), 1016–1035.

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. Physics of the Earth and Planetary Interiors 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. Academic Press, Rochester, NY, pp. 279–285.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. Institute of Electronics and Electrical Engineers Transactions on Neural Networks 11 (3), 586–600.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 2000. SOM Toolbox for Matlab 5. Helsinki University of Technology, 52pp.

Vesanto, J., Sulkava, M., 2002. Distance matrix based clustering of the self-organizing map. In: Proceedings of the International Conference on Artificial Neural Networks 2002, Madrid, Spain, pp. 951–956.

Vidale, J., 1986. Complex polarization analysis of particle motion. Bulletin of the Seismological Society of America 76 (5), 1393–1405.

Wald, A., Wolfowitz, J., 1940. On a test whether two samples are from the same population. The Annals of Mathematical Statistics 11 (2), 147–162.

Wang, J., Teng, T., 1997. Identification and picking of S phase using an artificial neural network. Bulletin of the Seismological Society of America 87 (5), 1140–1149.

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. Bulletin of the Seismological Society of America 88 (1), 95–106.