Institut für Geowissenschaften, Universität Potsdam

Recognition and Investigation of Temporal Patterns in Seismic Wavefields Using Unsupervised Learning Techniques

Dissertation zur Erlangung des akademischen Grades "doctor rerum naturalium" (Dr. rer. nat.) in der Wissenschaftsdisziplin Geophysik

eingereicht an der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Potsdam

von
Andreas Köhler
geboren am 08.04.1980 in Magdeburg

Potsdam, im Januar 2009

Published online at the
Institutional Repository of the University of Potsdam:
http://opus.kobv.de/ubp/volltexte/2009/2970/
urn:nbn:de:kobv:517-opus-29702
[http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-29702]

Als Dissertation genehmigt von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Potsdam

Tag der mündlichen Prüfung: 30.April 2009

Vorsitzender der Prüfungskommission:

Professor Dr. R. Bousquet
Institut für Geowissenschaften
Universität Potsdam

Erstberichterstatter:

Professor Dr. F. Scherbaum
Institut für Geowissenschaften
Universität Potsdam

Zweitberichterstatter:

Dr. M. Ohrnberger
Institut für Geowissenschaften
Universität Potsdam

Drittberichterstatter:

Professor Dr. M. Joswig
Institut für Geophysik
Universität Stuttgart

"If the whole universe has no meaning, we should never have found out that it has no meaning: just as, if there were no light in the universe and therefore no creatures with eyes, we should never know it was dark. Dark would be without meaning."

"Reason is the natural order of truth; but imagination is the organ of meaning."

by C. S. Lewis (* 1898, Belfast - † 1963, Oxford)

Zusammenfassung

Die Anzahl der weltweit kontinuierlich aufzeichnenden seismischen Messstationen ist in den vergangenen Jahren immer weiter angestiegen. Aus diesem Grund steht eine große Menge von seismischen Datensätzen zu Forschungszwecken zur Verfügung. Insbesondere betrifft dies passive Verfahren zur geologischen Strukturerkundung entweder mittels transienter Ereignisse wie Erdbeben oder unter der Verwendung der permanent vorhandenen natürlichen seismischen Bodenunruhe. Die Bearbeitung dieser Daten erfordert neben der klassischen manuellen Seismogrammanalyse verstärkt auch den Einsatz automatischer Detektionssysteme. Mit Hilfe von *überwachten* Lernverfahren, d.h. unter Verwendung von seismischen Signalen deren Auftreten bekannt ist, ist es möglich, unbekannte Muster zu klassifizieren.

Im Gegensatz dazu hatte die vorliegende Arbeit zum Ziel, ein allgemeines, *unüberwachtes* Verfahren zur quantitativen Zerlegung seismischer Wellenfelder zu entwickeln. Dies wird mittels einer automatischen Clusterung von Seismogrammzeitfenstern bzw. über die Visualisierung von zeitlichen Mustern auf unterschiedlichen Zeitskalen erreicht. Als unüberwachtes Lernverfahren, das neben der Clusterung auch eine einfach interpretierbare Visualisierung hoch-dimensionaler Datensätze über eine zweidimensionale Darstellung ermöglicht, wurde der *Self-organizing-map* Algorithmus (SOM) gewählt. Für automatische Lernverfahren ist die Parametrisierung der Seismogramme mittels Merkmalsvektoren erforderlich. Im vorliegenden Fall wurden möglichst viele potentielle Wellenfeldmerkmale unter Verwendung von verschiedenen seismischen Einzel- und Mehrstationsanalyseverfahren für aufeinanderfolgende kurze Zeitfenster berechnet. Um eine datenadaptive und effiziente Parametrisierung zu erreichen, wurde darüberhinaus ein quantitatives Auswahlverfahren für geeignete Merkmale entwickelt, das über einen mehrstufigen Filter bestehend aus einem Signifikanztest und einer SOM-basierenden Korrelationsanalyse redundante und irrelevante Eigenschaften aussortiert.

Mit den neu implementierten Techniken wurden verschiedene Arten von seismischen Datensätzen unter Berücksichtigung verschiedener seismologischer Fragestellungen bearbeitet. Die Algorithmen und deren Parameter wurden zunächst intensiv und quantitativ mit Hilfe synthetischer Daten getestet und optimiert. Anschließend wurden reale Aufzeichnungen regionaler Erdbeben und vulkanischer Seismizität verwendet. Im ersten Fall konnten geeignete Merkmale zur Detektion und Klassifizierung von Erdbebenwellenphasen gefunden und die Diskriminierung dieser Signale mit Hilfe der SOM-Darstellung untersucht werden. Unter Verwendung des zweiten Datensatzes wurden Cluster typischer vulkano-seismischer Signale am Vulkan Mount Merapi (Java, Indonesien) detektiert, die sich zur Vorhersage von Eruptionen eignen. Beide Anwendungen haben gezeigt, dass, verglichen mit einzelnen Methoden, automatisch gefundene Kombinationen von Merkmalen

verschiedener Parametrisierungsverfahren deutlich bessere Klassifizierungsraten zur Folge haben. Zudem können die Erkenntnisse über die Clusterung von seismischen Signalen dazu verwendet werden, verbesserte automatische Klassifizierungssysteme zu entwickeln. Abschließend wurden Aufzeichnungen der natürlichen seismischen Bodenunruhe bearbeitet. Insbesondere konnte der Einfluss kurzzeitiger und längerfristiger Variationen im Wellenfeld auf Methoden zur passiven Strukturerkundung untersucht werden. Es hat sich gezeigt, dass in einzelnen Fllen tageszeitabhängige Muster und lokale seismische Quellen die Ergebnisse negativ beeinflussen können. Die Wellenfeldzerlegung mittels Clusterung hat es erlaubt, diese Signale zu identifizieren und somit von der Analyse auszuschließen.

Abstract

Modern acquisition of seismic data on receiver networks worldwide produces an increasing amount of continuous wavefield recordings. Hence, in addition to manual data inspection, seismogram interpretation requires new processing utilities for event detection, signal classification and data visualization. Various machine learning algorithms, which can be adapted to seismological problems, have been suggested in the field of pattern recognition. This can be done either by means of supervised learning using manually defined training data or by unsupervised clustering and visualization. The latter allows the recognition of wavefield patterns, such as short-term transients and long-term variations, with a minimum of domain knowledge. Besides classical earthquake seismology, investigations of temporal patterns in seismic data also concern novel approaches such as noise cross-correlation or ambient seismic vibration analysis in general, which have moved into focus within the last decade. In order to find records suitable for the respective approach or simply for quality control, unsupervised preprocessing becomes important and valuable for large data sets.

Machine learning techniques require the parametrization of the data using feature vectors. Applied to seismic recordings, wavefield properties have to be computed from the raw seismograms. For an unsupervised approach, all potential wavefield features have to be considered to reduce subjectivity to a minimum. Furthermore, automatic dimensionality reduction, i.e. feature selection, is required in order to decrease computational cost, enhance interpretability and improve discriminative power.

This study presents an unsupervised feature selection and learning approach for the discovery, imaging and interpretation of significant temporal patterns in seismic single-station or network recordings. In particular, techniques permitting an intuitive, quickly interpretable and concise overview of available records are suggested. For this purpose, the data is parametrized by real-valued feature vectors for short time windows using standard seismic analysis tools as feature generation methods, such as frequency-wavenumber, polarization, and spectral analysis. The choice of the time window length is dependent on the expected durations of patterns to be recognized or discriminated. In general, the length should not be less than four to five times the period of the signal of interest. We use Self-Organizing Maps (SOMs) for a data-driven feature selection, visualization and clustering procedure, which is particularly suitable for high-dimensional data sets.

Our feature selection method is based on a relevancy filter performing significance testing for individual features and on a redundancy filter applying correlation hunting in feature subsets. In particular, we iteratively reduce the number of features by first assessing the temporal randomness of a feature time series using the Wald-Wolfowitz runs test and by comparing the observed and theoretical variability of features. Thresholds

are employed for temporal significance and variability. For the redundancy filter, the in-between feature dependencies are assessed using Self-Organizing Maps to group correlated features. Representative features from each group are obtained from a ranking given by the runs test.

Finally, using the selected features, a SOM is learned and clustered. Various SOM visualizations using color scales for individual feature values, distances between neighborhood data points, and similarity of features vectors allow the identification and investigation of patterns in the data set, i.e. clusters of seismic signals or wave phases.

Using synthetics composed of Rayleigh and Love waves and three different types of real-world data sets, we show the robustness and reliability of our unsupervised learning approach with respect to the effect of algorithm parameters and data set properties. Furthermore, we approve the capability of the clustering and imaging techniques. For all data, we find improved discriminative power of our feature selection procedure compared to feature subsets manually selected from individual wavefield parametrization methods. In particular, enhanced performance is observed compared to the most favorable individual feature generation method, which is found to be the frequency spectrum.

The method is applied to regional earthquake records at the European Broadband Network with the aim to define suitable features for earthquake detection and seismic phase classification. For the latter, we find that a combination of spectral and polarization features favor S wave detection at a single receiver. However, SOM-based visualization of phase discrimination shows that clustering applied to the records of two stations only allows onset or P wave detection, respectively. In order to improve the discrimination of S waves on receiver networks, we recommend to consider additionally the temporal context of feature vectors.

The application to continuous recordings of seismicity close to an active volcano (Mount Merapi, Java, Indonesia) shows that two typical volcano-seismic events (VTB and Guguran) can be detected and distinguished by clustering. In contrast, so-called MP events cannot be discriminated. Comparable results are obtained for selected features and recognition rates regarding a previously implemented supervised classification system. Furthermore, patterns in the background wavefield, i.e. the 24-hour cycle due to human activity, are intuitively visualized using SOMs.

Finally, we test the reliability of wavefield clustering to improve common ambient vibration analysis methods such as estimation of dispersion curves and horizontal to vertical spectral ratios. It is found, that in general, the identified short- and long-term patterns have no significant impact on those estimates. However, for individual sites, effects of local sources can be identified. Leaving out the corresponding clusters, yields reduced uncertainties or allows for improving estimation of dispersion curves. Furthermore, it is shown that these disturbing patterns can become important, particularly at night when the overall energy of the wavefield is reduced due to the 24-hour cycle.

Contents

Chapter 1

Introduction

1.1 Signal versus Noise: The Importance of Data in Modern Seismology

Almost all research in seismology is based on recordings of ground motion due to propagating seismic waves. These data sets provide important information about location and occurrence of earthquakes and allow us to estimate the properties of the medium through which the waves are propagating. This kind of knowledge is mandatory for seismic hazard assessment and to understand structures and dynamic processes within our planet. Thus, seismic data is important to people working in the field of geosciences and, moreover, to society in general.

After the pioneering work on instrumental seismic registration by people like Palmieri in 1856 or Cecchi in 1875, the first seismograms became available after 1880. In early observatory practice, only earthquake recordings, which were detectable using the limited seismometer sensitivity and the available processing capacity of the equipment, had been considered of interest. Prominent milestones during this period include the first ever event registration, by Verbeck, Milne and colleagues in Japan (1873), or the first teleseismic record, observed by Ernst von Rebeur-Paschwitz in Potsdam (1889). Over the next decades, until 1950, gradual advances in theory and instrumentation allowed for more and more systematic investigations on the nature of earthquakes. In most of those studies, the remaining records, before and after an event, were of no interest and were therefore, considered and treated as "noise". Only a few studies, before 1950, about the nature of the seismic background wavefield, are known (see review of Bonnefoy-Claudet et al., 2006). Furthermore, during this early time period of seismological research, the number of available seismic receivers was limited, and manual analysis of earthquake recordings by seismologists was standard procedure.

Since that time, more permanent and temporary seismic networks have gradually been installed. Furthermore, due to technical advances in semiconductor electronics, more storage capacities, computation power, digital instead of analog records, and modern seismometers have become available. Therefore, an increasing amount of highly-resolved, continuous data is produced daily. Today, in order to find recorded earthquakes or any other temporal patterns of interest, large seismic data sets can no longer be completely processed by hand. Manual seismic phase picking alone, can be too time consuming, for instance when real-time processing is required for early warning systems. Furthermore,

visual seismogram interpretation involves the risk of subjectivity. Therefore, assistant analysis tools are required in order to gain new insights.

When we talk about patterns in seismic recordings, we usually mean distinct arrivals of waves characterized by suddenly increasing amplitudes and a changing frequency content compared to the background wavefield. These signals can be generated by earthquakes but also by man-made sources. With respect to energy the latter can range from footsteps close to the instrument to nuclear explosions. One challenge in seismology is to distinguish between these events and to detect earthquake onsets and seismic wave phases.

Beside intensified activity on transient signal detection, the availability of long, continuous recordings from locations worldwide enabled us to address another new aspect of seismic data analysis. Over the last two decades, increasing attention has been paid to the permanently-measured background wavefield, which is known as *seismic noise*, and is, in fact, a superposition of waves, excited by natural and man-made sources. Therefore, the term "noise" may be misleading. The inter-event data cannot be characterized as random, in a statistical sense (e.g. White or Gaussian noise), although the distribution of sources can be modeled, on some scales, as random in time and space. Another definition, often used to characterize noise, is the absence of coherency or correlation. However, since we are dealing with time series recorded at different locations, this is not correct in all cases. Random noise can be correlated between two receivers if it is generated by the same process. On the other hand, deterministic ground motion need not be coherent at all stations (Scales & Snieder, 1998). Thus, there is no simple definition of seismic noise. In fact, distinguishing between signal and noise in an arbitrary seismogram is subjective and depends on the particular aim of the research. For observatory seismologists who monitor global, regional, or local seismicity, the appropriate, conservative definition for noise is the absence of an earthquake signal. However, if we consider a contentiously-propagating wavefield, then also seismic noise provides the potential to extract medium properties. Therefore, the background wavefield becomes the signal, which is no longer a very short part of the record, but is the entire data set itself. Moreover, sometimes the contrary is the case. Transients might be called noise in this context as they may disturb the analysis of the wavefield. For that reason, the term *ambient seismic vibrations* is often used instead of seismic noise. However, in the scientific community the latter is still quite common.

The discussion may be summarized in the statements: "Someone's noise is another one's signal." (Name of a session at the American Geophysical Union meeting 2006) or "Noise is what we do not want to explain". (Scales & Snieder, 1998). For instance, one may want to explain only one particular wave type or phase in the data, hence, the signal is the part of the data that fits the chosen model. For ambient seismic vibration analysis, the widely-used model is a wavefield of planar surface waves. By applying a common concept in seismology, known as *stacking*, one can deal with short-term deviations from this assumption and can also deal with random noise. By averaging analyses over time or by using different records (e.g. a receiver array) researchers are able to reduce the effect of such "ambient vibration noise".

Hence, the use of ambient vibration wavefield records would allow passive investigations of crustal structures, especially 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 transients, which we will call *short-term patterns*, changes in wavefield characteristics over longer time scales, ranging from

hours to days or to months (*long-term patterns*), can become important. Of course, the transition between both kinds of patterns is continuous. Whenever a transient event is lasting longer than expected, it may also be called a long-term pattern. On the other hand, short-term variations in the wavefield might not be necessarily characterized by increased amplitudes. In general, all these patterns may affect the quality or usability of site-structure information, extracted from the ambient vibration wavefield.

Another problem associated with large data sets is the assessment of data quality. Since detailed, manual inspection of all waveform data might not be possible, instrument failure or non-seismic signals are hard to detect. Moreover, even in the case of visual control, corrupted data may not be distinguishable from real seismic recordings.

Therefore, summarizing all the problems and also their consequent potentialities, which occur when using long, continuous recordings from modern data acquisition systems, the following questions can be asked:

- How can we process large data sets to find short-term and long-term patterns?
- How, and under what assumptions, can we make use of the noise wavefield?
- How do temporal variations in noise wavefield affect results?

In the next section, we will formulate, in more detail, the objectives of this thesis with respect to these three questions.

1.2 Research Objectives and Thesis Outline

The first two questions, set out in Section 1.1 above, have gained importance in the seismological community in the last 10 to 20 years. The next two chapters will introduce basic terms and techniques, which summarize, in detail, the actual state of the research in relation to both questions and highlight any adaptations that this study makes. While Chapter 2 discusses pattern recognition and its application in seismology, Chapter 3 explains ambient seismic vibrations in more detail and introduces the common analysis methods.

In this work, we want to suggest, apply, and evaluate techniques for unsupervised investigations on seismic wavefield recordings. The goal of such an approach is the automatic recognition of patterns on different time scales. In particular, this should be achieved by data abstraction, i.e. realization of data grouping (clustering) and fast interpretable data visualization. We intend to achieve an intuitive imaging or highlighting of recognized, temporal patterns within seismograms. For the data processing, we want to use as little domain knowledge or preconceptions as possible. In other words, we aim to "let the data speak for itself". As a consequence, the techniques used for this approach should be generic and adaptable with respect to the given data set. For instance, they should automatically select, out of as many potential attributes as possible, adequate wavefield properties that are suitable for pattern recognition. All employed techniques are introduced in Chapter 4.

The information concerning temporal patterns, which will be retrieved by our suggested approach, should support further analysis. For instance, it can be used to develop more sophisticated, supervised classification systems or simply to gain new insights into the

nature of wavefields. We demonstrate and assess the reliability of our approach in Chapter 5 using synthetic (Section 5.1) and recorded waveforms. In particular, using real data, the importance of identified short-term or long-term patterns are investigated with respect to different seismological problems. We employ regional earthquake (Section 5.2), volcano-seismic wavefield (Section 5.3), and ambient seismic vibration (Section 5.4) recordings. While earthquakes and volcanic seismicity are investigated regarding phase and event discrimination (short-term patterns), the latter application is used with emphasis on effects of short- and long-term patterns, on the results of common ambient vibration analysis techniques.

Chapter 2

Pattern Recognition

2.1 Introduction

Before we review automatic signal detection research in seismology in Section 2.4, we will first define and explain the most important concepts and terms in the general field of pattern recognition. We will follow the introduction given by Bishop (2006) and Theodoridis & Koutroumbas (1998). Some topics are discussed in more detail because they are integral subjects of this work.

Pattern recognition is the task of automatic searching and the discovering of regularities in data and grouping data, into different categories. On computers, these problems are approached by means of *machine learning* algorithms. Learning, which is also called *training*, means the phase where an algorithm adjusts a model to a given data set, and, where available, to additional *target information* (e.g. class labels). The training procedure involves the *generalization* of given information to potential new, unseen data without target values. The application of a trained algorithm is called the *testing* phase.

We distinguish between two different approaches of pattern recognition. For *supervised learning*, the training data has target information, e.g. a class membership or some other (continuous) value, whereas for *unsupervised learning* the training data is unlabeled. The first approach primarily involves *classification* and, in the case of continuous target variables, *regression*. For classification, a data set with known class memberships is used for training. The trained algorithm is applied to new, unlabeled data to predict class-memberships. On the other hand, unsupervised learning automatically identifies patterns in the data distribution like the natural grouping (*clustering*) or visualizes data by projections. Thus, there is only the training phase. However, based, for example, on a clustering, then testing or classification can be carried out by finding the (k-)nearest neighbor(s) and, thus, the most likely cluster-membership of new data. Doing so, clustering can be a first step in the process of supervised learning. Furthermore, unsupervised techniques are often employed in the field of *data mining*, where relevant information is automatically extracted from large data sets. In Section 2.3 we will discuss unsupervised learning techniques in more detail.

For supervised and unsupervised learning, a large amount of different machine-learning techniques exist. It goes beyond the scope of this work to give an overview of all approaches. However, most techniques can be formulated within the framework of probability theory (Fukunga, 1990; Bishop, 2006). The data itself and the learned model can be

considered as random variables. Pattern recognition aims for the estimation of probability distributions and, thus, also includes uncertainties of the data and the chosen model. The decision to identify a class or to define a cluster is based on the minimization of an objective function, which describes the risk of misclassification or the misfit between data and model, respectively.

For both the supervised and the unsupervised learning approach, the challenge of generalization is to avoid the overfitting of a learned model (e.g. a clustering) with respect to the training data. For instance, consider a model that explains all training data instances including their natural scatter, or a clustering, where each data item defines its own cluster. In those cases, test data will probably not be classified correctly or the model will be too complex and make no sense for further applications. Furthermore, to ensure its efficacy, training data must relate directly to the identified problem. One way to select an appropriate model and to consider the effect of missing data is to evaluate learning by *cross-validation* (CV) or to constrain it by *regularization*. For the first option the training data is repeatedly divided into new training and test data sets. Since several classification errors are obtained for different data set splits (*folds*), a mean error, and its uncertainty, can be estimated. The model with the lowest, average classification error would be the best choice. In Section 4.5, we will explain in more detail how we compute cross-validated classification errors.

In practice, input variables for pattern recognition problems are represented as multidimensional pattern vectors. Each data instance is presented by a vector of *features*. Computing features from raw data (e.g. measurements) is called *feature generation* and is an important part of pattern recognition. In Section 4.1, we will explain how and which features are generated from the raw seismic wavefield recordings using different parametrization methods. After features have been computed, the goal is to find an optimal data representation for the learned algorithm. A common way of doing this is to apply *feature extraction*, where a generated feature set, or the raw data vectors, are processed, e.g. by linear transformation. On the other hand, *feature selection* is the process of searching for the best combination or best subset out of all computed features. We will discuss feature transformation and unsupervised feature selection in more detail in the next section.

2.2 Unsupervised Feature Selection

For many pattern recognition problems, the number of all given, potential features can be very high. For instance, this number may be 1000, or more, in industrial applications (Fogelman-Soulié, 2008). Please note that often, these features have also already been selected a priori, based on domain knowledge. However, the information content or relevance of individual features, e.g. for clustering or imaging patterns in the data, may vary considerably. Furthermore, strong correlations between features will hide important information, which is encoded in less- or non-redundant components of the feature vector. Thus, computation time or occupied disk space may be unnecessarily increased and the quality of the final results may suffer from having useless features. Moreover, the higher the dimension of the data, the more data is needed for learning, and the less suitable is the Euclidian distance as a measure of similarity, due to the curse of dimensionality (Bellman, 1961; Bishop, 2006). Furthermore, interpretation of the results and characterization of

the data is much easier for a low number of features.

Again, a large number of methods for feature selection can be found in the literature. See, for instance Dash & Liu (1997) for the most important techniques. In general, we may distinguish between *wrapper*, *filter*, and *embedded* methods. Wrapper algorithms use a forward or backward selection procedure to search for the feature subset most relevant to the chosen learning method, according to a particular evaluation criterion. There are a lot of variants, which differ in the subset generator, the evaluation and stopping criteria. The computational complexity is very high for that approach since the learning phase has to be repeated for all potential subsets. Therefore, wrappers are not suitable for high-dimensional data sets. While filter methods reduce the feature set, i.e. by applying a threshold after feature ranking or by feature grouping based on redundancy, the selection of suitable features is integrated into the chosen learning method for the embedded techniques. One possibility is to learn weights for the features, according to their relevance.

It should be noted that useless features need not necessarily impair learning performance, aside from practical aspects such as speeding up training, saving disc space, and better interpretation ability. In fact, feature selection is mostly done with the goal of not losing classification or clustering accuracy (which is, strictly speaking, only a weak requirement for such an approach). However, for high-dimensional data sets of real-world problems, the computational aspects and the further use of results become more and more important.

Assessing the performance of a feature selection algorithm is not the complete story. Saeys et al. (2008) pointed out that robustness and stability, e.g. with respect to missing or later-added data, also has to be evaluated. In particular, Saeys et al. (2008) referred to problems of analyzing small sample sizes with many features. Moreover, these authors suggested so called ensemble feature selection, where multiple techniques are applied to the given data set. Finally, all results are considered to define suitable features. Such an approach might be necessary, since feature selection methods may be only appropriate to one particular hypothesis concerning the underlying model.

While a lot of approaches exist for supervised learning due to availability of labeled training data, unsupervised feature selection is a more recent topic of research. However, the subdivision into wrappers and filters is the same. Several approaches have been proposed to reduce the number of features, including dimensionality reduction algorithms, like the Principal Component Analysis (PCA). PCA, which is also known as the Karhunen-Loève Transform, is, strictly speaking, a feature extraction method. The goal of PCA is the minimization of correlation between features by transformation into a vector space with a new basis. More details are given in Section 4.6. Besides common, linear transformations such as PCA, there are also other, more recently developed methods for dimensionality reduction. Non-linear techniques are, for example Kernel PCA (Schoelkopf et al., 1997), the ISOMAP method (Tenenbaum et al., 2000), or local-linear methods, such as Laplacian Eigenmaps or Local-Linear Embedding (LLE, Roweis & Saul, 2000). However, 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.

An unsupervised wrapper method for clustering, has, for instance been presented by Dy & Brodley (2004) (Fig. 2.1). In that study, a forward selection was employed for the

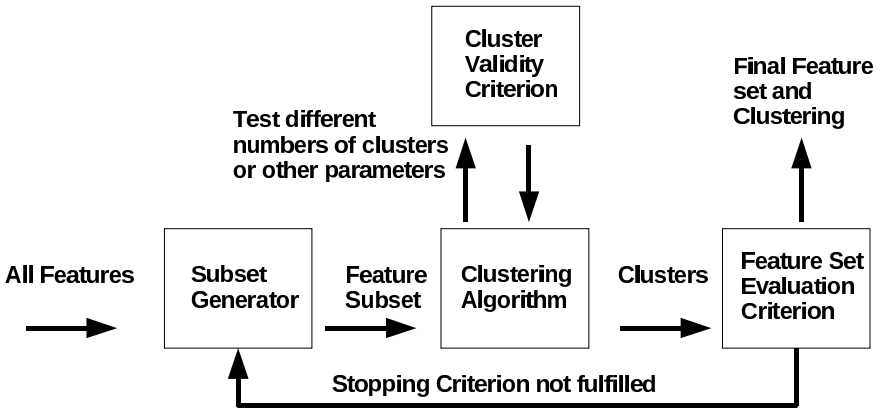

Figure 2.1: Wrapper approach for unsupervised feature selection after Dy & Brodley (2004).

feature subset generator. Starting with a single feature, all possible features were tested and the best one selected. Afterwards, a second feature was added and again, all possible combinations were tested. A new feature was added as long as the quality of the obtained clustering improved (stopping criterion). For cluster validity and feature set evaluation criteria, the DB index (Davies & Bouldin, 1979) and a normalized scatter separability criterion were used, respectively. In Section 4.5 we will give more details about those criteria.

In Basak et al. (1998), a fuzzy feature evaluation index for feature sets was used, which did not require clustering. Feature selection was done by finding the feature subset with the smallest index. For a second method this evaluation index was minimized using a Neural Network approach in order to find the relative importance of individual features. For the first method a search algorithm was still necessary. A technique requiring no search was suggested by Mitra et al. (2002). This method reduces feature redundancy by grouping features based on a pairwise feature similarity measure, called maximum information compression. Both approaches, Mitra et al. (2002) and Basak et al. (1998), were combined by Li et al. (2007), suggesting a two-level filter technique (Fig. 2.2). Feature selection is achieved by first reducing redundancy and then assessing relevance of each feature for clustering using the fuzzy feature evaluation criterion.

A very recent approach for unsupervised feature selection was introduced by Guérif (2008). The idea behind it is that randomly-ranked features belong to an equivalence class of irrelevant features. Whenever a feature set cannot be ordered properly, i.e. features are uniformly distributed with respect to their relevancy ranking, the irrelevancy class is found.

Figure 2.2: Filter approach for unsupervised feature selection after Li et al. (2007).

2.3 Unsupervised Learning

2.3.1 Clustering

For many real-life problems, clustering is a widespread tool. The most common application areas are in the fields of biology (species, gene or protein grouping), marketing (finding groups of customers with similar behavior), or in mining the World Wide Web. The latter comprises, for example, the organization of social networks and documents for search engines. In fact, data in the digital universe grows exponentially. Therefore, clustering becomes more and more important within this context.

The term "clustering" was first used by Tryon (1939). It can be explained as a meaningful grouping of unlabeled objects into respective categories. Meaningful clustering is obtained if the degree of association between two objects is maximal if they belong to the same group (intra-cluster similarity is high) or minimal if not (inter-cluster similarity is low). Clustering aims to discover the intrinsic, natural structure in the data without explaining why it exists. Performing clustering can be divided into several steps (after Jain et al., 1999). As it is for all pattern recognition approaches, the first step that is taken is feature extraction and selection. Besides conceptual clustering, where objects are grouped according to their fit to descriptive concepts, most clustering methods require a second step to be taken, which is the choice of a similarity measure to compare feature vectors. Most common are distance measures from the family of the Minkowski Metric:

$$\|\vec{x}\|_p \hat{=} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} . \tag{2.1}$$

If $\vec{x}$ is the difference between two feature vectors, $p = 1$, for instance, is the Manhattan or city block distance, which is taken for binary data. By setting $p = 2$, the most common Euclidean distance is obtained. Furthermore, more advanced measures, such as the Mahalanobis distance (no Minkowski Metric), take into account correlations between features. Finally, after defining a distance measure, the clustering algorithm itself, relying on a suitable model or objective function, has to be chosen.

The choice of features, distance measure, and algorithm depend on the particular problem or data set itself. Depending on the size of the data set, the dimensionality of feature vectors, and the expected cluster shapes, different algorithms are required. In fact, this is a very important challenge for clustering. A minimum amount of domain knowledge is required to choose techniques and to determine input parameters. However, in his so-called impossibility theorem, Kleinberg (2002) pointed out that there is no clustering approach that satisfies all of the three below-listed conditions. In that formulation a

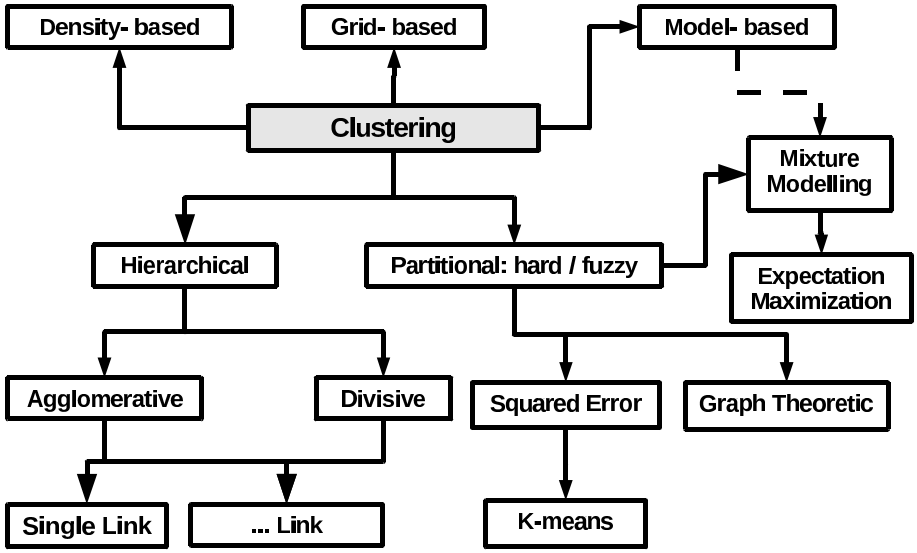

Figure 2.3: One possible subdivision of clustering approaches after Jain et al. (1999).

clustering function f takes a distance function d and generates a partition $\vec{P}$ (cluster labels) of the data $\vec{X}$: $f(d, \vec{X}) = \vec{P}$.

- Scale-invariance: For any d, f, and $a > 0$: $f(d, \vec{X}) = f(a \cdot d, \vec{X})$.
- Richness: The range of the clustering function is equal to the set of all partitions of the data.
- Consistency: For a partition $f(d, \vec{X}) = \vec{P}$ and two distance measures d and d', where d' is a transformation of d so that $d' \leq d$ within all cluster and $d' \geq d$ between clusters for all combinations of data items, also $f(d', \vec{X}) = \vec{P}$.

Thus, it is impossible to design an all-in-one device suitable for every purpose, which can discover clusters with arbitrary shapes and densities and which is able to deal with noise and outliers, and, moreover, is suitable for low as well as for high-dimensional data sets of arbitrary size. Furthermore, there are no general, theoretical guidelines for selecting a distance measure or algorithm for a particular application.

In order to overcome the problem of being too subjective when being limited to one particular algorithm, *multi-objective clustering* can be carried out. In this process, different algorithms are applied on the same data set. Finally, all groupings are put together to obtain an impression about possible solutions. Furthermore, many recent studies focus on so-called *semi-supervised learning* where constraints are used to limit the possible range of clustering solutions by employing a priori knowledge, e.g. class labels for some data instances.

Fig. 2.3 shows an overview of common clustering approaches modified after Jain et al. (1999). The most often applied concepts are *hierarchical* and *partitional* clustering.

Table 2.1: Linkage rules for hierarchical clustering.

Linkage	Description
Single	Distance between two closest objects of two clusters
Complete	Distance between the two most distant objects of two clusters
Average	Average of all distances between all pairs of objects not belonging to the same cluster
Centroid	Difference between means of both clusters
Ward's	Increase in the "error sum of squares" after fusing two clusters into a single cluster

Hierarchical Clustering

In hierarchical clustering, we distinguish between *agglomerative* and *divisive* approaches. For the agglomerative approach, the objects are merged into successively larger clusters using some measure of distance. Finally, a hierarchical structure called a *dendrogram* (a nested cluster tree) is obtained. At the base of the tree, each sample is a cluster; whereas at the top of the tree, one cluster contains all the data. Merging objects can be realized by using a distance threshold for each level, which discretely increases; or they can be realized by simply joining the closest objects at each level. Diversive algorithms work in a similar way. However, they begin with one cluster and successively split into smaller ones.

It is common not to compute the complete dendrogram. A stopping criterion can be introduced, for example when clusters are too far apart to be merged or when a sufficient number of clusters exist. Often, one is only interested in a single partition of the data. It is possible to cut the edges of the cluster tree for a particular number of clusters (*k-splitting*) or for a distinct distance threshold.

In order to compute a dendrogram, we obviously need a measure, which allows to compute the distance between a data point and a set of points (cluster). These measures are called *linkage rules* and are listed in Table 2.1. Different linkage rules are appropriate for different problems. While, for example the single-linkage rule has a tendency to create long and non-isotropic clusters, complete linkage produces compact clusters. Average linkage is suitable for both cluster shapes.

There are a lot more advanced variants of hierarchical clustering, for example by combining the hierarchical tree generation and other clustering methods. One example of a widespread method is the BIRCH algorithm (Zhang et al., 1996).

K-means Clustering

Partitional clustering produces one distinct grouping of the data. The best known method within this context is *k-means*. This algorithm was independently developed by different researchers (e.g. McQueen, 1967). See Jain (2008) for a more detailed overview. K-means generates exactly k clusters of the greatest possible distinction by iteratively adjusting the cluster centroids, i.e. the mean over all objects within a cluster. The underlying goal

is to minimize variability within clusters and maximize variability between them. This corresponds to the minimizing of an objective function, which is the squared error function (error sum of squares):

$$F = \sum_{i=1}^{k} \sum_{x_j \in P_i} (\vec{x_j} - \vec{\mu_i})^2 \ , \tag{2.2}$$

where P_i is one of k clusters and $\vec{\mu_i}$ the mean (centroid) of all data within. The processing scheme can be summarized as follows:

1. Initially start with k centroids (e.g. randomly chosen from the data).
2. Assign each object to the closest centroid.
3. Recalculate the positions of the cluster centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move or a convergence criterion is fulfilled, e.g. minimum decrease in squared error.

Since k-means does not necessarily find the global objective function minimum and is sensitive to the initial, randomly-selected centroids, the algorithm should be run multiple times. K-means is suitable to find isotropic clusters. For more complex data sets it may fail, given a simple Euclidian norm was employed. Furthermore, since it depends on the unknown parameter k, validation of different clusterings, employing different values for k, is required. There are a lot of variants of k-means such as the ISODATA and K-Medoids algorithm. Furthermore, there is a fuzzy extension called Fuzzy C-means. Fuzzy clustering methods produce a degree of class membership for each object and not a hard partitioning.

The theoretical background of clustering can also be explained using the framework of probability theory (Bishop, 2006). Using the concept of *mixture modelling*, it is assumed that the data is a sum of several, multi-dimensional distributions (e.g. Gaussians). The goal is to estimate the parameters of all distributions (e.g. means and standard deviations) by maximizing the likelihood considering data and model. Finally, grouping is given by the probabilities of cluster memberships, where a hard clustering can be easily obtained by choosing the largest classification probability of each object. An example is the *Expectation Maximization* (EM) algorithm, suggested by Dempster et al. (1977). It can be shown that for Gaussian distributions, the EM algorithm is similar to k-means. In fact, the objective function of k-means relies on the same theory, i.e. maximizing the likelihood results in the squared error function (Equation 2.2). The difference is that k-means, based on Euclidean distances, considers only the means and not the covariances of clusters.

Other Techniques

Besides hierarchical clustering and k-means, various new and advanced approaches have been published over the last decade (see Fig. 2.3). *Density Based Clustering* such as DBSCAN (Density based spatial clustering of application with noise, Ester et al., 1996) finds clusters based on connectivity of data. A cluster is defined as the maximal set of density-connected points. The strength of DBSCAN is that it works for arbitrary-shaped and overlapping clusters. It is able to identify noise and outliers, and automatically finds the number of clusters. However, other parameters have to be given instead, i.e. the

radius of neighborhood and minimum number of points in that radius, both defining density connectivity. A weakness is that these parameters are hard to adjust because they are less intuitive compared, for example to parameter k for k-means or hierarchical methods. Furthermore, DBSCAN only works for low dimensions (< 10).

Another, recent approach is nearest neighborhood clustering, e.g. the SNN method (Shared Nearest Neighborhood, Ertöz et al., 2003), which is also able to find clusters of different sizes, densities, and shapes. Moreover, it also works in cases of high dimensionality. However, as for DBSCAN, the parameters (minimum number of shared nearest neighbors and strength threshold) are more difficult to adjust.

Furthermore, *Graph-Theoretic Clustering* makes use of the *Minimum Spanning Tree* (MST) of the data. Clusters are generated by deleting the MST edge with the largest length or by grouping highly-connected edges. In fact, a single linkage dendrogram is a sub-graph of the MST. *Subspace clustering* partitions the data by using only relevant features for each cluster (Parsons et al., 2004). By employing local, linear embedding (e.g. ISOMAP, Laplacian Eigenmaps) and k-means, *spectral clustering* finds partitions of more complex structures than k-means, alone, can afford.

The existence of input parameters for all clustering algorithms (e.g. k for k-means) require the validation of clusterings. Since this is strongly connected to visual assessment, we will discuss this topic in more detail in Section 2.3.3.

2.3.2 Vector Quantization

Vector quantization is very similar to clustering. In fact, one realization is the application of k-means using a high number of clusters. In general, a set of so-called *prototype vectors* is learned for vector quantization, which is a representation of the probability density functions of the original data distribution. The method can be used for data compression. Each prototype approximately represents the same number of data points.

2.3.3 Validation and Visualization

For data compression, the representing prototypes, i.e. centroids or boundary points of clusters, are the finally-desired output. However, the found (set of) clustering(s) or the obtained vector quantization is often not the end of the story. For many applications, evaluation and further processing of the results is required. This can be done either by quantitative *validation* or by qualitative *visualization*.

Quantitative validation is mandatory for clustering, since most algorithms produce clusters, regardless of whether the data contains clusters or not. Thus, it is necessary to assess either the cluster tendency of the data set before clustering or to compute a quality measure afterwards. The following approaches can be carried out (Halkidi et al., 2002; Strehl, 2002):

- *Internal criterion*: Compare clustering with the inherent data structure
- *Relative criterion*: Compare various clusterings, e.g. to find suitable algorithm parameters.
- *External criterion*: Comparison with an assumed a priori structure, e.g. labeled or manually-classified data.

For internal criteria, the obtained partition can be compared with the proximity matrix $\mathbf{D} = d_{ij}$ by means of statistical hypothesis-testing techniques, where d_{ij} is the distance between data item i and j (Halkidi et al., 2002).

Relative criteria are suitable to compare clusterings generated by the same algorithm using different input parameters. The objective function of the employed clustering algorithm can be used (e.g. sum of squared errors for k-means, Equation 2.2). Furthermore, there are various other indices, which are all in the end based on averaged distances within and between clusters (Halkidi et al., 2002). For instance, consider the *Davies-Bouldin index* (DB, Davies & Bouldin, 1979) and the *scatter-matrix* (e.g. Halkidi et al., 2002). *Cluster stability* (CS, Lange et al., 2004) adapts the idea of cross-validation by evaluating the variability of results for missing data. More details and definitions for all three measures are given in Section 4.5.

Finally, there are various external (semi-supervised) validity criteria. Besides simply computing the classification error as the ratio of incorrectly classified and all data points, more advanced measures have been proposed. Strehl (2002) discussed several criteria, e.g. *purity* and *entropy*, which are also connected to classification accuracy. However, these criteria have been found to favor large numbers of clusters. Furthermore, using concepts from the information theory, Strehl (2002) introduced *precision*, *recall*, and the *F-measure*. Within this context, the author suggested a new criteria called *normalized mutual information* (NMI), which is less biased by the number of clusters. In Section 4.5 the NMI index and our employed classification accuracy measures are introduced in more detail.

Returning to the relative criteria, the most important algorithm parameter is probably the number of clusters k. Determination of the correct number is not trivial. In fact, often there is no unique number of clusters in the data and, furthermore, the risk of overfitting exists. The latter means that the best relative criteria are found for large number of clusters. Moreover, a consensus is reached by experts that there is no best, all-purpose relative measure for clustering (Vesanto & Alhoniemi, 2000). In order to overcome these problems, cross-validation-based approaches can be used (CS criteria). Furthermore, for the EM algorithm, adding a penalty term can avoid too large a number of clusters. Nevertheless, often there is no single answer to the "k-question". It depends on the aim of the application and domain knowledge, which single partition, or whether more solutions are considered as the final result.

A qualitative way to assess the validity of obtained clustering, is visualization, because the human cognition is well suited to such problems (see e.g. Shneiderman, 2002; Nattkemper et al., 2003). Intuitive visualization of data or models is also done in supervised learning, e.g. in graphical models for Bayesian Networks (Bishop, 2006). Since the distribution and grouping of vectors cannot be visualized in a conventional data plot for dimensions higher than three without loss of information, a visual assessment of clusterings is difficult. As mentioned, validity methods, such as internal or relative clustering criteria, can help to decide whether a particular clustering is a meaningful representation of the data or not. Most criteria rely on the assumption of relatively simple cluster characteristics. However, since shape and density of clusters can differ greatly and the overall number or even the existence of clusters is not known in advance, cluster validity measures may also 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 PCA or non-linear techniques. Furthermore, in cases of hierarchical clustering, presenting the dendrogram can be suitable to evaluate data partitioning.

Self-Organizing Maps (SOM Kohonen, 2001) is a convenient, unsupervised learning method, which is widespread in various scientific fields (see references in Kohonen, 2001). 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. Based on the relatively-simple map illustration, further processing can be done and easily validated. The SOM can be clustered using common methods, such as k-means or hierarchical clustering algorithms (Vesanto & Alhoniemi, 2000). The clustering of the SOM, or the SOM itself, can be used to label the data set according to cluster membership of a data point or its projected position on the SOM. Furthermore, SOMs can be used to assess correlations between features (Vesanto & Ahola, 1999). We choose the SOM technique as the basic framework within this study, and introduce it, in more detail, in Section 4.4.

2.4 Pattern Recognition in Seismology

Due to the increasing amount of data available from networks that monitor seismicity on local, regional or global scales all over the world, automatic detection and classification of seismic events is becoming more and more important. Consider, for example the preparation of input data for advanced investigation of the earth's subsurface, such as 3D tomography, the development of automatic warning systems at volcanoes, or monitoring the compliance with the nuclear test band treaty (CTBTO). Regarding the literature in this field, great progress has been made, since the emergence of simple short-time over long-time average (STA/LTA) triggers in the early 1960s (see review Withers et al., 1998). Today, inspired by pattern recognition or data mining techniques developed in various research fields such as speech recognition, many approaches have been applied to seismic recordings. Nevertheless, in observatory and research practice simple algorithms are still quite common.

The simplest goal is signal detection that corresponds to a two-class problem (signal and noise). One is primarily interested in detecting or identifying the onsets of earthquakes. In doing so, false alarms or missing an event should be avoided. One challenge is that noise is not a well-defined or very heterogeneous class; please refer to the next chapter for more details. In contrast to detection, classification of different types of events and seismic phases is more difficult and requires advanced techniques due to the complex nature of seismograms.

The idea of signal detection by introducing a threshold, as for STA/LTA, has been adapted by several studies. Schimmel & Gallart (2003) used a measure, for the degree of polarization, instead of amplitude ratios. Bai & Kennett (2000) introduced thresholds for a combination of different wavefield properties, whereas Roberts et al. (1989) used a coherency threshold. Coherency is computed between the considered data window and a P or S wave model using properties of the three-component covariance matrix of the seismogram. In Christoffersson et al. (1988) the authors employed a Maximum-Likelihood estimator, instead of a threshold, to evaluate the degree of fit for specific wave type models.

More recent, techniques well established in other research fields, such as pattern matching (Joswig, 1990), artificial neural networks (Dowla et al., 1990; Dai & MacBeth, 1995;

Wang & Teng, 1997), hidden Markov models (Ohrnberger, 2001), and dynamic Bayesian networks (Riggelsen et al., 2007), have been employed and adapted to seismic data.

However, automated seismic data processing still has to rely on expert knowledge of seismologists. In fact, selecting a suitable data parametrization (e.g. Bai & 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, e.g. clustering (Bardainne et al., 2006; Esposito et al., 2008). Here, no expert knowledge (class labels for the training data) is required in the learning phase. 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, is still a task of the seismologist. The latter can be addressed, for example, by unsupervised feature selection.

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 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.

For both the supervised and the unsupervised learning approach, features have to be generated from the raw recordings because the number of seismogram samples is too large, dependencies between all three components have to be considered, and the amplitude should not be the dominant pattern. There are a lot of common feature generation approaches in seismology, most based on time-frequency spectra and polarization attributes. Features can be either computed for a short-time window of the seismogram (e.g. phase discrimination) or for an entire, preselected event (event discrimination). More details about feature generation methods are given in Section 4.1.

Self-Organizing Maps have already been applied, in different contexts, in seismology (Maurer et al., 1992; Musil & Plešinger, 1996; Tarvainen, 1999; Plešinger et al., 2000; Esposito et al., 2008) and for active seismic data sets (Essenreiter et al., 2001; Klose, 2006; de Matos et al., 2007). The first mentioned studies in seismology are mainly investigating event discrimination, thus, presenting a single feature vector for each event. While the aim of Tarvainen (1999) was to recognize mining explosions using amplitudes, P-S time differences, and azimuths of array recordings, Musil & Plešinger (1996) and Plešinger et al. (2000) investigated micro-earthquakes and mining events, employing the frequency spectra and amplitude parameters. The recent study of Esposito et al. (2008) concerns the discrimination of very-long-period events at the Stromboli Volcano.

Finally, aside from grouping temporal patterns, the clustering of the spatial distributions of earthquake locations is also undertaken in Seismology. However, this goes beyond the remit of this study.

2.5 Objectives

In this study, we want to employ unsupervised learning techniques, i.e. SOMs and clustering (k-means and hierarchical techniques) because seismologists often deal with unknown, complexly-composed data. As an initial learning step, unsupervised learning has the potential to aid further processing. While Bardainne et al. (2006) and the above-mentioned SOM-using studies 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 parametrizations of short-time windows (beginning in the range of seconds). For this purpose, we use different feature generation methods, which are common in seismology (see Section 4.1). From this large set of potentially-suitable attributes, a smaller, optimal subset, including discriminative and significant features, should be automatically selected and combined in a single feature vector.

Since an exhaustive wrapper search, based on repeated clustering, is not appropriate for our real-world problem that has a large number of features, a filter approach for unsupervised feature selection is more promising. Furthermore, we want to keep features that might show no clear cluster tendency but will have significant patterns in their time history, which is typical for continuous transition between signals and noise. Therefore, using similar ideas to Li et al. (2007), we will introduce a multi-level feature selection procedure in Section 4.7. In particular, we will use significance testing, using the Wald-Wolfowitz runs test (Wald & Wolfowitz, 1940), which will be introduced in Section 4.2, as a temporal, context-dependent, feature-relevance measure. Furthermore, we will use SOMs for redundancy reduction.

Chapter 3

Ambient Seismic Vibration Analysis

As mentioned in Chapter 1, the **A**mbient **S**eismic **V**ibration wavefield (ASV wavefield) is the permanently-existing seismic ground motion. Due to its potential to extract subsurface, elastic properties, ASV analysis has been the topic of several research projects within the last few years (e.g. SESAME, Bard, 2004). Since investigations on the ASV wavefield are strongly connected to the common analysis techniques, we will first introduce the methods and afterwards summarize what we know about its nature.

3.1 Techniques

Two kinds of wave propagation properties are usually considered for ASV analysis (see e.g. Milana et al., 1996; Ohmachi & Umezono, 1998; Bard, 1999). The first one is the apparent seismic wave velocity on the earth's surface. If it is frequency dependent, this propagation characteristic is called *dispersion curve*. Furthermore, the spectral amplitude ratio between horizontal and vertical components is of particular interest (*H/V spectral ratio* or *HVSR*). To estimate both the dispersion curves and the H/V ratios, stacking (averaging) in time is done to reduce random noise. The meaning or interpretation of results depends on the assumptions made about the nature of the wavefield.

3.1.1 Array Methods

In order to estimate horizontal seismic velocities and their propagation directions, a network (array) of several receivers is required. The more stations are available, the more reliable is the estimation. In particular, the aperture (size) of the array determines the lowermost and the lowest station distance the uppermost usable frequency. Thus, since often a limited number of receivers is available, there is a tradeoff between aperture and station coverage for a good frequency resolution. Realization of several arrays of different apertures, using the same stations, can help to overcome this problem (Kuehn et al., 2006).

Several array methods were developed between 1950 and 1970. The most common technique to be used is Frequency-Wavenumber (f-k) analysis (Lacoss et al., 1969). Here, beamforming is done for all possible, discrete horizontal wavenumber vectors. Wavenumber vector are determined for different frequency bands by selecting the maximum power

or coherency from the 2D wavenumber grid. Finally, the wavenumber can be easily translated into velocity or slowness. Another method is the Spatial Autocorrelation Method (SPAC, Aki, 1957). For SPAC, the horizontal slowness is derived from the correlation functions at zero lag computed between all receiver pairs. A couple of modifications and extensions of the traditional SPAC technique have been developed since Aki (1957). Bettig et al. (2001) suggested the modified vertical SPAC method (MSPAC), which allows for arbitrary array layouts. Cho et al. (2006) presented an extended three-component array method including traditional SPAC as a special case. For a more advanced uncertainty measure for autocorrelation functions, Asten (2006) suggested the use of the complex autocorrelation function. The non-zero imaginary part can be considered as a measure for the deviation from the assumption of multi-directional, planar surface waves or for insufficient station coverage. Finally, Köhler et al. (2007) extended MSPAC to all three wavefield components.

Using the assumption of planar surface waves, the estimated dispersion curves can be interpreted (Aki, 1957; Lacoss et al., 1969; Asten & Henstridge, 1984; Horike, 1985; Tokimatsu, 1997). In general, the mixing of Rayleigh and Love waves occurs on the horizontal components. In contrast to SPAC on three components, there is no clear decomposition on horizontal components into Rayleigh and Love waves for f-k, except when the direction of wave propagation is known. Using the SPAC method, it is also possible to estimate the proportion of Rayleigh and Love waves on the horizontal components (Köhler et al., 2006, 2007).

3.1.2 Single-Station Method

Nakamura (1989) suggested the use of the H/V ratio of the ASV wavefield at single stations to determine the frequency-dependent site transfer or amplification function, respectively. Since this was done originally for earthquake recordings (body waves) to cancel out source effects, it was assumed that the H/V ratio corresponds to SH wave resonance at the site. However, when the ASV wavefield is dominated by surface waves (e.g. Tokimatsu, 1997; Arai & Tokimatsu, 2004), the amplification is not directly obtained. For pure Rayleigh waves, H/V corresponds to the site-dependent ellipticity. Furthermore, when there is a contribution of Love waves, only the H/V peak frequency can be considered as the maximum ellipticity. However, Malischewsky & Scherbaum (2004) showed that the frequency of maximum SH amplification and highest ellipticity are very similar for most types of sites

3.2 Origin and Nature of Ambient Seismic Vibrations

As Bonnefoy-Claudet et al. (2006) pointed out, most of the studies that were published after 1970, were on applications and not on the origins of ASV wavefields. However, within the last decades, some basic knowledge and assumptions have been derived from observations. We will summarize the findings following the review of Bonnefoy-Claudet et al. (2006).

The ASV wavefield is generated by natural as well as by man-made sources. Generally speaking, the natural sources excite frequencies below 1 Hz, which is called *microseism*. The low-frequent sources (0.005-0.6 Hz) are mainly due to earth tides and the coupling

of water waves with the ocean bottom. The latter effect is connected to large-scale meteorological phenomena, i.e. atmospheric turbulence characterized by wind, temperature, and air pressure changes. Examples of the most important source regions of high activity are the Pacific Ocean or the Labrador Sea (Toksöz & Lacoss, 1968). Closer sources are ocean waves striking at nearby coasts, which produce more high-frequent ground motion (about 0.8 Hz). Regional or local atmospheric turbulence is also responsible for energy close to 1 Hz. For instance, consider the transfer of wind energy to the ground by trees or buildings. While at very low frequencies, source distribution is directional, for short periods it becomes more heterogeneous.

On the other hand, anthropogenic sources dominate above 1 Hz. Here, the term *microtremors* has been established. Microtremors are mainly produced by industry and traffic. Therefore, they are observed particularly in urban areas. Depending on the site, the distribution of their sources can be very heterogeneous. In simulations, sources are often modeled as random in time and space.

Often, a mixture of both the microseism and the microtremor part is observed around 1 Hz. Depending on the wind conditions, natural sources can excite ground motion above 1 Hz. On the other hand, microtremors are also observed below 1 Hz, e.g. in deep sedimentary basins.

The composition of ASV has been controversially discussed within the context of the H/V method. While microseism are considered dominated by Rayleigh waves due to ground coupling with ocean waves, some authors assumed that microtremors contain a significant proportion of body waves. If this is the case, the H/V spectrum can be interpreted as the amplification function due to SH wave resonance. However, today, most researchers assume a mixture of surface and body waves, depending on site structure and sources. The dominating waves, however, are assumed to be surface waves. Nevertheless, it is not easy to give an average number of proportion. The same applies to the contribution of Rayleigh and Love waves. However, it seems that Love waves dominate microtremors (Aki, 1957; Köhler et al., 2006). Higher-mode surface waves are also often observed, depending on local conditions (layers and sources).

The uncertainties about proportion of body, Rayleigh, Love waves, and higher mode contribution, affect the interpretation of H/V spectra and dispersion curves. While the H/V peak frequency is a robust site parameter, care must be taken in the use of the H/V shape (amplitudes). Furthermore, it is a challenge to identify higher modes in dispersion curves because, for example, of the overlapping and intersections of mode branches. Within this context, recently-developed techniques, such as automatic decomposition into different modes (Cho et al., 2006) or direct determination of the Rayleigh-wave ellipticity from P-SV wavelets (Fäh et al., 2001), can be used.

3.3 Use of Ambient Seismic Vibrations

The main target sites for ASV measurements are urban areas of high seismic hazard for which strong amplification, due to soft-sediments in the subsurface, are expected. Thus, knowledge about the structure is mandatory for the forecasting of ground motion, either by stochastic models, including a parameter for the local structure, or by numerical simulations of earthquake wavefields. In particular, the depth-dependent S wave velocity is required as an input parameter. As a low-cost alternative to, or in combination with,

active geophysical experiments or other geotechnical methods, this information can be jointly, or separately, inverted from the estimated dispersion curves and H/V spectral ratios (e.g. Herrmann, 2001; Scherbaum et al., 2003; Wathelet et al., 2004; Parolai et al., 2005). Since one is mainly interested in the shallow structure (up to a few hundreds of meters), the employed frequency band ranges from about 0.1 to 30 Hz.

Analysis of ASV is also done on larger scales. Over the last few years, a lot of studies have been published dealing with regional, surface wave tomography for deeper structures (e.g. earth's crust), based on a method suggested by Shapiro et al. (2005). The idea is to extract the medium Greens functions by cross-correlation of the ASV wavefield between two stations. Using a receiver network, this then corresponds to a data set of travel times on different source-receiver paths. In fact, it can be shown that this approach is very similar to the SPAC technique. Moreover, noise cross-correlation has also been developed and employed in other research fields, which are not dealing with the solid earth, e.g. in under-water acoustics (Roux & Kuperman, 2004) and helioseismology (Duvall et al., 1997). The latter example uses oscillation of the sun as input data. Roux et al. (2005) presented a generic formulation for cross-correlation, which is applicable in all three fields (seismology, acoustics, and helioseismology).

3.4 Temporal Patterns in Ambient Vibration Wavefields

For array methods, H/V ratios, and noise cross-correlation, it is usually assumed that the ASV wavefield is stationary in time and space. In order to avoid effects of short-term variations, results are averaged over long time periods. Depending on the frequency band of interest, about one hour (local subsurface) up to several months (global and regional tomography) are common practices.

In Chapter 1 we introduced the terms long-term and short-term patterns. While the first one corresponds to transients, the latter describes continuous variations on a longer time scale. We will now briefly discuss the importance of both patterns regarding the nature and analysis of ASV wavefields.

3.4.1 Long-Term Patterns

Various studies about variations in ASV wavefield have been published (see again review of Bonnefoy-Claudet et al., 2006). For microtremors, daily changes (between day and night) and weekly variations in energy contribution (between working days and weekends) are observed due to the rhythms of human life. Patterns in microseism at around 1 Hz exist due to changes in local weather conditions, such as wind speed. Variations of longer time scales for lower frequencies exist due to seasonal, meteorological changes. For instance, the mean ocean wave heights change between winter and summer.

Since ASV measurements are mainly carried out over very short time periods and noise cross-correlation is a rather new technique, no studies are known, which systematically deal with the impact of long-term patterns on results. Considering only one particular wavefield property and a data set limited to three sites and 24 hours, Köhler et al. (2006) investigated the proportion of Rayleigh and Love waves, using the 3c-MSPAC method. No significant changes could be found for the two urban sites. However, for the rural site (Colfiorito, Italy), a slightly increased contribution of Love waves during the daytime

was observed. On the other hand, variation of the cross-correlation function of the ASV wavefield on larger scales were attributed to changes within the medium itself (Brenguier et al., 2008), which would allow analysts to monitor volcanoes, for example.

3.4.2 Short-Term Patterns

Transients, generated by close sources, are part of the microtremor wavefield. Since they are mostly directly triggered by human activity, transients are more frequent in the daytime. Parolai & Galiana-Merino (2006) investigated the effect of those signals on H/V spectra. Comparing the ASV analysis results obtained with and without including the transients, the authors found no significant effect for different sites and frequencies, even for a short data set of 30 minutes. When only transients were considered, a large variability was observed in the H/V shape, but less for the H/V peaks themselves. However, due to insufficient data, results became unstable for low frequencies. Mucciarelli et al. (2003) used actively-triggered transients for the estimation of H/V spectral ratios. Finally, Roberts & Asten (2008) investigated the effects of close sources on the SPAC method, i.e. wave-front curvature. By means of theoretical modeling and field studies, the authors showed that the effect was minimal, given that the distance to the center of the array was larger than two array radii for the majority of sources. If this assumption is violated, the study showed that shear wave velocities are underestimated

3.5 Objectives

Our goal is to find and investigate short- and long-term patterns in ambient vibrations by clustering time windows in order to obtain further insights into the nature of the wavefield (Section 5.3 and 5.4). Furthermore, we aim to test whether this wavefield decomposition has the potential to improve common analysis methods (Section 5.4). In particular, we want to find out whether resulting clusters correspond to the following classes:

- Pure Rayleigh and Love waves
- Fundamental and higher mode surface waves
- Pure surface waves
- More and less suitable time windows for f-k, SPAC, and H/V analysis

We will conduct these investigations on small-scale array measurements carried out over the last 10 years at many urban and a few rural sites. For a limited number of sites, longer records are available (one or more days). Furthermore, the developed techniques will be applied to an ASV wavefield, which was recorded close to an active volcano (Section 5.3).

Chapter 4

Methods For Our Unsupervised Approach

In this chapter, we introduce the techniques which we will employ for the unsupervised investigation of seismic wavefield recordings. Section 4.1 is about feature generation, i.e. the parametrization of seismic data. The assessment of the randomness of an individual feature is done using the Wald-Wolfowitz significance test, which is explained in Section 4.2. In Section 4.3 the employed clustering algorithms are presented. A detailed introduction into Self-Organizing Maps is given in Section 4.4. In Section 4.5 several validation and evaluation measures are introduced, which we will use in Chapter 5 to assess the reliability of our processing approach. Finally, we will suggest a novel unsupervised feature selection method in Section 4.6, which combines the techniques described in the before-mentioned sections of this chapter.

4.1 Feature Generation

For the 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 event is easily identified by considering a few characteristic parameters of the seismogram, such as, e.g., 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 4.1). The manual combination of features from different parametrization methods has already been proposed in the context of supervised classification of seismic events (Jepsen & Kennett, 1990; Wang & Teng, 1997; Bai & Kennett, 2000; Ohrnberger, 2001). Since we want to implement a generic,

Table 4.1: Name of method, corresponding references and number of features for each feature generation approach. Number of features refers to the use of three frequency bands for Methods 1-4 and 7. For Methods 5 and 6, 10 bands are employed.

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

unsupervised approach using only a minimum of domain knowledge, we collect features from all proposed methods. Note that we do not use features which are directly dependent on amplitudes in our study (e.g. absolute power of f-k analysis, seismogram envelope), since they may differ strongly for a particular type of signal or seismic phase. Nevertheless, for other application fields such as active seismic experiments, those features could easily be included in our processing scheme. Furthermore, the azimuth of arriving waves is not used due to its cyclic nature, which is not suitable for clustering based on feature vectors. However, the variance of the instantaneous azimuth within a time window (*vazi*) and the *H/E* ratio are employed.

We summarize all parametrization methods and the corresponding features, including instantaneous attributes from complex trace analysis, polarization from covariance matrix eigenvalues, spatial coherency, spectral properties, and amplitude ratios, in Table 4.1 and 4.2. All features are calculated for short three-component time windows whose lengths T is depending on the considered frequency band: $T = WINFAC \cdot 1/f_{cent}$. Here, f_{cent} is the center frequency of the overall frequency band of interest (logarithmic scale) and $WINFAC$ is a free parameter. Thus, to ensure that at least one period of the signal is present in a window, $WINFAC$ should not be lower than one. Each time window is used to compute the f-k spectrum (Method 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. For our parametrization they are averaged over the complete time window. The feature *spacim* of

Table 4.2: Features and short names for each parametrization method. Superscript numbers refer to the references in Table 4.1.

Method	Feature Description	Short Name*
1	semblance: vertical, radial and tangential component	*pr*
2	real and imaginary (absolute value), frequency band averaged autocorrelation coefficients: vertical, radial and tangential component	*spac*, *spacim*
3	degree of polarization [1]	*dopII*
	ellipticity, strength of polarization, angle of incidence, planarity [2]	*ell*, *sop*, *inc*, *plan*
	linearity, planarity [3]	*rect*, *planII*
	linearity (two definitions), stability of direction cosine, enhanced linearity [4]	*linII*, *linIII*, *sdc*, *elin*
	enhanced linear polarization [5]	*elip*
	degree of polarization [6]	*dopIII*
4	instantaneous frequency and variance: vertical and horizontal component	*if*, *vif*
	phase difference, ellipticity and tilt between vertical and horizontal components [7]	*pdiff*, *ell*, *tilt*
	variance of azimuth, ellipticity [8]	*vazi*, *3cell*
	component-averaged instantaneous frequency [9]	*FQ1*
	degree of polarization, linearity [10]	*dop*, *lin*
5	normalized horizontal and vertical sonogram	*sono*
	dominant spectral frequency and bandwidth: vertical and horizontal component	*domf*, *bb*
	logarithm of ratio between sum of lower and higher sonogram bands	*ratiolf*
6	normalized semi-mayor 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*

* Suffixes for short names:
Component: z (vertical), e (east), n (north), h (horizontal), r (radial), t (tangential)
Frequency band index: 1, 2, ..., 3, (..., 10)

Method 2 describes the quality of SPAC coefficient estimates, which is related to deviations from the assumption of multi-directional, planar surface waves.

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 ($tt > 0.5 \cdot T$). The expected travel time is estimated empirically assuming an appropriate seismic velocity. If sub-arrays become necessary, only a single receiver combination is used for feature averaging.

Except for the sonogram and the method after Pinnegar (2006), where we choose 10 bands, most features are computed for three different frequency bands. The number of features given in Table 4.1 corresponds to this particular number of bands. However, in general, more frequency bands can be used for all features (see e.g. Section 5.4). The lowermost and uppermost frequency limits are chosen according to the frequency content of the data or, if a priori information is available, by choosing the band including the expected patterns. Note that there are also features which take into account the entire frequency spectrum up to the Nyquist frequency (*domf*, *bandwidth*). For simplicity, the frequency band index and the component suffix are omitted for the short feature names in Table 4.2.

We expect strong redundancy between features since of course not all features represent different wavefield properties. For instance, polarization information is also implicitly included within the three-component frequency spectra. Furthermore, the probability is high that features of the same type, but computed for close frequency bands, are correlated. This problem is considered by applying automatic feature selection (Section 4.7).

4.2 Wald-Wolfowitz Runs Test

In the following sections we will use a simple, three-dimensional toy data set $\vec{x}_T = (X, Y, X)$ to demonstrate the introduced techniques. The data set consists of three clearly separated clusters defined by their centroids. The distribution within each cluster is Gaussian. Furthermore, a temporal context is assigned to each sample vector. Fig. 4.1 shows that the clusters are occurring one after another in time.

The nonparametric Wald-Wolfowitz runs test (Wald & Wolfowitz, 1940) can be used to assess the randomness of a two-valued time series by considering the distribution of runs. A "run" of a series is a maximal segment of adjacent equal elements. In general, any time series can be transformed into a two-valued one by considering whether a data item is smaller or larger than the median of the series. This process is called *median dichotomization* in statistics. In Fig. 4.1 this is shown for five time series. In particular, we use our toy data set $\vec{x}_T$ and two random sequences V (Gaussian) and W (uniform). Whenever the background coloring in Fig. 4.1 changes with time, a new "run" starts. In order to check whether a time series shows significant, non-random temporal patterns, we evaluate the test statistic of the runs test:

$$Z_{test} = \frac{r - \mathrm{E}[R]}{\sqrt{\mathrm{Var}[R]}}\,, \tag{4.1}$$

where R is a random variable corresponding to the number of runs of a random time

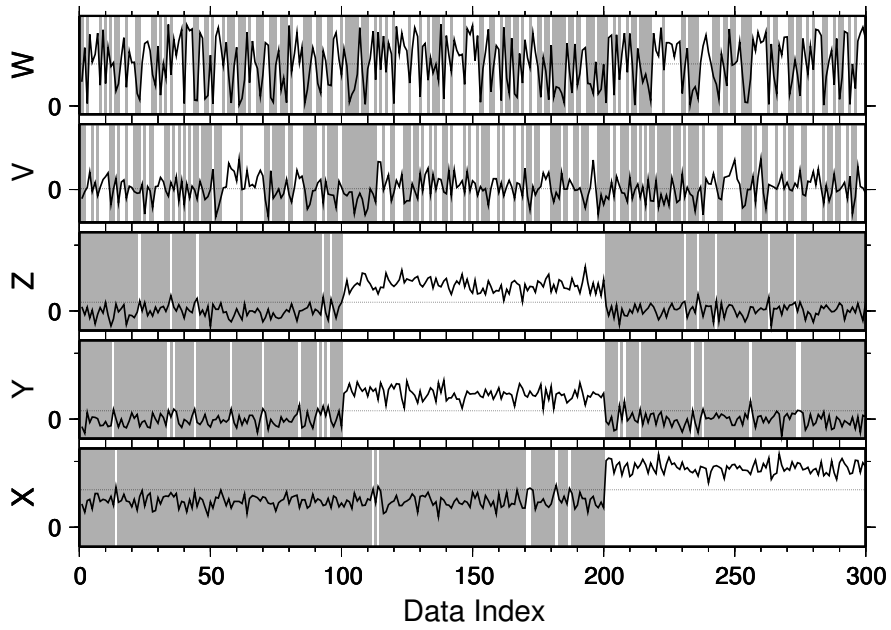

Figure 4.1: Example of the runs test for five time series. Horizontal lines correspond to median. Background colorings highlight values above and below median. Whenever coloring changes with time, a new "run" is starting.

series which has the same length N as the series under investigation. The variable r is the number of observed runs for that sequence. The mean,

$$\mathrm{E}[R] = \frac{2N^-N^+}{N} + 1 \;, \tag{4.2}$$

and the variance,

$$\mathrm{Var}[R] = \frac{2N^-N^+(2N^-N^+ - N)}{N^2(N-1)} \;, \tag{4.3}$$

of R is computed using the number of data items larger and smaller than the median for the observed time series (N^+ and N^-). Whenever the hypothesis of randomness is not rejected ($Z_{test} < 1.96$ for a significance level of 5%), or the number of observed runs r agrees with the expected value R, respectively, the corresponding sequence shows no significant patterns and, therefore, has no information content with respect to the goal of our approach. For instance, for a completely random time series (V, Gaussian), we would obtain $Z_{test} = 0.35$.

Application of the runs test allows the decision whether time series have information about non-random, temporal variations. Moreover, a relative significance ranking with respect to Z_{test} is yielded. The higher Z_{test}, the more likely is the non-random character of the time series. As we will show in the following experiment, this ranking of sequences of same duration N depends on the pattern length L and the number P of observed patterns. For that, we have to use a very simple model for a temporal pattern. We simply add an offset (here: 5) to a normally distributed base time series with zero mean and unit variance (like V) for L succeeding samples. For instance, depending on what is considered as the base series, feature Z of the toy data set shows one or two patterns with $L = N/3$.

Time series of different lengths are generated (20 between $N = 30$ and $N = 20000$). Each sequence is characterized by a number of occurring patterns ($P = 1, 5, 10, 50, 100$) and pattern length (9 different lengths). The onset times of patterns are randomly chosen. We repeat the time series generation 100 times for different onsets for each combination of N, L, and P. For each combination, the test statistic Z_{test} is computed and averaged for all 100 sequences. In Fig. 4.2 pattern length is indirectly shown by the ratio L/N, which corresponds to different colors on a gray scale.

There is a clear power law dependency (logarithmic scale) between number of samples N and Z_{test} for each ratio L/N, and all tested number of patterns. Furthermore, offsets are observed between curves having different ratios and number of patterns. The results for the pure random time series (triangles) are clearly below $Z_{test} = 2$ for all number of samples, thus confirming the chosen significance level of 5%. Furthermore, Z_{test} increases with length of pattern for a fixed number of samples. Hence, the longer the duration of a pattern (e.g. the number of time windows of a sequence with high linearity), the higher is the test statistic for the time series. Note that, as for our toy data set, what can be identified as patterns or base time series becomes ambiguous, when the pattern length is large compared to N. Furthermore, when an offset was added to the majority of samples, the patterns must be considered as the base series. This is the reason why we observe decreasing values for Z_{test} for high ratios L/N in Fig. 4.2c,d,e. Fig. 4.2 allows to inter- or extrapolate a rough estimate for Z_{test} for a given pattern length, number of patterns, and total sample number.

4.3 Cluster Algorithms

Within this study, two simple and widespread clustering techniques are employed to implement a generic and robust processing scheme. First, we use the k-means algorithm introduced in Section 2.3.1. In order to take into account the dependency on the initially-selected position of cluster centroids, five to 10 clustering runs are carried out using different, randomly-chosen locations for a given number of clusters k. The partition with the lowest sum of squared errors, given by Equation 2.2, is chosen as the final solution. Furthermore, we apply hierarchical clustering using all introduced linkage rules (see Table 2.1).

In order to obtain a particular partition of the data set, different numbers of clusters k have to be tested for both clustering methods. For hierarchical clustering, the k-splitting criterion is used to obtain a partition from the dendrogram, i.e. the cluster-tree is split for a given number of edges k. We test clustering between an appropriate k_{min} and k_{max} and choose the most favorable parameter based on quantitative validation and qualitative visualization (see discussion in Section 2.3.3). The employed visualization method and the computed cluster validity measures are introduced in Section 4.4 and 4.5.

4.4 Self-Organizing Maps

In Fig. 4.3 the SOM technique is demonstrated using the toy data set $\vec{x}_T$ (red symbols in Fig. 4.3a). The SOM learning algorithm combines vector quantization, i.e. the generation of prototype vectors (black symbols in Fig. 4.3a), and an ordered, non-linear mapping into a space of lower dimension (maps in Fig. 4.3a). Usually, SOMs are built on a regular

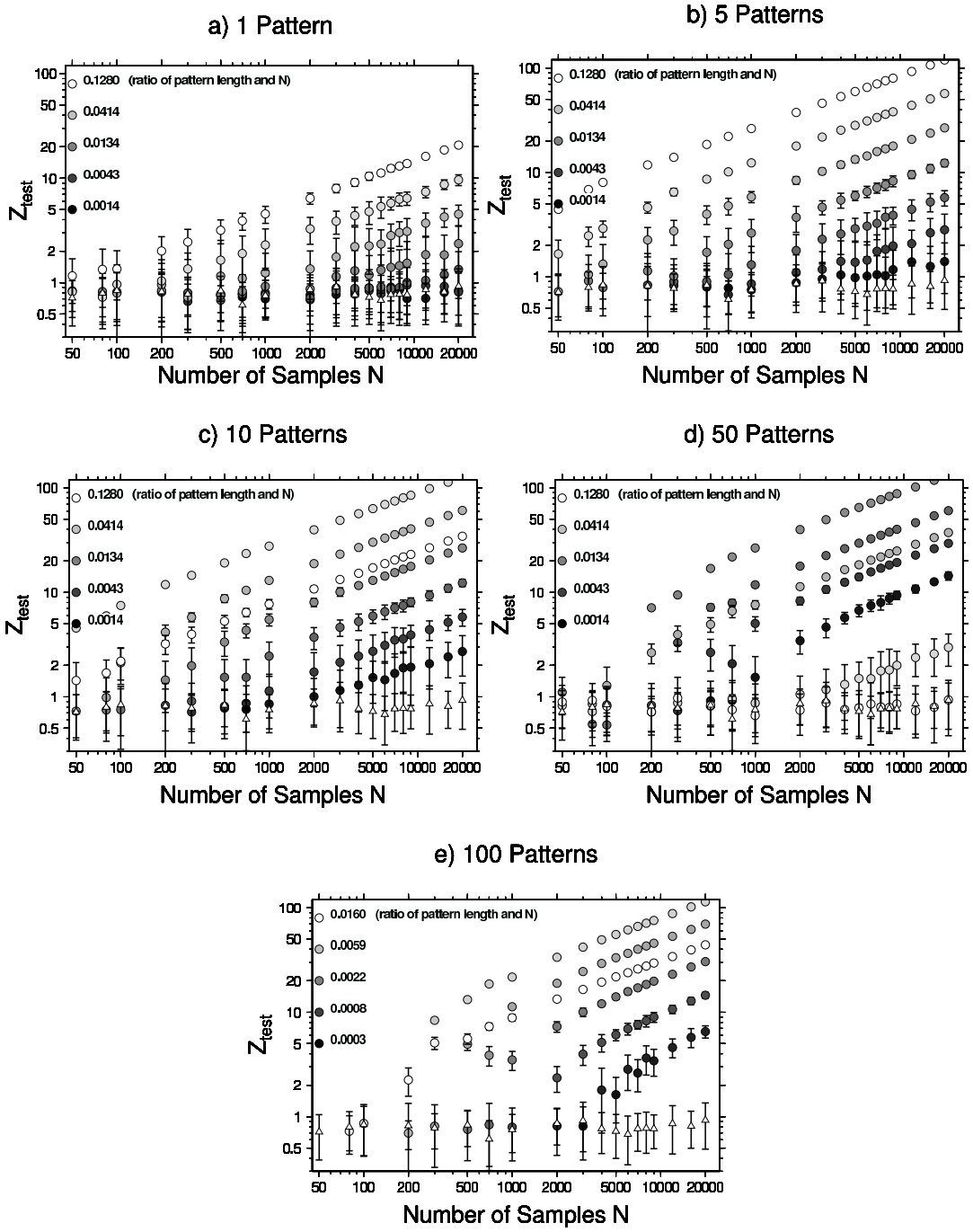

Figure 4.2: Calibration of the runs test. Each panel shows the computed test statistic Z_{test} for different number of patterns P added to a random base time series. Z_{test} is plotted over number of samples N. The coloring of symbols corresponds to different ratios L/N between pattern length L and N.

grid with hexagonal units. Each grid unit n is represented by a prototype vector $\vec{m}_n$. The number of SOM units M and the aspect ratio of the SOM depend on the amount of data and the corresponding ratio of the two largest eigenvalues. The following heuristic formula is employed (Vesanto et al., 2000):

$$M = 5 \cdot \sqrt{N} \ . \tag{4.4}$$

For each data sample $\vec{x}_i$ $(i = 1, ..., N)$ of dimension J, the closest prototype vector $\vec{m}_c$ can be found, where c is called the best matching unit (BMU):

$$c = \arg\min_n \{||\vec{x}_i - \vec{m}_n||\} \ . \tag{4.5}$$

For the distance, we use the (weighted) Euclidean norm:

$$||\vec{x}_i - \vec{m}_n||^2 = \sum_{j=1}^{J} w_j (x_{ji} - m_{jn})^2 \ , \tag{4.6}$$

where $\vec{w}$ is a weight vector giving the importance of individual vector components (features).

In Fig. 4.3a, the size of the red, overlaying hexagons symbolizes the frequency a SOM-unit is a BMU of a data sample after learning. At each learning step t, the prototype vectors in the neighborhood of unit c are moved towards a selected data vector $\vec{x}_t$:

$$\vec{m}_n(t+1) = \vec{m}_n(t) + \alpha(t) h_{cn}(t) (\vec{x}_t - \vec{m}_n(t)) \ , \tag{4.7}$$

where $h_{cn}(t)$ defines the a neighborhood function around unit c and $\alpha(t)$ is the learning rate, both decreasing with time. A common choice for $h_{cn}(t)$ is a Gaussian function:

$$\begin{aligned} h_{cn}(t) &= \exp^{-d_{cn}^2 2\sigma_t^2} \\ d_{cn} &= ||\vec{r_c} - \vec{r_n}|| \ , \end{aligned} \tag{4.8}$$

where σ_t is the neighborhood radius at learning step t and $\vec{r_n}$ are the 2D coordinates of SOM unit n.

As an alternative to the *sequential* implemention of SOM learning in Equation 4.7, we will use the so-called *batch mode* in this study:

$$\vec{m}_n(t+1) = \frac{\sum_{i=1}^{N} h_{cn}(t) \vec{x}_i}{\sum_{i=1}^{N} h_{cn}(t)} \ , \tag{4.9}$$

where all data vectors $\vec{x}_i$ are employed at each learning step t. In contrast, data index i is equal to t for the sequential implementation.

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 so-called unified distance matrix (U-matrix). For this purpose, each SOM-unit is divided into seven sub-units (for hexagons) which are colored according to the prototype vector distance of directly connected units (U-matrix in Fig. 4.3b). The center unit stands for the averaged distance. Here, a color scale from warm (red), meaning large distances, to cold (blue)

colors, meaning small distances, is used. Thus, regions of low data density are represented by warm, and regions of high density by cold colors. The U-matrix visualization therefore allows the identification and manual definition of clusters in the data space (3D for toy data) by simply considering the two-dimensional SOM as apparent from Fig. 4.3a and 4.3b.

However, for many applications, automatic clustering using standard algorithms such as k-means or hierarchical techniques is required since manual grouping is often subjective and inaccurate. As each SOM prototype vector itself can be regarded as a cluster centroid, these clustering algorithms can directly be applied to the set of all prototype vectors to obtain a final clustering of the data set (Vesanto & Alhoniemi, 2000). In this study, we use a hierarchical clustering approach applied to the SOM prototype vectors (Vesanto et al., 2000). The advantage of this compared to direct clustering of the original data set is that the computation time is significantly reduced without loss of information. Furthermore, the clustering can be visualized directly on the SOM by coloring each unit according to its cluster-membership (Fig. 4.3b). By comparison with the U-matrix, this allows for a fast and simple visual assessment of the clustering validity. For our toy data set, the number of found clusters clearly fits the structure of the U-matrix.

Alternatively, agglomerative clustering based on the SOM U-matrix was suggested by Vesanto & Sulkava (2002). In a first step, local minima of the SOM U-matrix are identified after smoothing over a given number of nearest neighbors NN. The minima are considered as prototypes of an initial clustering. Afterwards, SOM is successively flooded starting within the minima. The growth of the clusters is controlled by the distance between unclustered and clustered points using a given linkage rule (Table 2.1). Furthermore, the rule "closest" can be used where the BMUs of the clustered SOM units, with respect to the unclustered SOM, controls the cluster growth. For another option ("neighf"), the SOM area defined by the employed neighborhood function (Equation 4.8) is additionally considered to compute distances between points.

The cluster memberships of BMUs (e.g. their color on the SOM) can be used to label the original data. If the data set is a representation of a time series, the temporal occurrence of clusters can be presented together with the raw data. However, automatic clustering is not always reasonable, e.g. when the U-matrix suggests no clear grouping on the SOM. In that case, the position of the BMU on the SOM can be used instead to characterize the time series and highlight patterns in the data. For instance, a similarity coloring can be generated by spreading a color scale on top of the SOM according to the 2D position of each SOM unit. Thus, SOM units of similar prototype vectors have similar colorings. A slightly modified technique, which we will use in the following, also takes into account that similarity between neighbor prototype vectors varies over the SOM (Vesanto et al., 2000). Instead of the SOM unit coordinate, the corresponding prototype vector projection, given by the principal component axes of the two largest eigenvalues of all vectors, are used for the color scale (Fig. 4.3b). Moreover, also the color scale generated for the SOM clusters takes into account the cluster centroid similarities. Hence, applied to seismograms, time windows with similar wavefield features get similar colors.

In order to reduce redundancy in the data space, which is known as correlation hunting, the so-called component planes of a SOM can be used (CPs, first three panels in Fig. 4.3c: red stands for high values). A CP ($\vec{cp}$) is built on the trained SOM (M units), where each unit n is represented by a particular component (feature) i of the corresponding prototype

vector $\vec{m}_n$, i.e. for component i the CP is given by $cp_n = m_{in}$. The components of the absolute correlation matrix $\mathbf{A}$ between all CPs is defined as:

$$a_{ij} = \frac{1}{M} \sum_{n=1}^{M} \|m_{in} \cdot m_{jn}\|. \tag{4.10}$$

As proposed by Vesanto & Ahola (1999), the correlation matrix can be used as input data for the training of a second SOM on a rectangular grid. The data vector $\vec{x}_t$ is then defined as:

$$\vec{x}_t \hat{=} \mathbf{a}_{\cdot j}, \tag{4.11}$$

where $\mathbf{a}_{\cdot j}$ is a column of $\mathbf{A}$. The so-called component plane SOM (CP-SOM) can be used to visualize intuitively the correlation or similarity between components (last panel in Fig. 4.3c). Guérif et al. (2005) proposed a related, embedded feature selection method, where the features are weighted during SOM-training based on a simultaneously generated CP-SOM.

Correlated features can be grouped by clustering the CP-SOM. In this study, we apply the U-matrix-based clustering to the CP-SOM prototypes (Vesanto & Sulkava, 2002; Barreto, 2007) (coloring of base map in Fig. 4.3c). Obviously, two clusters are found on the CP-SOM for the toy data. Thus, since component Y and Z are correlated, only two features are necessary to find 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).

4.5 Cluster Validity and Performance Measures

We will now introduce the relative cluster validity criteria and external measures for the classification accuracy which are employed in this study. In particular, we are mainly interested in two problems. The first is to find the best fitting number of clusters k for hierarchical clustering and k-means. Furthermore, we want to compare a found clustering with a manually or theoretically labeled data set.

4.5.1 Relative Criteria

We use the Davies-Bouldin (DB) index, suggested by Davies & Bouldin (1979):

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left\{ \frac{D_i + D_j}{d_{ij}} \right\}, \tag{4.12}$$

where d_{ij} is the distance between cluster centroids i and j, and D is the average distance to the cluster centroid (mean) within a cluster. For each cluster i, the maximum ratio between within- and between-cluster distances is chosen considering all other clusters $j \neq i$. Finally, all computed ratios are averaged. Hence, the higher d_{ij} and the lower D_i and D_i (separated and compact clusters), the lower is the DB index. Therefore, the tested partition with the lowest index is the one best fitting an existing, natural data grouping. As suggested by Vesanto & Alhoniemi (2000), we use the DB index to find a meaningful number of clusters k on a SOM.

Another criterion, which has been employed within the context of feature selection, is the normalized scatter separability (Dy & Brodley, 2004). This index is obtained from scatter separability Sc, which is composed of the within-class scatter matrix $\mathbf{S_w}$, and the between class scatter matrix $\mathbf{S_b}$ given by:

$$\mathbf{S_w} = \sum_{i=1}^{k} \frac{N_i}{N} \mathrm{E}\{(\vec{X} - \vec{\mu}_i)(\vec{X} - \vec{\mu}_i)^T | \omega_i\} \,, \tag{4.13}$$

$$\mathbf{S_b} = \sum_{i=1}^{k} \frac{N_i}{N} (\vec{\mu}_i - \vec{M}_0)(\vec{\mu}_i - \vec{M}_0)^T \,, \tag{4.14}$$

$$Sc = \mathrm{trace}(\mathbf{S_w}^{-1}\mathbf{S_b}) \,, \tag{4.15}$$

where $\mathrm{E}\{\cdot\}$ is the expectation over all data $\vec{X}$ in cluster ω_i with N_i instances, $\vec{\mu}_i$ is the cluster centroid, and $\vec{M}_0$ the total mean of all N samples. In order to compare two partitions P_1 and P_2 obtained from two features sets F_1 and F_2 of different dimensions, both indices have to be normalized:

$$\begin{aligned} S(F_1, P_1) &= Sc(F_1, P_1) \cdot Sc(F_2, P_1) \\ S(F_2, P_2) &= Sc(F_2, P_2) \cdot Sc(F_1, P_2) \,. \end{aligned} \tag{4.16}$$

The feature subset with the largest S generates the best clustering.

Another cluster validity criterion is cluster stability (CS), suggested by Lange et al. (2004). Here, the best clustering (e.g. with respect to k) is the one most stable regarding missing data or reproducibility for other data from the same source. For this purpose, the data set is randomly divided into two sections, $\vec{X^1}$ and $\vec{X^2}$, of same size. This is repeated 10 times. Both data sets are clustered within each fold into partitions $\vec{C^1}$ and $\vec{C^2}$, where C_i^1 is the cluster membership of X_i^1. Within this work, this phase corresponds to SOM learning and clustering of prototypes. Furthermore, a third grouping is generated by classifying $\vec{X^2}$ using the clustering $\vec{C^1}$, i.e. by finding the BMUs of $\vec{X^2}$ on the clustered SOM trained with $\vec{X^1}$. The clustering solutions are compared over the distance d_H using the Hamming distance function H:

$$d_H = \mathrm{E}\Big\{ \min_{\pi} \frac{1}{N} \sum_{i=1}^{N} \mathrm{H}\{\pi(\phi(X_i^2)), C_i^2\} \Big\} \,, \tag{4.17}$$

where ϕ is the classifier with respect to partition $\vec{C^1}$, $\mathrm{E}\{\cdot\}$ is the expectation over all 10 folds, and $\mathrm{H}(a, b) = 1$ if $a \neq b$, and zero otherwise. Since cluster labels of two partitions do not have to be in the same order, the function π tests all possible combinations. In our implementation we simplify this search for the correct distribution of cluster labels by simply comparing the best matching cluster centroids of both clusterings. Finally, the result is normalized:

$$CS = \frac{d_H}{d_H^{rand}} \,. \tag{4.18}$$

The normalization coefficient d_H^{rand} is obtained from Equation 4.17 for random clusterings (i.e. random labeling of data) of the same data set and same k. The smaller CS, the more stable is the clustering for a given number of clusters.

4.5.2 External Criteria

If theoretical data labels exist, external criteria can be computed. The simplest measure is misclassification rate defined as the ratio of misclassified and all data instances. In this study, classification of data is done by finding the class memberships of the BMUs of unlabeled data on a previously-labeled SOM. Therefore, the first step is to determine a class label for each SOM unit or SOM cluster. For this purpose, a labeled data set is projected on the SOM (finding BMUs). The most frequent class label is assigned to each SOM unit or SOM cluster, depending on whether classification should be done based on SOM prototypes or clusters. The number of ambiguous units or clusters M_{amb} (same number of BMU hits for two or more classes) is counted.

In Section 5.1 (synthetic data) we compute a single class error as the ratio of misclassified data, i.e. where the observed cluster labels $\vec{C}^o$ do not match with theoretical time window labels $\vec{C}^t$), and the total number of samples employed for testing. This can be again formulated using the Hamming distance:

$$CE = \frac{1}{N} \sum_{i=1}^{N} \mathrm{H}(C_i^o, C_i^t) \ . \tag{4.19}$$

For real data sets, where the number of instances within each class need not be similar, a modified classification error is computed to ensure that each class has a similar weight. First, false positive CE_j^+ (classified as j but presented as other) and false negative classification errors CE_j^- (presented as j but classified as other) are computed for individual classes j:

$$CE_j^+ = \frac{1}{N} \sum_{C_i^o = j} \mathrm{H}(C_i^o, C_i^t) \ . \tag{4.20}$$

$$CE_j^- = \frac{1}{N} \sum_{C_i^t = j} \mathrm{H}(C_i^o, C_i^t) \ . \tag{4.21}$$

If a class is not present on the SOM after labeling, CE_j is set to 1. Finally, the mean over all classes CE_{av} is penalized by the ratio between the number of ambiguous M_{amb} and all SOM units or clusters:

$$CE = CE_{av} + (1 - CE_{av}) \cdot \frac{M_{amb}}{M} \ . \tag{4.22}$$

In the following, results for CE are given in percent. Furthermore, since theoretical labels cannot be considered as the absolute ground-truth for most applications within this study, the classification error should be considered rather as a relative measure. Labels are manually chosen for real data or are produced by a simplification of the input model for synthetics. For instance, thresholds are used to deal with superposition and continuous transition between seismic signals.

Normalized mutual information (NMI, Strehl, 2002) compares two groupings with number of clusters k_i and k_j:

$$NMI = \frac{\sum_{i=1}^{k_i} \sum_{j=1}^{k_j} N_{i,j} \ \log(\frac{N \cdot N_{i,j}}{N_i \cdot N_j})}{\sqrt{(\sum_{i=1}^{k_i} N_i \ \log \frac{N_i}{N})(\sum_{j=1}^{k_j} N_j \ \log \frac{N_j}{N})}} \ , \tag{4.23}$$

where N is the number of data points, N_i is the number of instances in cluster i, and $N_{i,j}$ is the number of instances within cluster i as well as in group j. We use NMI as an external criterion and compare an obtained clustering (i) with the grouping given by the theoretical class labels (j). In contrast to computing CE, comparison is done without labeling all obtained clusters with the most frequent class label within each category. Hence, k_i and k_j need not necessarily to be identical. If the clustering and theoretical class labels match perfectly, we obtain $NMI = 1$.

In Fig. 4.4, DB, CS, CE, and NMI are plotted over number of clusters using the SOM of the toy data set. Both relative criteria show a minimum for the real number of clusters ($k = 3$). While the difference compared to $k = 2$ is more pronounced for the DB index, CS is less vulnerable to overfitting as expected. In contrast, classification error CE shows a clear overfitting behavior which could be avoided by cross-validation. On the other hand, for the NMI measure, a clear maximum at $k = 3$ is obtained.

4.6 Principal Component Analysis

A feature vector $\vec{X}$ is considered as a random variable with zero mean. The eigenvalues λ_j of the covariance matrix $\mathbf{\Sigma}$ of $\vec{X}$ are represented in the diagonal matrix $\mathbf{\Lambda}$ in descending order. The corresponding eigenvectors build the orthogonal matrix $\mathbf{\Gamma}$, which leads to the following equation:

$$\mathbf{\Lambda} = \mathbf{\Gamma^T \Sigma \Gamma} \; . \tag{4.24}$$

The linear transformed feature vector is obtained by:

$$\vec{Y} = \mathbf{\Gamma^T \tilde{X}} \; . \tag{4.25}$$

It can be shown that the variance of the principal components Y_j is λ_j. Hence, Y_1 has the biggest proportion of the total data variance. Thus, in order to select the components explaining most of the data variance, a threshold can be introduced for dimensionality reduction.

4.7 An Unsupervised Feature Selection Procedure

In the previous sections we introduced different techniques which we will now combine for an unsupervised feature selection procedure. In order to keep significant features and reduce redundant information for a feature set generated by different approaches, we propose a three-level feature selection approach which iteratively reduces the number of features. The processing flow is illustrated in Fig. 4.5. In the first level we chose potential feature candidates by assessing the information content of each feature individually (relevancy filter), while in the second and third level, dependencies between features are considered (redundancy filter). In the next sections we discuss each level in more detail.

4.7.1 Level 1: Within Individual Features

In this level we first compute three criteria for each feature:

- Ratio R_{obs}/R_{exp} between observed variability $R_{obs} = \max(F) - \min(F)$ and reasonably expected range R_{exp} of a feature F derived theoretically from physical or data processing parameters.
- Wald-Wolfowitz test statistic Z_{test} (Equation 4.1).
- Lowest DB index (Equation 4.12) computed from a 1D k-means clusterings allowing 2 to N_{clus} clusters (e.g. N_{clus}=5).

The first two criteria are used to exclude features. We reject those features providing no significant discrimination between time windows due to small observed ranges ($R_{obs}/R_{exp} < r_{limit}$, e.g. $r_{limit} = 0.1$), and which show no significant temporal patterns ($Z_{test} < Z_{limit}$). For instance for the ellipticity, the natural limits (0 and 1) imply $R_{exp} = 1$. Furthermore, for the instantaneous frequency $R_{exp} = f_{max} - f_{min}$ follows from the lower (f_{min}) and upper frequency limit (f_{max}). For the (normalized) amplitude features from generation Methods 5, 6 and 7, no physical limits can be given. Therefore, we set r_{limit} equal to zero (accepting all features). As mentioned in Section 4.2, $Z_{limit} = 1.96$ is an appropriate threshold for the runs test. However, if the duration of expected temporal patterns is longer, increasing this value may improve performance.

The DB index is computed to assess the cluster tendency of the feature. This criterion is used together with Z_{test} to rank features in the next levels. For a discussion on the sensitivity of parameters N_{clus}, r_{limit} and Z_{limit}, see Chapter 5.

4.7.2 Level 2: In-between Features of Individual Subsets

In the second level, we consider seven feature subsets corresponding to the different feature generation methods given in Table 4.1. Only features accepted by Level 1 are used. We first learn a SOM and subsequently a CP-SOM for each subset. Finally, the CP-SOM clustering is applied. From each CP-SOM cluster, the features with the lowest DB index and the highest test statistic Z_{test} are chosen as representative features for the particular cluster. Thus, we keep features with best cluster tendency and most significant temporal patterns. If both selected features have the same BMU on the CP-SOM, only the one with the highest Z_{test} value is selected.

4.7.3 Level 3: In-between all Remaining Features

From Level 2 we obtain a reduced subset for each feature generation method. In the third level, we learn a single SOM and CP-SOM, combining all subsets in order to assess correlations between methods. The splitting of correlation hunting into two levels is done to ensure a similar importance for each feature generation method since each approach has a different number of features. Finally, we choose the features from the CP-SOM clusters as in Level 2. The obtained set of features is then used for further processing, i.e. to learn the final SOM and to cluster the data set.

4.7.4 Simple Example

In Table 4.3 we demonstrate our feature selection procedure using a simple data set of five features ($N = 300$). Values for X, Y and Z, together defining three clusters, correspond

to our toy data set. Furthermore, the random features V and W from Section 4.2 are used. The temporal context of all features is given in Fig. 4.1. We omit the range test in Level 1 and only use a single subset (no Level 3) because the features have no physical background.

Table 4.3: Results of our feature selection method applied to the toy data set and two random feature time series.

Feature	X	Y	Z	V	W
Observed runs r	70	77	78	149	154
Runs test statistic Z_{test}	9.37	8.56	8.44	0.23	0.35
DB Index	0.35	0.46	0.47	0.63	0.56
Selected after Level 1	yes	yes	yes	no	no
Index of CP-SOM cluster	1	2	2	-	-
Selected after Level 2	yes	yes	no	-	-

Features V and W are correctly rejected by the runs test ($Z_{test} < 1.96$) because of their temporal randomness. The result of CP-SOM clustering is shown in Fig. 4.3c. Features Y and Z belong to the same CP-cluster. Thus, features X and Y, the second one due to its higher runs test statistic Z_{test} and lower DB index, are finally selected.

Table 4.4: Results of a wrapper feature selection algorithm applied to the toy data set and two random feature time series. Best feature subset (bold) is found when normed scatter separability criterion S for the next subset becomes smaller than the previously tested: $\text{sign}\,(S - S_{prev}) < 0$

	$N_{clus} = 10$		$N_{clus} = 3$	
Search Step	Feature Subset	$\text{sign}\,(S - S_{prev})$	Feature Subset	$\text{sign}\,(S - S_{prev})$
1	**V**		X	
2	V, W	-1	X, Y	+1
3	V, W, X	-1	**X, Y, Z**	**+1**
4	V, W, X, Y	+1	X, Y, Z, W	-1
5	V, W, X, Y, Z	-1	X, Y, Z, W, V	0

We also test a wrapper approach for feature selection (Dy & Brodley, 2004) and show the results in Table 4.4. The forward search based on the normalized cluster scatter separability criterion S (Equation 4.16) results in a feature subset including only V ($N_{clus} = 10$) or X, Y and Z ($N_{clus} = 3$), respectively. Thus, the random features are correctly rejected only for the second run. However, no redundancy reduction is obtained and the maximum number of clusters has to be limited to avoid overfitting.

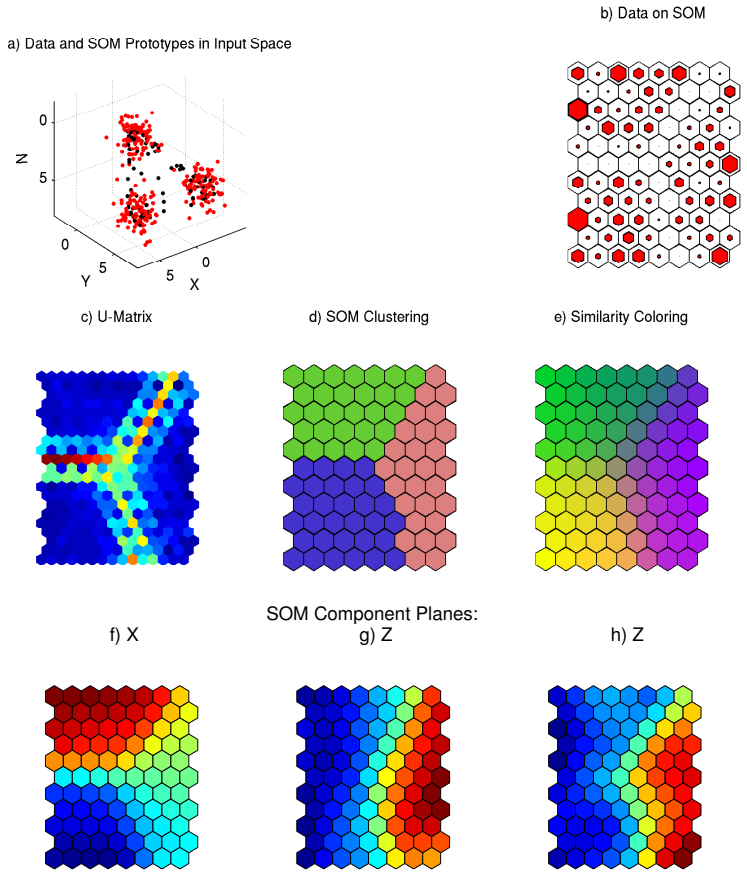

Figure 4.3: Illustration of an application of Self-Organizing Maps and related techniques to a simple toy data set of three features. In a) red symbols (circles in input space and hexagons on SOM) correspond to data, where size on SOM symbolizes the frequency a SOM-unit is a best matching unit (BMU) of a data sample. Black symbols correspond to prototypes. In b) several SOM visualizations are shown. Panel c) presents component planes (CP) for each feature. CP-SOM shows similarity and grouping (background coloring) of features.

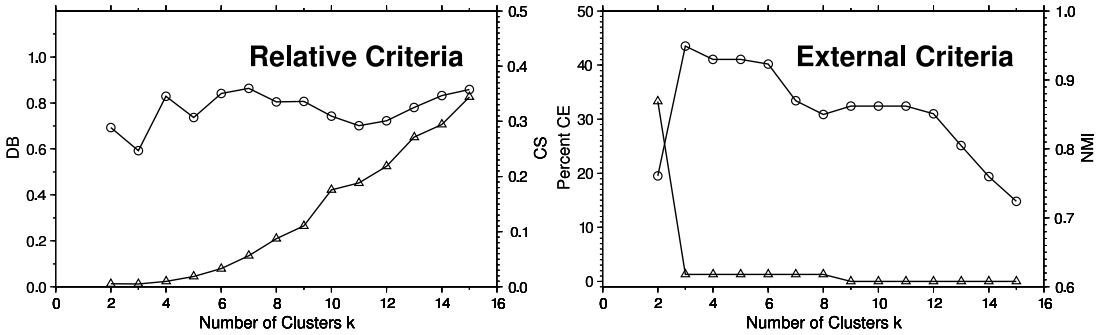

Figure 4.4: Internal cluster validity criteria (left) and classification performance (right) for different numbers of clusters. Davies-Bouldin index (DB, circles) and cluster stability (CS, triangles) do not make use of theoretical labels. Classification error (CE, triangles) and Normalized Mutual Information (NMI, circles) compare clustering with given class labels.

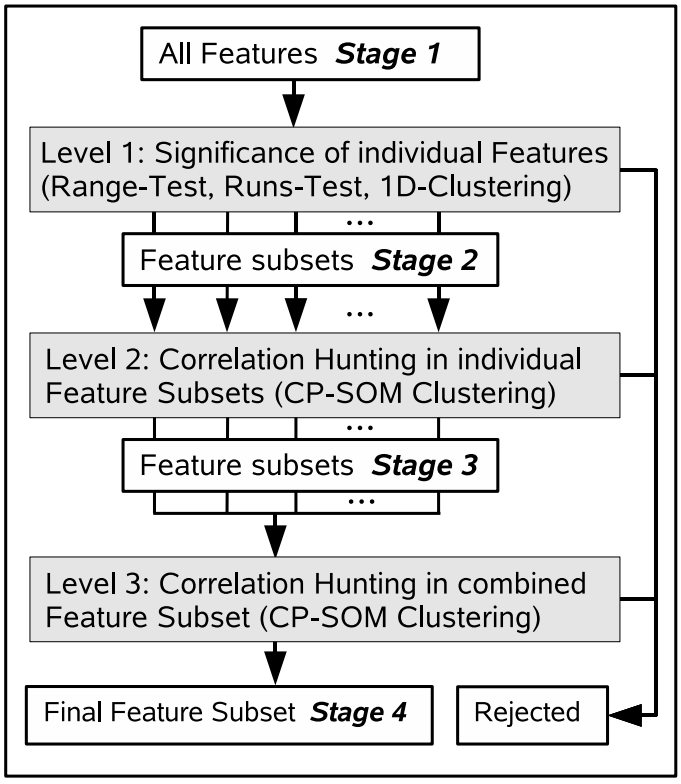

Figure 4.5: Three-level feature selection procedure. Stages 1-4 correspond to different feature subsets after or before particular processing steps. Feature subsets at Stage 2 and 3 correspond to different feature generation methods.

Chapter 5

Experiments

In the previous chapter, we have introduced unsupervised feature selection, visualization, and clustering techniques for the analysis of seismic wavefield recordings. In this chapter, these approaches will be applied to several data sets with emphasis on different objectives. We conduct experiments using synthetic seismic data as well as real-world recordings. Synthetic data is used to validate the techniques and show the potentials of unsupervised learning in seismology. We employ three kinds of real-world data sets. Regional earthquake recordings are used to find features for the detection and classification of seismic phases and to investigate the natural discrimination between. Additionally, further sensitivity tests are carried out. Furthermore, clustering is tested for its potential for discrimination of volcanic signals and background noise. Finally, we study the clustering of ambient seismic vibration data and the potential to improve existing techniques for estimation of dispersion curves and H/V spectra.

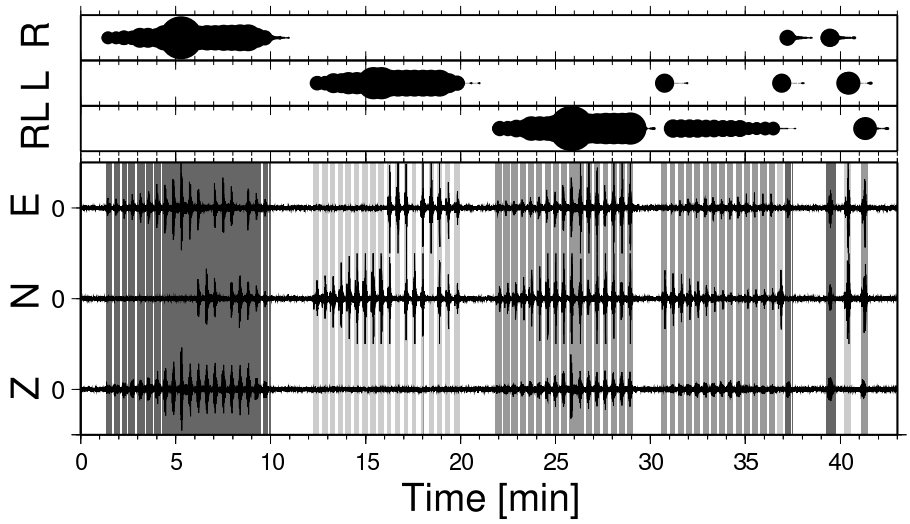

Figure 5.1: Three-component seismogram for array center station of data set 1. On top source history (R: Rayleigh wave, L: Love wave, RL: Mixture of both) and relative amplitudes are given. Background color corresponds to time window labels (R: dark gray, RL: gray, L: light gray).

5.1 Synthetic Data

In order to assess the reliability of our methodology, check the sensitivity of its parameters, and demonstrate a potential application, we generate two synthetic data sets of different complexity using a sampling rate of 50 Hz

For the first data set (data set 1, Fig. 5.1), we use a simple surface source setting that generates clearly separated wave packages of Rayleigh waves (vertical point force), Love waves (point force tangential 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). For the tangential point forces, a minor contribution of Rayleigh waves is not avoidable on receivers not lying on the connecting line between source and array midpoint, due to small, non-zero, radial force components. The maximum contribution on the most distant station is about 13% of the Love wave amplitude. For the majority of receivers, contribution is clearly less. Since we average over all stations for feature computation, we can assume that Love waves are dominating.

As the intention of this study among others is related to the investigation of ambient vibration wavefields, the character of such source setting is simulated for the second data set (data set 2, Fig. 5.2). A number of about 50 point sources, randomly distributed (normal) in time and space (surface sources only), are employed (distance range: 2 - 6 km). The orientations of point forces are randomly chosen, either vertical, tangential or arbitrary, all with random amplitudes.

For both data sets, the seismic velocity model is a deep basin structure, and the receiver configuration consists of 12 stations with an aperture of about 400 meters (Fig. 5.3). Fig. 5.3 also shows the receiver boxes (sub-arrays) for feature averaging (see Section 4.1). For the two lowest frequency bands, time lags allow averaging over all stations (dashed box). However, for the highest frequency band, splitting into five boxes, from which we use only

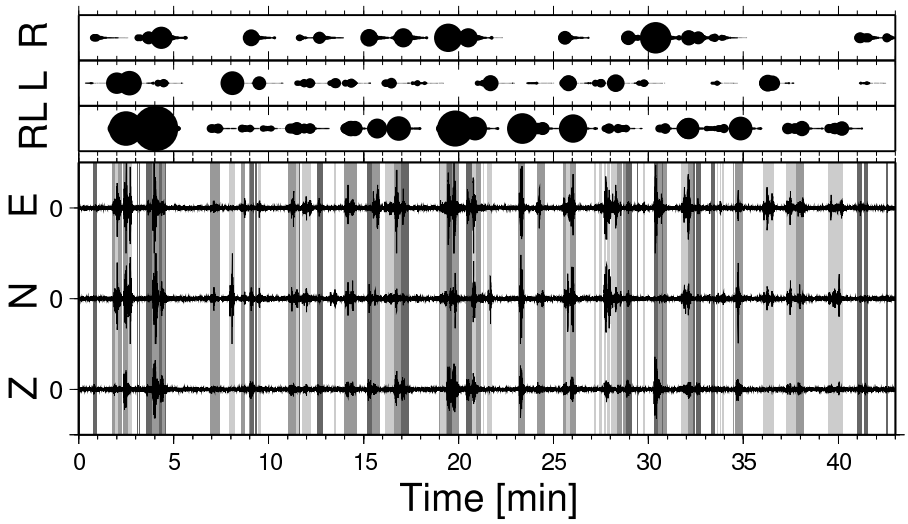

Figure 5.2: Three-component seismogram for array center station of data set 2. On top source history (R: Rayleigh wave, L: Love wave, RL: Mixture of both) and relative amplitudes are given. Background color corresponds to time window labels (R: dark gray, RL: gray, L: light gray).

one in the following, is required. For more details on the velocity model see the Appendix (Table 8.2) and Köhler et al. (2007). Since we employ the mode-summation technique of Herrmann (2001) and sources on the earth's surface, the fraction of generated body waves is negligible. Finally, we add white noise to the waveforms with a variance corresponding to 1% of the maximum receiver amplitude. The time window length for further processing is 3.7 seconds ($WINFAC = 5$, $f_{cent} = 1.34$ Hz), which results in 691 time slices or feature vectors. The frequency bands for features generation are located between 0.24 and 18 Hz (Appendix, Table 8.2), which is the range of dominant energy.

For quantitative validation we need the ground-truth information for each time window. Hence, it is necessary to label the data according to the theoretical source setting. For this purpose, we define four classes: pure Rayleigh waves (class 1), pure Love waves (class 2), mixture of Rayleigh and Love waves (class 3) and pure, random noise (class 4). The amplitudes, travel time, and duration (non-zero amplitudes) of a signal, with respect to a given source-receiver distance, is extrapolated from a previously-evaluated seismogram section for single sources. Furthermore, time windows with signal amplitudes lower than 2% of the maximum amplitude are assigned to the noise class. For data set 2, overlapping between classes occurs. The corresponding time windows are assigned to class 3 in the following. However, whenever the vertical force component dominates compared to the horizontal components, class 1 is chosen (horizontal represents 10% of vertical amplitude). Fig. 5.1 and 5.2 show the occurrence of sources on top and the time window class labels (class 1 - 3) in the background of the seismograms. The sizes of the symbols on top correspond to the relative receiver amplitudes.

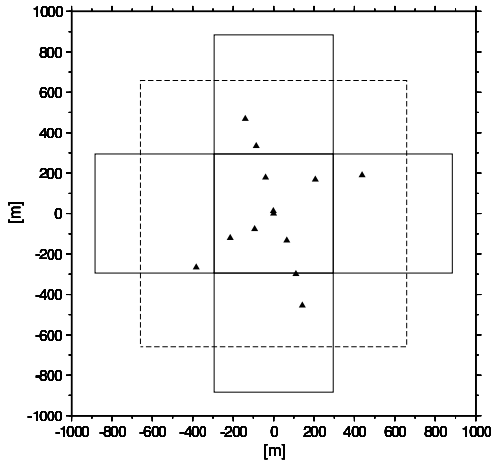

Figure 5.3: Receiver array geometry for synthetic data sets. Boxes correspond the station grouping for feature averaging (see Section 4.1).

5.1.1 Data Set 1: Evaluation of Techniques

Data set 1 is used for an extensive study on algorithm parameters and reliability of unsupervised learning with SOMs. We first evaluate our feature selection strategy and compare results with PCA. Afterwards we conduct investigations on SOM training and clustering using the finally selected feature subset.

Evaluation of Feature Selection: Assessing the Improvement

For the following computations, we use empirically found or default algorithm parameters listed in Table 5.1 as a start. Variation of the parameters, including those for feature selection, SOM learning and clustering, is discussed within the next sections.

In order to assess the validity and performance of our feature selection procedure, we apply a 10-fold cross-validation (CV) technique following the approach of Dy & Brodley (2004). Validation is based on the obtained cluster membership for each SOM prototype vector. As discussed in the previous chapter, hierarchical clustering with a k-splitting criterion is employed, where the number of clusters, between k_{min} and k_{max}, is chosen according to the lowest Davies-Bouldin (DB) validity index. 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 processing, including CP-SOM clustering for feature selection (Level 2 and 3) and final clustering, the remaining section serves as a test or validation data set. The selection of the test data set is repeated 10 times, so that each section is used for validation once. Note that the (ground-truth) labels of the training data set are not used in the training phase, i.e. for feature selection, SOM learning, and SOM clustering. This information is used to assess the performance of the procedure after training. Each cluster is classified with respect to the most frequent label within each group, which is given by projecting the theoretical class labels of the training data on the SOM (see Section 4.5.2).

Level 1 of the feature selection procedure is applied to the complete data set (training and test data) in order to keep the temporal context for the runs test. Rejecting random

Table 5.1: Empirically found or default algorithm parameters for the unsupervised learning techniques.

Feature Selection							
N_{clus}	r_{limit}	Z_{limit}	*mapsize*	*linkage*	NN		
5	0.1	1.96	big	average	1		
Final SOM learning and Clustering							
mapsize	*algorithm*	*linkage*	*lattice*	*norm*	*kernel*	k_{min}	k_{max}
big	batch	average	rect.	logistic	Gaussian	3	15

features or determination of the ranking, respectively, is conducted before the random data permutation for CV. Thus, CV only concerns SOM-related learning. It would be possible to apply the runs test to the training data set without permutation. However, for synthetic data set 1 this is not advisable since there would be no uniform distribution of classes within each fold. We refer to Section 5.2, where CV is applied to real-world data including Level 1.

In the testing phase we compute the BMUs, and thus the cluster-memberships, of the test data set on the training data set SOM. The class error is computed for each fold as the percentage of misclassified data with respect to the total number of samples of the test data set (see Equation 4.19). Since there is a tradeoff between the found number of clusters and the classification error, the final cross-validated classification error ($CVCE$, Dy & Brodley (2004)) is presented as the mean of all individual fold errors. The mean can be used because fold errors are approximately normal distributed (See Appendix, Fig. 8.2).

It should be noted here, that we do not expect to achieve a $CVCE$ tending to zero since the transition between seismic wave types and noise can be smooth, although we introduced a threshold to obtain the theoretical class labels. Furthermore, the results show similarities between clusters belonging to class 1 and class 3 (see Fig. 5.6 and 5.8). This is expected due to changing dominance of Rayleigh waves in the mixture wavefield, depending on distance and force orientation.

In order to quantify the improvements made by our new feature selection approach, we first compute $CVCE$ for several feature subsets obtained at four stages of our procedure (Fig. 4.5) and for particular feature generation methods (see Table 4.1). The results for $CVCE$, standard deviation, and number of features (averaged number for Stage 3 and 4) are summarized in Table 5.2.

For each feature generation method, the classification errors are similar within their uncertainties for all stages of feature selection. Therefore, no classification accuracy is lost for decreasing number of features. Moreover, there is a clear decrease of the error for Method 2 (more than one standard deviation). Furthermore, the $CVCE$ decreases significantly (about two standard deviations) when all feature generation methods are combined

Table 5.2: Results of CV for data set 1. Cross-validated Classification Error ($CVCE$) and averaged number of features for different stages of feature selection and different feature generation methods are given (see Fig. 4.5 and Table 4.1).

Method	1	2	3	4	5	6	7	all
Percent $CVCE$								
Stage 1	24.4	46.1	28.0	21.5	24.2	33.8	28.1	17.1
	±5.7	±8.6	±7.5	±5.1	±6.2	±4.9	±9.3	±3.8
Stage 2	24.4	46.1	28.0	18.3	24.2	33.8	28.1	18.1
	±5.7	±8.6	±7.5	±5.2	±6.2	±4.9	±9.3	±5.0
Stage 3	25.4	35.5	26.8	25.7	24.1	41.9	26.1	14.8
	±5.8	±6.1	±4.5	±4.6	±8.2	±5.6	±4.3	±2.5
Stage 4	-	-	-	-	-	-	-	15.7
	-	-	-	-	-	-	-	±4.3
Averaged Number of Features								
Stage 1	9	18	39	42	25	20	9	162
Stage 2	9	18	39	40	25	20	9	160
Stage 3	5.0±0.0	7.8±1.1	12.7±1.5	13.4±1.4	10.1±1.5	9.0±1.6	4.4±0.7	62.4±2.1
Stage 4	-	-	-	-	-	-	-	24.9±3.2

at each stage compared to the individual feature subsets. Compared to all methods, the spectral features (Method 5) provide the best discriminative power for clustering, since lower misclassification rates are obtained. The best overall performance (14.8%) is achieved with about 62 features from all methods at Stage 3. However, after assessing correlation between all feature generation methods at Stage 4, the $CVCE$ is still within the range of standard deviations of Stages 1 to 3, and slightly lower than taking all features. Due to the relatively simple synthetic wavefield, most features show significant patterns and are therefore accepted in feature selection Level 1. However, assessing the correlations between features in Level 2 and 3, significantly reduced the set of features for all feature generation methods. This reduction does not significantly increase the misclassification rates, except for feature generation Method 6 (spectrum of polarization ellipsoid).

Fig. 5.4 visualizes the redundancy reduction in Level 3 by means of the CP-SOM for one fold (Level 2 is shown in the Appendix Fig. 8.3). It becomes clear that similar CPs, and thus correlated features, are grouped within each CP-SOM cluster, and an appropriate and representative feature set is finally obtained.

In Fig. 5.5 the results for $CVCE$ are summarized for different stages of our feature selection procedure. Furthermore, we show results for a stepwise reduction of the final feature set (right hand-side of vertical dashed line in Fig. 5.5). For each step the feature with the lowest Z_{test} is omitted. The final feature set (before reduction) is obtained with the complete data set. In other words, CV is only conducted for final SOM training and clustering. Additionally, as an alternate performance measure, the normalized mutual information estimates (NMI, Equation 4.23), averaged over all CV folds, are plotted for the same stages or number of features. For the computation of NMI, we compare the ground-truth class labels of the data (four classes) and the found grouping of k clusters (see Section 4.5.2).

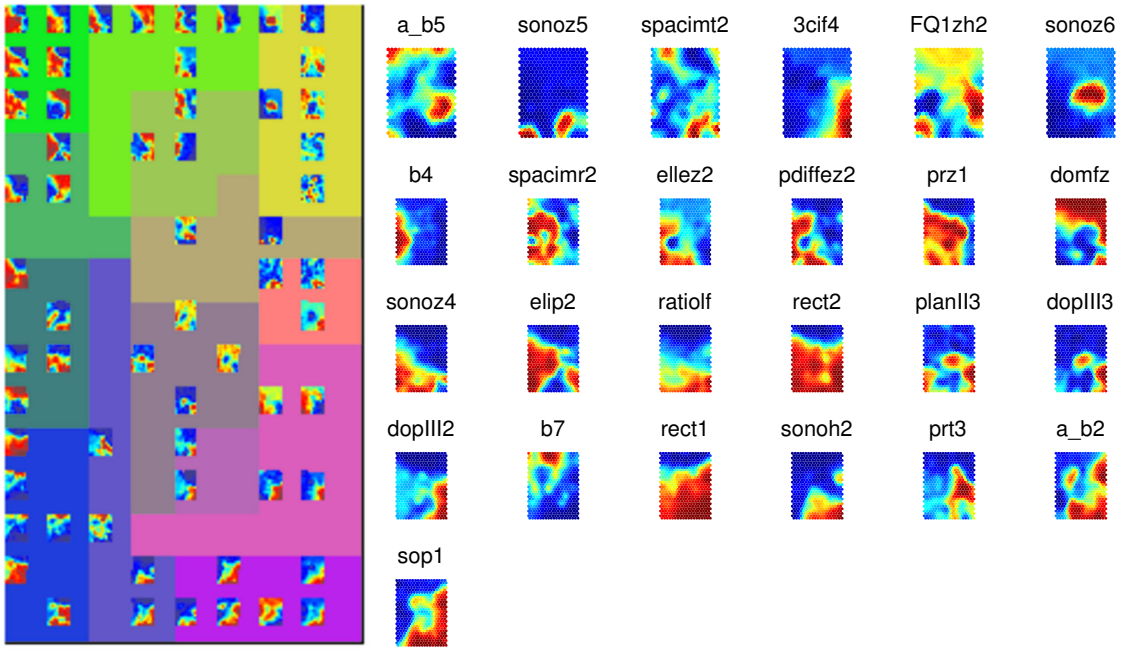

Figure 5.4: CP-SOM for feature selection Level 3 (all generation methods, left hand) and component planes for final feature set (right hand).

Taking into account the standard deviations, $CVCE$ as well as NMI show similar classification performance for decreasing number of features down to a number of about 15 to 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 ($CVCE$) or to decrease performance (NMI). 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.

In Section 4.7 we suggested to carry out a two-level correlation hunting in order to give a similar importance to different feature generation methods. In Table 5.3 the result for a single redundancy filter level is shown (2 level FS). The corresponding $CVCE$ does not change significantly compared to the three-level approach (3 level FS). However, a comparatively large feature set is obtained. Since we know from Fig. 5.5 that significantly less features are sufficient to achieve similar classification rates, the two-level correlation hunting strategy confirms our expectation.

Since we obtain similar results for the complete and the automatically selected feature set, the question arises whether a randomly-selected feature set would perform equally well. This set can be considered as a “baseline” quality for feature selection. For each CV fold we randomly chose 25 features out of the complete set without repetitions. The

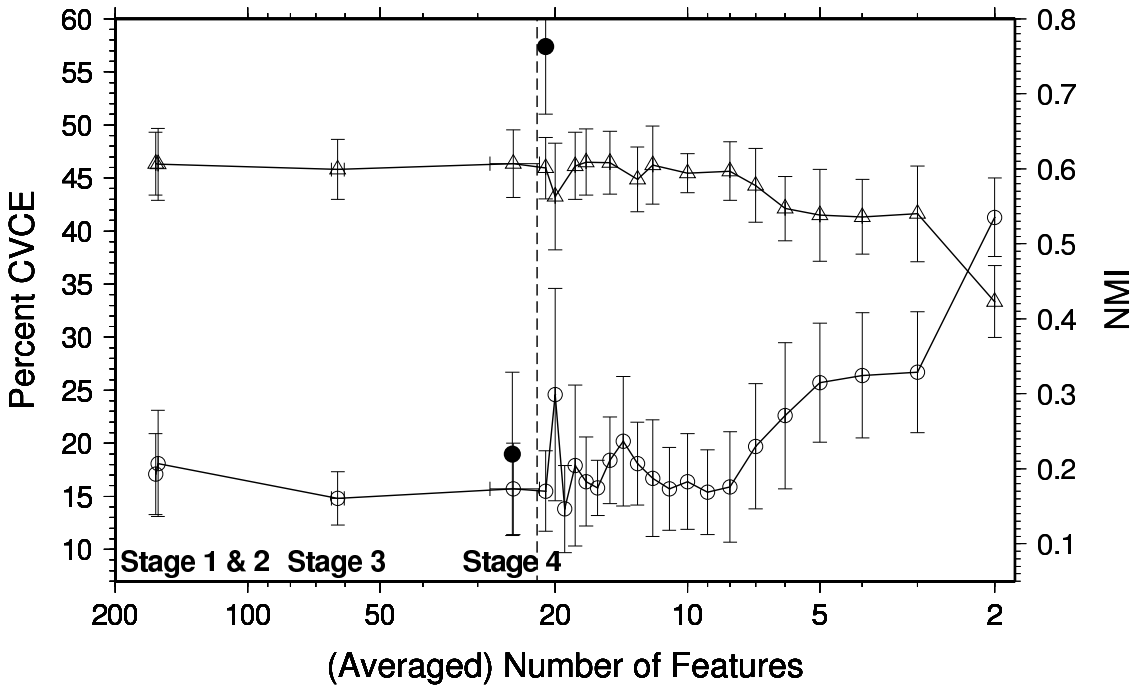

Figure 5.5: CV results: cross-validated class errors ($CVCE$, circles) and averaged normalized mutual information (NMI, triangles) for different number of features. Vertical dashed line marks size of automatically selected, final feature set. Filled symbols close to dashed line correspond to feature selection baseline (random feature set, on bottom) and clustering baseline (random labeling, on top).

achieved error in Table 5.3 (random FS) is located at the upper bound of the uncertainty of our procedure (3 level FS). Furthermore, the standard deviation is significantly higher for the random selection. Therefore, we observe a clear evidence that the improved results (3 level FS) are not just obtained due to the decreased number, but rather due to the correct ranking and grouping of features.

Evaluation of Feature Selection: Sensitivity of Parameters Level 1

We investigate the sensitivity of algorithm parameters by applying the same CV procedure as in the previous section. Except for one, we keep all parameters constant, as given by Table 5.1, in order to investigate the effect on the classification accuracy.

Table 5.3: CV results for different feature selection variants: Our suggested feature selection method (3 level FS), using only one level for correlation hunting (2 level FS), and randomly selecting a feature set (random FS).

	3 level FS	2 level FS	random FS
Percent $CVCE$	15.7±4.3	15.9±3.2	19.0±7.7
Number of Features	24.9±3.2	49.8±3.7	25

Table 5.4: Results for $CVCE$ and number of features for algorithm parameter variation in feature selection Level 1. Significance levels for $Z_{limit} \geq 10$ are very small and, therefore, approximate zero. It is save to say that the probability of the corresponding feature time series to be random can be ruled out.

N_{clus}	2		5		10		15	
Percent $CVCE$	17.4±4.2		15.7±4.3		17.1±3.7		16.2±4.0	
No. of Features	24±2.1		24.9±3.2		24.6±1.9		23.6±1.6	
r_{limit}	0.0		0.1		0.2		0.5	
Percent $CVCE$	15.7±4.3		15.7±4.3		17.0±3.4		17.5±5.7	
No. of Features	24.9±3.2		24.9±3.2		25.1±2.7		24.4±2.4	
Z_{limit}	1.65	1.96	2.58	3.0	5.0	10.0	15.0	17.0
Sig. Level	10%	5%	1%	0.3%	4‰	≈0%	≈0%	≈0%
Percent $CVCE$	15.7 ±4.3	15.7 ±4.3	16.2 ±5.8	19.3 ±7.0	14.4 ±4.4	21.7 ±7.3	16.7 ±3.5	32.0 ±7.5
No. of Features	24.9 ±3.2	24.9 ±3.2	25.1 ±3.3	24.2 ±1.9	22.9 ±3.9	17.1 ±1.9	11.2 ±1.5	6.0 ±0.9

First, the impact of maximum number of clusters for 1D k-means clustering (N_{clus}) and the variability limit in Level 1 (r_{limit}) are investigated. Table 5.4 shows that the effect on $CVCE$ is small compared to the standard deviation. Furthermore, the size of the selected feature set is fairly the same for all parameters. However, using $N_{clus} = 5$ yields slightly lower misclassification rates. This might be due to the number of theoretical classes, which is four for our data set. Thus, we may generalize that an appropriate choice of N_{clus} can be guided by the expected number of classes, if it is known. An good choice for the feature variability limit seems to be $r_{limit} \leq 0.1$. However, the ratio R_{obs}/R_{exp} is not lower than 0.1 for any feature. Therefore, feature rejection based on variability has no effect for this data example.

Feature rejection with the runs test is controlled by the chosen significance level. Up to $Z_{limit} = 5$, similar results for both the number of features and the classification errors are obtained (Table 5.4). There is a trend to higher classification errors when Z_{limit} is further increased or significance level is decreased, respectively. Obviously, features are missing, which have relevant information for class discrimination, when the number of features decreases below the critical size of about 15 to 10 components (see Fig. 5.5).

However, similar as for the range limit, the majority of features show non-random behavior ($Z_{test} > 1.96$). Thus, we can conclude, that the effect of the relevancy filter (Level 1) on feature reduction is negligible for this simple data set. Setting r_{limit} and Z_{limit} to zero results in similar $CVCE$ (14.8±4.7%) and number of features (25.2±2.6). Therefore, we will investigate and discuss the sensitivity of Level 1 parameters in more detail in Section 5.2 and 5.3 using real-world data. Nevertheless, since Z_{test} is also successfully used for feature ranking in Level 2 and 3, we want to emphasize that the suitability of the runs test in principal is shown for the synthetic data.

Table 5.5: Results for $CVCE$ and number of features for algorithm parameter variation in feature selection Level 2 and 3.

Mapsize		big		normal		small	
Percent $CVCE$		15.7±4.3		19.0±5.1		19.3±4.7	
No. of Features		24.9±3.2		22±3.9		20.3±1.9	
Linkage	single	centroid	average	complete	closest	neighf	ward
Percent $CVCE$	16.7±3.3	17.4±4.4	15.7±4.3	18.8±4.4	18.8±5.0	15.7±4.3	16.2±3.9
Features	24.6±2.6	25.0±2.6	24.9±3.2	25.3±2.7	25.1±2.3	24.9±3.2	25.2±2.0

NN	1	2	3	4	5
Percent $CVCE$	15.7±4.3	22.0±5.5	23.2±3.4	25.1±4.3	28.6±8.2
No. of Features	24.9±3.2	11.1±1.5	6.5±1.5	4.0±0.0	3.8±0.6

Evaluation of Feature Selection: Sensitivity of Parameters Level 2 and 3

For feature selection Level 2 and 3, several parameters control SOM learning and CP-SOM clustering (Table 5.1). In the implementation of Vesanto et al. (2000), the number of SOM units or prototype vectors, respectively, is controlled by three different options. The base number of units (Equation 4.4) is multiplied by 0.25 ("small"), 1 ("normal") or 4 ("big"). Table 5.5 shows that slightly improved classification rates, about one standard deviation, are obtained for a large SOM. It is intuitive that using a higher number of SOM units, more information about features can be quantized by the SOM prototypes. However, an upper limit of SOM units is of course necessary. Otherwise, the advantage of vector quantization (data reduction) is not given anymore.

For CP-SOM clustering two parameters are investigated. As introduced in Section 4.4, an agglomerative clustering approach on the SOM distance matrix is used, where a linkage rule controls the growth of clusters. Five linkage rules and the options "closest" and "neighf" are tested. Although the differences are within the uncertainties, Table 5.5 shows that the closest and complete linkage yield results close to the upper bound of one standard deviation of the average linkage. The second parameter is the number of nearest neighbors NN, which controls indirectly the number of clusters. Increasing NN has a strong, negative effect on the results. Involving more than one neighbor, results in too large clusters and, therefore, rejects relevant features.

Evaluation of Feature Selection: Comparison to PCA

In this section we compare our feature selection procedure with the PCA method (Section 4.6). In the context of pattern recognition, PCA can be employed in two different ways. First, we compute the $CVCE$ for the principal component projections of the data, considering the largest eigenvalues only. This is the common way to make use of PCA as an unsupervised features extraction approach. However, as already discussed, interpretation of the results is difficult since the generated components have no physical meaning. Thus, in a second step, we use PCA for unsupervised feature selection. We take those

Table 5.6: CV results for SOM learning with PCA projections. Number of components is controlled by the percent threshold T_{PCA} (see text).

T_{PCA}	50%	10%	5%	2%
Percent $CVCE$	19.7±6.5	21.7±7.2	18.3±3.7	21.0±3.3
No. of Components	2	5	8	22

Table 5.7: CV results for SOM learning and clustering with features from two largest PCA eigenvalues. Number of features is controlled by the percent threshold T_{PCA2} (see text).

T_{PCA2}	90.9%	83.3%	66.7%	55.6%	50%	33.3%	20%
Percent $CVCE$	27.2±8.3	29.0±9.5	14.9±3.8	16.1±5.1	13.8±2.8	18.0±5.6	22.5±2.2
Features	14.5±1.5	30.7±1.4	81.3±0.9	101.8±1.7	115.7±2.0	136.8±0.8	153±2.0

features having the highest weights in the projection matrix $\mathbf{\Gamma^T}$, considering only the two largest eigenvalues (Equation 4.25). These features are most significant for both principal components.

For the first approach, the number of components is limited by a percentage threshold T_{PCA}. The principal component Y_i with eigenvalue λ_i is selected given that $\lambda_i > \lambda_{limit}$, where $\lambda_{limit} = (\lambda_{max} - \lambda_{min}) \cdot (T_{PCA}/100)$. Table 5.6 shows that PCA performs slightly worse compared to our method ($CVCE = 15.7\%$) for different number of principal components. In order to obtain the features with largest weights (second approach), we use the threshold T_{PCA2}. The feature X_i with weight $\omega_i = \Gamma_{ij}$ (evaluated for $j = 1$ and $j = 2$) is selected given that $\omega_i > \omega_{limit}$, where $\omega_{limit} = (\omega_{max} - \omega_{min}) \cdot (T_{PCA2}/100)$. Using a threshold between 50 and 70%, we obtain results within the error bars of our procedure (Table 5.7). However, the number of features is clearly higher than for our feature selection method. Therefore, we are using many unnecessary, redundant features. Furthermore, taking less or more features clearly impair performance.

Evaluation of SOM Learning and Clustering

In the previous sections we have applied CV including feature selection within each fold. In this section we take the final feature set (21 features), which is obtained by applying feature selection to the complete data without CV. We now conduct CV only for SOM learning and clustering and investigate the sensitivity of the corresponding parameters (Table 5.1).

Before we do this, to get a feeling for the performance of classification by clustering, we compute the $CVCE$ for a random clustering, i.e. four cluster labels are randomly distributed. For a second test, a random permutation of the theoretical class labels is done

Table 5.8: CV results for SOM learning and clustering using different clusterings: Hierarchical (averaged linkage), hierarchical with randomly distributed data labels, 4 randomly defined clusters, and k-means.

	hierarchical	random labels	random clusters	k-means
Percent $CVCE$	15.5±3.8	52.0±3.2	74.6±5.3	24.1±9.6

Table 5.9: CV results for SOM learning and clustering using different algorithm parameters for SOM training.

Mapsize	big	normal	small	*lattice*: rect
Percent $CVCE$	15.5±3.8	19.4±4.8	17.0±3.1	18.0 ±7.4
Norm	no norm	varnorm	rangenorm	logisticnorm
Percent $CVCE$	34.9 ±4.1	21.9 ±8.2	18.3 ±6.7	15.5±3.8

before processing. The results for $CVCE$ in Table 5.8 can be considered as a baseline which all clustering approaches have to underbid. This is clearly achieved by our technique (first column of Table 5.8). Furthermore, k-means clustering is tested instead of the hierarchical approach. Using the same parameters (k_{min} and k_{max}), a clearly higher classification error is obtained. This result is expected since k-means cannot deal with non-isotropic cluster shapes, given that the Euclidean distance is used. Therefore, fit of the natural grouping is more difficult. Nevertheless, using, e.g., the Mahalanobis distance, also k-means would be able to fit non-isotropic clusters.

We first test parameters controlling the SOM training phase (Table 5.9). As for SOM learning in the feature selection algorithm, a big mapsize is favorable. Employing option "small" does not result in significantly higher errors. However, using more prototypes can become important for other data sets. For instance, consider discrimination for two very similar seismic phases (such as Pn and Pg) in case of large data sets consisting of many other types of seismic signals. Furthermore, using rectangular shaped SOM units (*lattice*) does not increase errors significantly. The choice of the neighborhood kernel (Equation 4.8) has also no effect on the results. We obtained exactly the same results using other kernels than the Gaussian one (not listed).

Normalization of data is very important for clustering or learning in general. Otherwise, the features with a broad dynamic range may dominate training, which would lead to biased results, given Euclidean distances were used. A misclassification rate of 35% is obtained without normalization for our data set. In Table 5.9 we show the results for three different approaches. We test normalization with data variance (varnorm), which is very common in Pattern Recognition, and linear normalization in the range between zero and one (rangenorm). Since there is a tendency to lower $CVCE$ and decreased uncertainties, we favor the logistic normalization (logisticnorm), which also scales the data

Table 5.10: CV results for SOM clustering using different linkage rules.

Linkage	single	centroid	average	complete	ward
Percent $CVCE$	52.5 ±5.1	21.0 ±10.9	15.5 ±3.8	24.8 ±10.6	20.3 ±5.9

between zero and one. However, here the nonlinear softmax transformation is used, which is asymptotically close to the minmum and maximum data limits and linear in between.

Additionally, we test the sequential implementation of the SOM algorithm (Equation 4.7). Compared to the batch mode (Equation 4.9) results are very similar ($CE = 17.2\%$). However, we do not carry out a CV, since the computation time for the sequential mode is very high. Thus, we employ the batch mode only for further analysis.

We employ an agglomerative hierarchical approach to cluster the SOM prototypes. For the computation of the dendrogram, we test again several linkage rules. The results in Table 5.10 show that the effect is rather high for the single linkage. The lowest misclassification rate and standard deviation is clearly obtained using average linkage. Since this rule yielded more favorable results also for CP-SOM clustering, we consider this linkage rule as the most suitable one for wavefield clustering.

Evaluation of SOM Learning and Clustering: The "k question"

The final partition is obtained by applying the k-splitting criterion to the SOM cluster dendrogram. For the CV experiments in the previous paragraphs, the number of clusters k with the lowest DB index was automatically selected between $k_{min} = 3$ and $k_{max} = 15$. We will now discuss the robustness of the DB criterion and the choice of limits k_{min} and k_{max}.

Choosing the correct number of clusters is a very critical and important step in clustering as discussed in Section 2.3.3. In their paper Vesanto & Alhoniemi (2000) pointed out that a visual inspection of DB indices for different values of k, and of the SOM itself, is mandatory for selecting the most meaningful clustering. Furthermore, it is known that there is no best relative cluster validity criterion suitable for all applications and all data sets. It is suggested to "use the index values as a guideline rather than absolute truth" (Vesanto & Alhoniemi, 2000). We will do both in the following, test performance for different values of k and visual SOM inspection.

For visual SOM inspection we consider the U-matrix representation for the complete data set. Fig. 5.6 shows that indeed a clear natural clustering exists. Counting the SOM areas with high data densities, we obtain a number of about five to 10 groups. This observation indicates that the four theoretical classes have sub-classes, or that each class cannot be represented by a single, compact cluster, respectively. However, defining a "true" k is difficult or even not possible and reasonable. Considering the U-matrix and the dendrogram plot in Fig. 5.7, there are obviously several, meaningful clustering solutions, depending on which distance scale is chosen for interpretation. Five clusters are an appropriate grouping (colored dendrogram branches). However, one could also merge the blue with the yellow and the violet with the red cluster, or even split them into further child clusters. Fig. 5.6 and 5.8 show the obtained clusterings for $k = 15$ (lowest DB index)

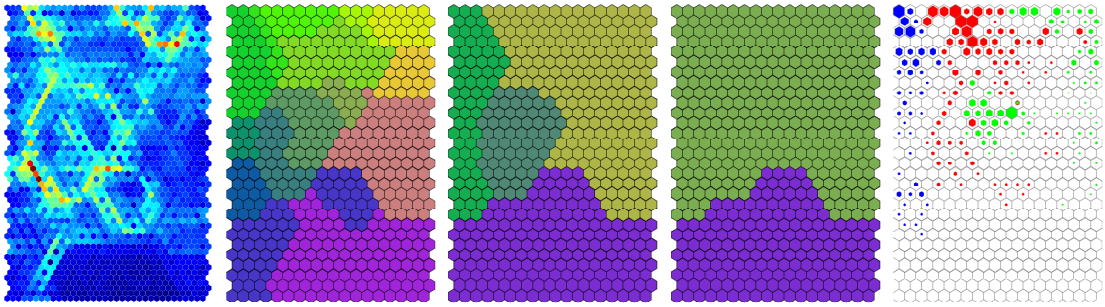

Figure 5.6: U-matrix, SOM clusterings (lowest DB: $k = 15$; furthermore $k = 4$ and $k = 2$) and hits of theoretical classes for Rayleigh waves (green), Love waves (blue) and mixture wavefield (red) for data set 1. Symbol size in the rightmost panel corresponds to signal amplitude.

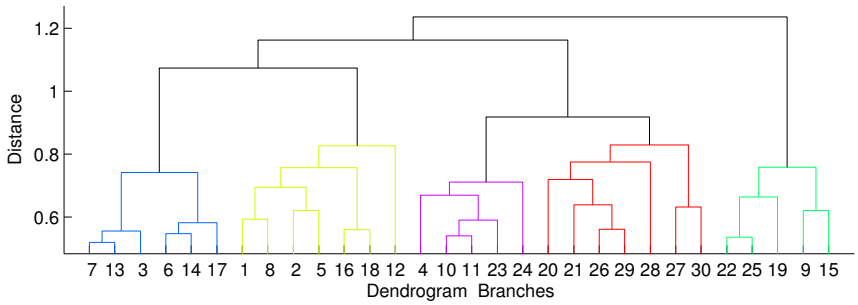

Figure 5.7: Dendrogram of SOM Clustering. Maximum number of clusters is 30. Colored branches correspond to one possible partition (5 clusters).

on the SOM, and by means of background colors for the seismogram time windows. The U-matrix structure is well fitted by the cluster borders. Furthermore, cluster memberships are in agreement with the waveform characteristics.

In order to quantify our observations, we plot two relative cluster validity criteria, DB index and cluster stability (CS), for different numbers of clusters (Fig. 5.9). The uppermost limit for k is the number of SOM prototypes (522), obtained from Equation 4.4, which corresponds to pure vector quantization without clustering. In contrast to CS, the DB criterion indicates better clusterings (lower values) with increasing k. Hence, we would obtain the best clustering for $k = 522$. This observation is an expected trade-off due to overfitting. The usual way to overcome the problem is CV. However, CV cannot be applied for relative criteria in the conventional way since labeled training data must be available. Thus, an upper limit k_{max} has to be defined for automatic clustering. However, no overfitting is observed for the CS criterion since a variant of CV is used for its determination.

When we consider both curves in more detail, we may identify weak local minima at $k = 2$ and $k = 4$. The minimum at $k = 2$, which, in contrast to DB, is the global minimum for CS, corresponds to the two-class problem "signal and noise" (see Fig. 5.6). In order to exclude this case, since SOM visualization suggests more than just two clusters, k_{min} was set to three previously. The local DB minimum at $k = 4$ fits with the expected number

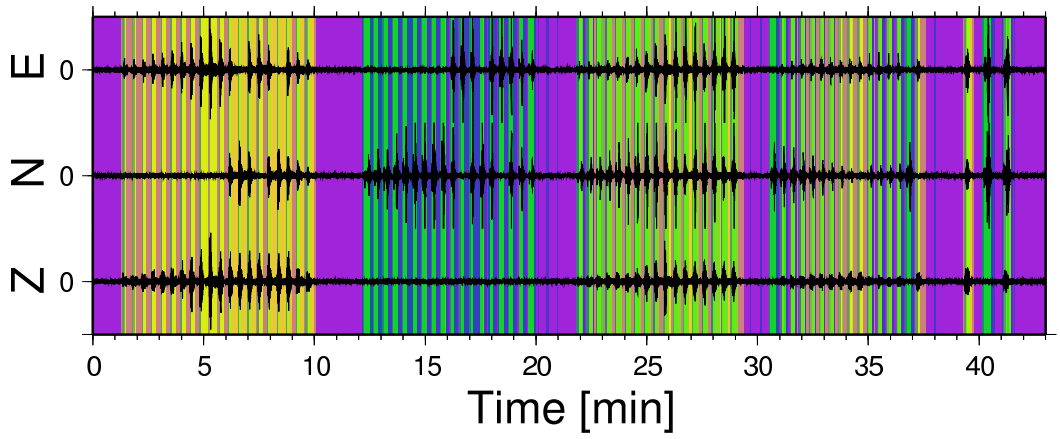

Figure 5.8: Three-component Seismogram for array center station of data set 1. Background coloring corresponds to SOM cluster colors in Fig. 5.6 ($k = 15$).

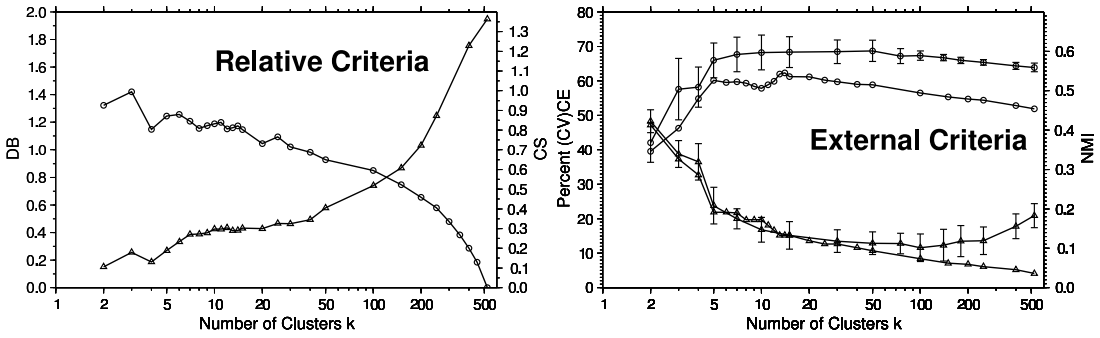

Figure 5.9: Relative cluster validity criteria (left) and classification performance (right) for different numbers of clusters. Davies-Bouldin index (DB, circles) and cluster stability (CS, triangles) do not make use of theoretical labels. Classification error (CE, triangles) and Normalized Mutual Information (NMI, circles) compare clustering with theoretical classes. The curves with error bars correspond to the cross-validation results ($CVCE$ and averaged NMI).

of classes. Furthermore, the observed number of clusters, based on U-matrix inspection, agrees with a third weak local DB minimum at $k = 8$.

Thus, we can conclude that the DB index is suitable for an automatic clustering approach. Using an upper limit for the number of clusters ($k_{max} = 15$), which can be considered as introduction of domain knowledge, leads to the visually observed number. We could also decrease k_{max} based on our expert knowledge to obtain the expected number of four classes. However, out intention is to fit the actually existing data grouping. Considering a statistic for the chosen k values from the previously conducted CV experiments ($k = 10, 11, 12, 15, 13, 14, 10, 15, 15, 14$), the automatically selected numbers of clusters are slightly higher than the U-matrix suggests without CV. However, for validation of our procedure this is acceptable. On the other hand, CS is not suitable for CV experiments, since CS estimation involves a high computational cost due to the 10-fold splitting for each CV fold. Furthermore, two clusters seem to be the preferred solution. Nevertheless,

for unsupervised learning in general, using CS additionally to the DB index helps to avoid overfitting.

Within the next paragraphs, we will make use of the theoretical class labels, which would be not available for unseen real-world data. The rightmost panel in Fig. 5.6 shows the class labels projected on the SOM scaled by the signal amplitude. This visualization allows to investigate, why the data can be fitted by more than four clusters. Signals with low amplitudes are affected by the random noise. Therefore, they form clusters localized on the SOM between strong signals and noise. Furthermore, the mixture class (red) can be either dominated by Rayleigh or Love waves. In contrast to $k = 2$, where a signal and a noise cluster is obtained, $k = 4$ does not yield clusters for the Rayleigh, Love, mixture wavefield and noise class. In fact, four clusters rather correspond to a decomposition into pure Love waves, low amplitude signals, a mixture wavefield containing the pure Rayleigh waves, and random noise.

Furthermore, two external performance criteria are plotted in Fig. 5.9 (see Section 4.5.2). We present results for classification error (CE) and normalized mutual information (NMI), with (curves with error bars) and without CV. The NMI values, averaged over all CV folds, and NMI without CV show a broad global maximum between $k = 5$ and $k = 50$. Performance decreases again for $k > 50$. Similar behavior is observed for $CVCE$. In contrast to NMI, a clear overfitting is observed without CV (CE). Taking into account the error bars, the results for both criteria show that $k > 5$ leads to a similar classification performance. Between $k = 10$ and $k = 20$, the best achievable accuracy is obtained. Using more clusters does not improve results significantly. Therefore, as for the relative criteria, it is adequate to limit k based on domain knowledge and SOM inspection, since the goal of clustering is to find the most meaningful grouping with lowest possible model complexity.

5.1.2 Data Set 2: Example of Unsupervised Analysis

The CV experiments in the previous sections showed that feature selection applied to all potential feature candidates yields a subset suitable for clustering. Furthermore, appropriate algorithm parameters for SOM learning and clustering were found. The second, more complex data set, should demonstrate the use of unsupervised wavefield analysis for a random signal sequence. We employ the complete data set and again all features for feature selection. A number of 18 automatically selected features are used for final SOM training. The corresponding SOM component planes are plotted in Fig. 5.10, together with the short name of each feature (see Table 4.2). Note that more features are rejected in Level 1 than for data set 1. In particular, 10 features from Methods 3 and 4 do not pass the runs test. Thus, the significance of Level 1 becomes more clear for data closer to real-world data. The U-matrix visualization of the trained SOM, the clustering, and the SOM similarity coloring are shown in Fig. 5.11. In Fig. 5.12 we present the labeled time windows together with the three-component seismogram of the center array station using the BMUs of each window and the color scales of Fig. 5.11b and 5.11c.

Interpretation Without Domain Knowledge

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. 5.11a) suggests that the data set can be divided

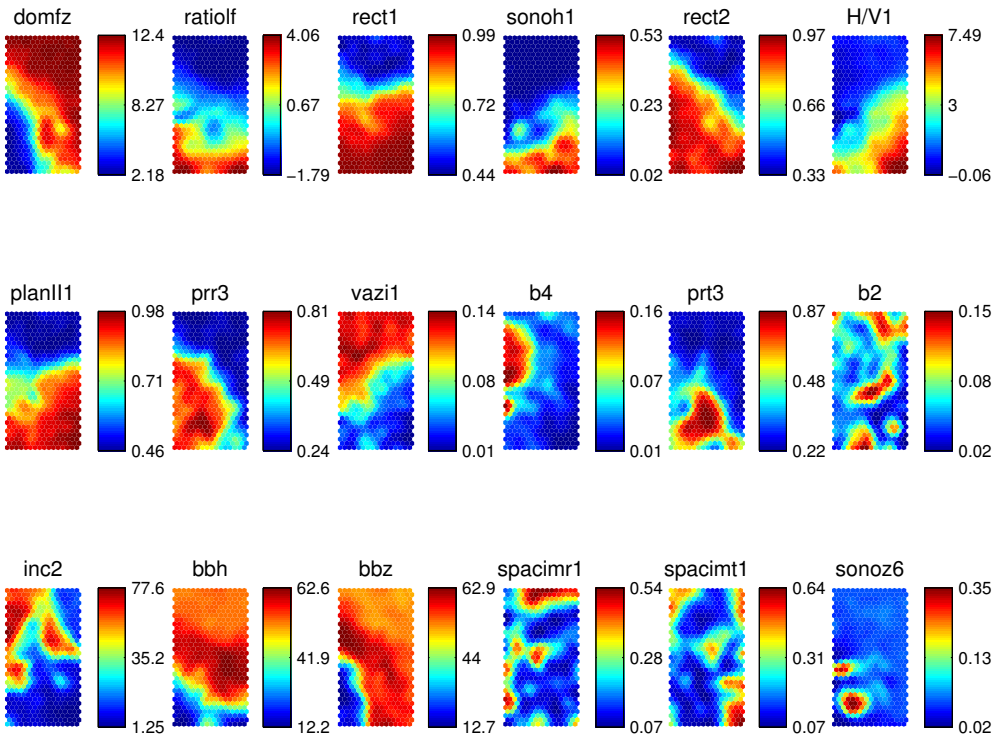

Figure 5.10: SOM component planes for synthetic data set 2.

into two major categories. The upper part of the map shows an uniformly-high data density without significant pattern. By contrast, the lower part appears more structured, i.e. several areas with high data densities are limited by regions of lower densities. However, definition of clusters is difficult since not all high-density areas are completely surrounded by distinct and sharp boundary regions.

The automatic clustering confirms our observations (compare Fig. 5.11a and 5.11b). The upper part of the SOM is assigned to a single cluster (green), whereas the remaining clusters more or less fit the structure of the U-matrix. Obviously, the borders are not chosen optimally due to the mentioned complex structure and lack of distinct cluster borders. However, the clustering provides a first order grouping and a meaningful representation of the dominant patterns of the data (Fig. 5.12). In order to resolve the fine structure and transitions between clusters, the SOM coloring technique introduced in Section 4.4 can be used. The coloring of SOM units in Fig. 5.11c allows to identify four dominant pattern, represented by green, blue, red and yellow colors. The corresponding background coloring of the seismograms in Fig. 5.12 intuitively highlights these pattern in their temporal context.

From the DB index using $k_{min} = 3$ and $k_{max} = 15$, we automatically obtain $k = 6$ as best clustering. Fig. 5.13 shows DB and CS criteria also for higher k values. In fact, there is a clear local minimum at $k = 6$ for the DB index. In contrast to CS, DB slightly decreases further for $k > 15$ due to overfitting. As for data set 1, there is also a clear minimum at $k = 2$ for DB as well as for CS, which is global for the latter. Nevertheless,

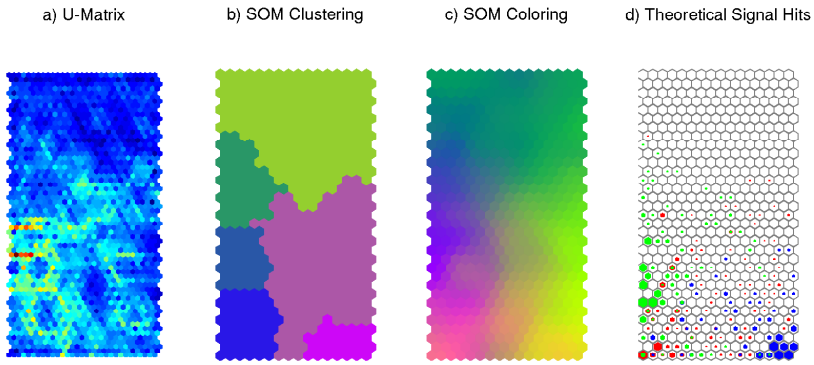

Figure 5.11: U-matrix, SOM clusterings (lowest DB), SOM coloring and hits of theoretical classes for Rayleigh waves (green), Love waves (blue) and mixture wavefield (red) for data set 2. Symbol size in the rightmost panel corresponds to signal amplitude.

visual SOM inspection suggests that $k = 6$ is an appropriate grouping. Hence, DB is more suitable for this data set.

Interpretation Using Domain Knowledge

Let us now introduce some more expert knowledge into the interpretation, i.e. that we only generated surface waves and added random noise. Considering the background coloring of time windows without transients (Fig. 5.12), noise is obviously represented by the top part of the SOM. The uniformly-high density shows that the noise is a homogeneous class, whereas the distribution of signal time windows has a more sparse and inhomogeneous character. Focusing on the dependencies between all three spatial components for a single time window, we can derive that the yellow colors (SOM-coloring) or the bottom right cluster (light violet), respectively, represent dominantly Love waves, since there is no signal amplitude on the vertical component. The bottom left area of the SOM, which corresponds to violet and magenta (SOM similarity coloring) or bluish time windows (cluster coloring), represent signals with vertical amplitudes. This is an evidence for Rayleigh wave contribution. The observation based on signal amplitudes is confirmed by the CPs (Fig. 5.10). Signal and noise are mainly discriminated by polarization (*rect*, *planII*, *vazi*), but also by spectral properties (horizontal bandwidth *bbh*). On the other hand, Rayleigh and Love waves are distinguished by their energy contribution on vertical and horizontal components (e.g. radial coherency *prr*, dominant vertical frequency *domfz* and vertical bandwidth *bbz*). Furthermore, values for angle of incidence (*inc2*) and semi minor axis of polarization ellipsoid (*b4*), which are theoretically zero for Love waves and vary for Rayleigh waves and noise, are consistent with our interpretation. 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 of course affect 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 (bottom of SOM) and dominant noise (top of SOM).

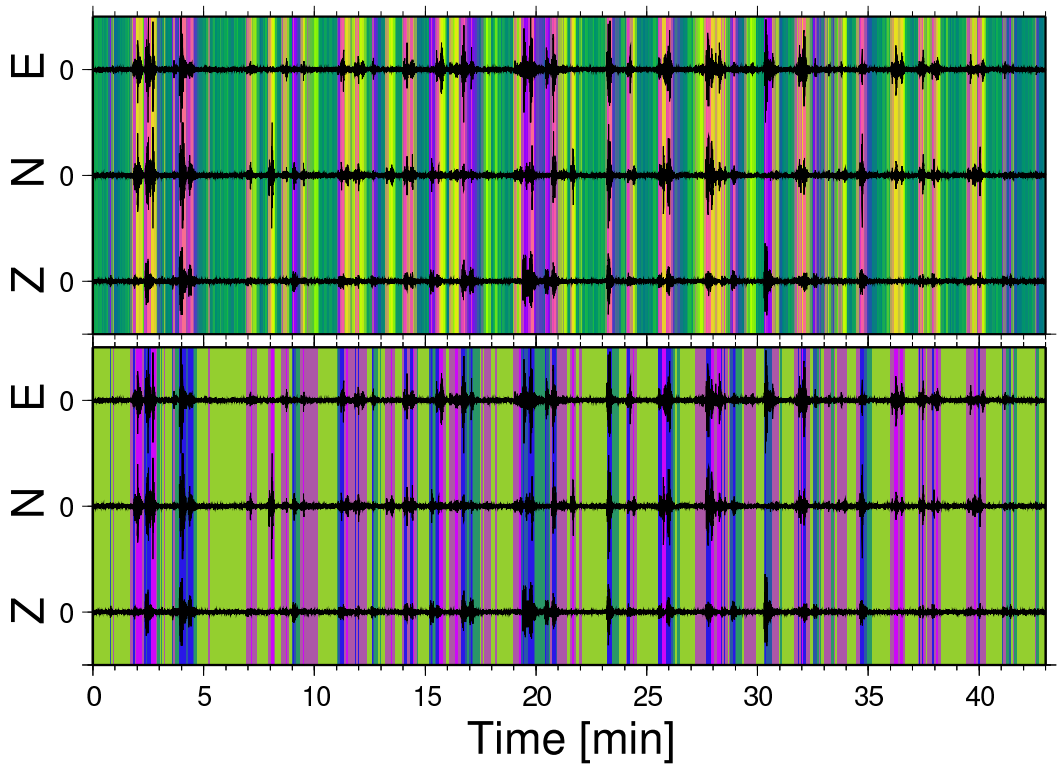

Figure 5.12: Three-component seismogram for array center station of data set 2. Background coloring corresponds to SOM cluster colors in Fig. 5.11b (lower panel) and SOM coloring in Fig. 5.11c (upper panel).

Evaluation Using True Class Labels

In a last step, we now consider additionally the employed source setting for interpretation. The projection on the SOM, i.e. finding the BMUs of the labeled feature vectors, is shown in Fig. 5.11d. The size of each 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 hits are located on the bottom left and pure Love dominantly on the bottom right areas of the SOM. The mixture wavefield is distributed between both sides depending on the force orientation. However, strong amplitudes only capture the leftmost (Rayleigh wave-like) area. Furthermore, it is confirmed that the more we approach the top of the SOM, the smaller are the signal amplitudes. There are no hits of signals on the uppermost part.

Finally, we carry out six CV experiments for SOM learning and clustering in the same way as for data set 1. For each CV the time window length for feature generation is varied to investigate the sensitivity of the parameter $WINFAC$. Taking into account the uncertainties, Table 5.11 shows that the differences are not significant for $WINFAC \leq 7.5$. However, for increasing values there is a trend to higher errors and uncertainties, since, as expected, discrimination of wave types is getting more difficult for longer time windows.

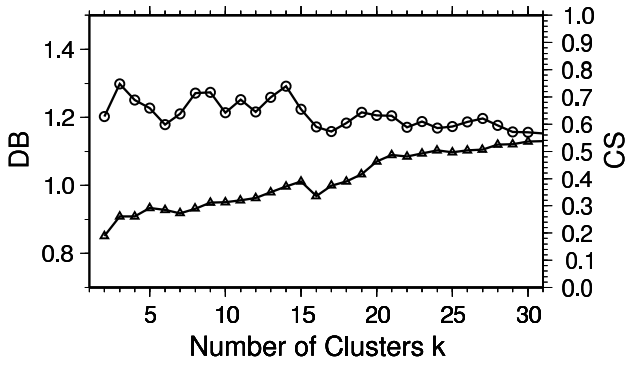

Figure 5.13: DB (circles) and CS (triangles) cluster validity criterion for different numbers of clusters.

Table 5.11: Results of CV for synthetic data set 2. Feature generation parameter $WINFAC$ is varied.

$WINFAC$	1	2.5	5	7.5	10	20
Percent $CVCE$	27.6±2.0	25.7±4.0	28.1±4.5	30.7±6.6	34.4±7.6	35.3±15.1
No. of Features	21	25	18	17	13	16

5.1.3 Discussion

Let us now summarize the results and what we have learned from SOM-based analysis of synthetic wavefield data. We aim at giving practical guidelines and suitable algorithm parameter for the real-world data analysis.

Realiability of Feature Selection

The histograms in Fig. 5.14 show how often a feature is selected during CV for three experiments on data set 1. In particular, we present applications of our feature selection procedure (Fig. 5.14a), the PCA selection approach (Fig. 5.14b), and several feature selection parameter tests (Fig. 5.14c). For the latter, we merge the CV feature statistics for different CP-SOM clustering parameters (linkage: five rules, nearest neighbors: five options). In order to obtain a feature set of comparable size, a threshold of $T_{PCA2} = 83.3\%$ is used for PCA feature selection (about 30 features). All three plots show that similar feature sets are selected in each CV fold. Thus, the procedure is stable and robust with respect to choice of data and parameters. There is less variance for PCA compared to our technique. However, overall feature distribution is similar. This observation shows that our procedure implicitly selects relevant features based on constraints similar to those of the approved and well established theory of PCA. Therefore, the reliability of our technique in this context is given. Nevertheless, remember that robustness has always to be interpreted together with classification performance (Saeys et al., 2008). In spite of being more stable, clearly higher misclassification rates are obatined for PCA using the chosen number of features ($CVCE = 29.0\pm9.5\%$, see Table 5.7). The white symbols in Fig. 5.14a correspond to the feature sets of data set 1 and 2 obtained without CV (see Table 5.12). As expected, the selected features are consistent with the histogram, what is

Table 5.12: Short names, generation methods and runs test statistic for features selected from data set 1 and 2.

Data set 1			**Data set 2**		
Feature	Method	Z_{test}	Feature	Method	Z_{test}
prz1	1	21.43	*domfz*	5	20.52
domfz	5	20.22	*ratiolf*	5	19.45
a_b2	6	19.76	*rect1*	3	19.45
ellez2	4	17.17	*sonoh1*	5	18.69
ratiolf	5	16.56	*rect2*	3	18.08
prr3	1	16.56	*H/V1*	7	17.32
sonoh10	5	16.26	*planII1*	3	17.02
elip2	3	15.34	*prr3*	1	16.71
rect1	3	14.89	*vazi1*	4	14.81
rect2	3	14.89	*b4*	6	14.12
b4	6	14.81	*prt3*	1	12.75
FQ1zh2	4	14.12	*b2*	6	11.08
sonoz7	5	13.06	*inc2*	4	10.55
sonoz6	5	11.99	*bbh*	5	10.32
prt3	1	11.84	*bbz*	5	8.72
dopIII2	3	11.8395	*spacimr1*	2	7.20
H/V2	7	11.69	*spacimt1*	2	6.89
spacimr2	2	11.54	*sonoz6*	5	5.85
sonoz5	5	6.97			
3cell3	4	3.69			
ellzn3	4	2.70			

a further evidence for stability.

The feature statistics in Fig. 5.14a and the final sets (Table 5.12) show that features are selected from almost all feature generation methods. Hence, features from each method show significant temporal patterns and provide information not given by the others. Nevertheless, there are methods whose features occur more frequently. Most features are directly computed from the frequency spectrum (Method 5). Furthermore, since features in Table 5.12 are ranked with respect to Z_{test}, obviously several features from Method 5 are among the most significant ones. The reliability of time-frequency features confirms common approaches on signal detection in seismology such as the one of Joswig (1990) and Riggelsen et al. (2007). However, in our final feature sets also polarization (linearity, degree of polarization, ellipticity, Methods 3 and 4) and coherency attributes (Methods 1 and 2) have a significant contribution.

Summing up what we have learned from the experiments in the previous sections, it is sufficient to consider only the finally selected feature subset (Stage 4). Employing the three-level approach with the parameters found to be appropriate, we obtain a meaningful feature ranking based on the Wald-Wolfowitz test statistic and an adequate feature

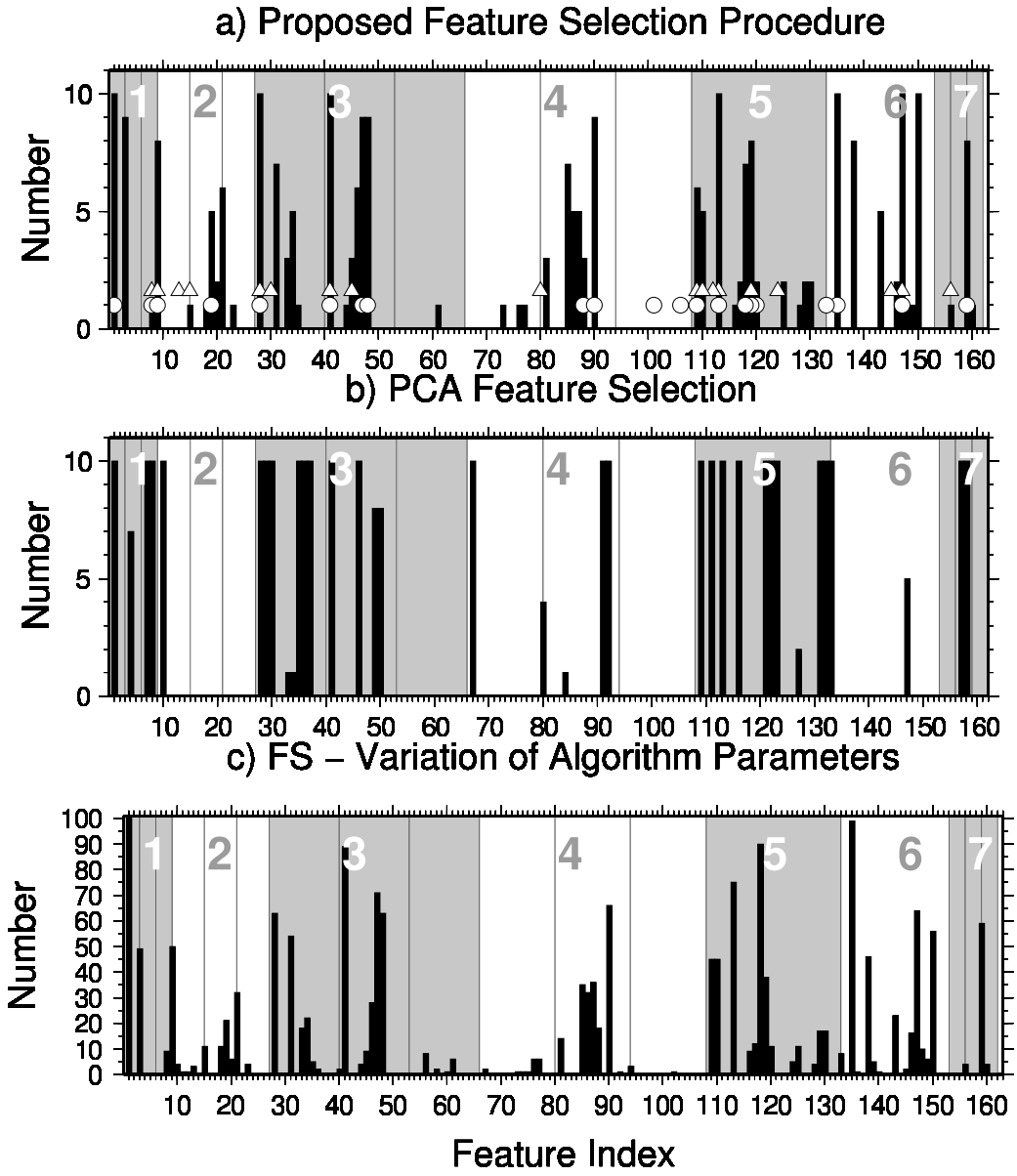

Figure 5.14: CV selection statistic for each feature using data set 1. Background coloring distinguishes different feature generation methods (see gray and white numbers) and vertical lines frequency bands (only for methods with three bands). a) Our feature selection procedure b) PCA feature selection c) Merged results for CV experiments on feature selection algorithm parameters. White symbols in a) correspond to the feature set without CV. Circles stand for data set 1 and triangles for data set 2.

grouping. Both properties nearly lead to the smallest possible feature set (about 10 to 15 features, see Fig. 5.5). Reducing this set further would increase misclassification rates. Thus, the dimensionality, and therefore also computation time and model complexity, is reduced considerably for further analysis and interpretation of the data set, without significantly losing discriminative power. However, we are dealing with relatively simple synthetic data sets, where almost all features carry parts of the relevant information, even though redundancy is very high. Thus, the significance of Level 1 for feature reduction is low. Furthermore, the classification performance is not significantly improved compared to the complete feature set. On the other hand, the improvement is significant compared to a random feature set of same size as the finally selected one. Furthermore, by combining features from different feature generation methods, performance is clearly improved compared to just using the common and most suitable single approach, which we find is the time-frequency spectrum (Method 5). Our feature selection procedure is more favorable than feature selection or feature extraction based on PCA, in the sense that the best performing feature set (lowest errors) is obtained for less features than using PCA.

Since the effect and significance of feature selection Level 1 is low for both synthetic data sets, more investigations on corresponding parameters have to be carried out in the next sections. In particular, the limit for the runs test statistic has to be examined more intensely.

Realiability of SOMs and SOM Clustering

The clustering of synthetic seismic wavefield data by means of hierarchical grouping, employing the average distance as linkage rule, performs clearly better than using other linkage rules, k-means and, not surprisingly, than a random labeling of the data. Therefore, we will use this technique, and also the appropriate algorithm parameters for SOM learning (e.g. big map size, logistic normalization), in the following sections. Finding average distance as the most favorable linkage rule is not unexpected, since we already mentioned in Section 2.3.1 that it can deal with compact as well as with anisotropic clusters. Other SOM learning parameters, such as choice of neighborhood kernel, have no significant effects on results. Both synthetic data sets confirm that visual inspection of the SOM (e.g. U-matrix) and cluster quality criteria are mandatory to find an adequate number of clusters fitting the natural grouping of the data. If automatic clustering is desired (e.g for CV), a k_{max} limit based on visual inspection and the expected number of clusters must be given. In this case, the DB index is appropriate to find a single meaningful data partition. The CS criterion is more suitable to avoid overfitting, but seems to favor low k instead ($k = 2$ for our test). Furthermore, it has a high computational cost. When the U-matrix shows no clustering, the SOM coloring technique based on PCA is well suitable to highlight pattern in the seismogram. Furthermore, we have to keep in mind that natural clusters in the data may not always correspond to the expected number of wave-type classes. This observation confirms the need of unsupervised learning for data inspection, e.g. before more advanced supervised learning can be carried out. In the following, we will apply the same processing techniques to real-world data, since we expect similar data characteristics.

5.2 Regional Earthquake Recordings

The following application example shows the SOM-based analysis of regional seismicity. In this section we apply our feature selection procedure to earthquake recordings in order to find suitable features, which allow to detect the onset of events and also to distinguish between different phases of arriving waves. Furthermore, the phase and event discrimination in the feature space will be investigated. We consider earthquakes recorded at the European broadband network with magnitudes larger than four, which occurred 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, and not for short time windows as done in our study. Nevertheless, a similar example is also given in this section. Wavefield analysis using the complete data is addressed in Section 5.3 and 5.4 using other real-world data sets.

We employ recordings lasting six minutes and starting two minutes before P wave onset. First, we select 44 earthquakes, for which we could identity and pick clear P and S onsets at the station RDO (See Appendix, Table 8.4). The picks will help us to evaluate our observations after SOM learning. For feature selection and SOM training, we prepare a single data set by computing the features for 6.56 seconds long, non-overlapping time slices ($WINFAC = 4$, $f_{cent} = 0.61$ Hz) for each event (55 slices each). Frequency bands are located between 0.09 and 3.5 Hz (see Appendix, Table 8.3). Subsequently, the individual vector time series are merged. Finally, we obtain a set of 2420 slices for 44 events. Since we do not use a receiver network, only Methods 3 to 7 can be employed for feature generation (129 features). Furthermore, in order to investigate the station-dependency of the observed patterns, a second data set is created in the same way for an additional receiver (KEK). At station KEK, 57 events with clear P and S onsets could be identified (See Appendix Table 8.5).

In the following, including processing of other real-world data sets in Section 5.3 and 5.4, SOM training is performed using a weight for each selected feature (see Equation 4.6). The weight of a feature corresponds to its runs test statistic Z_{test}. Therefore, also the ranking of features will affect the SOM training. Furthermore, we start our analysis of the earthquake data set without SOM clustering. The resulting SOM prototype vectors are directly used to investigate seismic phase discrimination.

5.2.1 Data Set Including Three Events

We start with a simple example using three recordings to demonstrate the procedure. The events were recorded at the same receiver RDO and occurred at different times (23/12/2003, 02/12/20041, 08/02/2006) in the same source region (39.3 - 40.7° N, 28.0 - 30.4° E). In contrast to the complete data set, the analyzed time segments for each event are lasting three minutes longer (9 min, 79 time slices). In Fig. 5.15 the three-component seismogram is shown for one event. The labels and the gray scale on top indicate different wave phases, which can be identified using theoretical arrival times (Pn, Pg, Sn, Sg)

Table 5.13: Features automatically selected for three-event data set. Short name and runs test statistic Z_{test} are given.

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
sonoh9	11.78	*sonoh6*	8.40	*planII3*	6.05
sonoh8	11.78	*a_b1*	6.84	*sonoz2*	5.01
FQ1zh3	10.48	*b1*	6.71	*b3*	4.36

and domain knowledge (Coda). Therefore, time slices can be members of six manually defined classes (Pn, Pg, Sn, Sg, coda, and noise). All 129 potential features are computed for the analysis. The feature selection procedure is applied using the suitable algorithm parameters found in Section 5.1.1. An exception is Z_{limit}, whose importance could not be investigated using the synthetics. Here, for now, we use $Z_{limit} = 4.0$. However, in Section 5.2.2 we will test and discuss the choice of Z_{limit} in more detail. In the following, classification errors (CE^+, false positive and CE^-, false negative) are computed using Equation 4.22. In particular, we consider the class memberships of SOM prototypes and average over all six class errors (Pn, Pg, Sn, Sg, coda, and noise). Note that no CV is carried out. Therefore, the following results cannot be generalized beyond these three event.

Feature Selection and SOM Training

Our automatic selection procedure finds a set of nine features (Table 5.13). Again, we observe that spectral features dominate regarding their number and their position in the ranking with respect to Z_{test}. However, also polarization is represented by planarity and features from Method 6. The first column of Table 5.14 shows that the best discrimination of wave phases, compared to other feature subsets, is obtained after feature selection with our selected feature vector. For comparison, using nine randomly selected features, results in clearly higher classification errors. Considering the feature generation methods separately, classification errors for Method 5 are similar compared with feature selection using all methods. Thus, features representing the time-frequency content of the wavefield are the most suitable. This is again a confirmation of classification approaches in seismology, where spectral features are often employed a priori (e.g. Joswig, 1990).

Additional Parameter Tests

We perform additional sensitivity tests on feature generation, feature selection, and SOM training parameters, which synthetic data could not account for.

Since we are dealing mainly with body waves, and synthetic signals were mainly consisting of surface waves, the effect of the time window length on classification rates may differ ($WINFAC$). From Table 5.15 we find that $4 \leq WINFAC \leq 6$ is favorable for seismic phase classification, which is comparable with the synthetics (Table 5.11). However, in contrast, not only increasing, but also decreasing $WINFAC$ increases the errors.

Table 5.14: Classification errors for three-event data using all features and subsets (feature generation methods), with and without applying feature selection. No feature from Method 7 passed Level 1. Instead results for a random feature set are shown in the lower panel. "F No." is the number of features. Best results are highlighted.

Full Features Sets						
	All	Meth. 3	Meth. 4	Meth. 5	Meth. 6	Meth. 7
CE^+	16%	12%	15%	23%	16%	16%
CE^-	19%	15%	19%	25%	11%	21%
F No.	129	39	36	25	20	9
Features Selection						
	All	Meth. 3	Meth. 4	**Meth. 5**	Meth. 6	Random
CE^+	**6%**	22%	13%	**6%**	15%	18%
CE^-	**10%**	29%	10%	**9%**	15%	24%
F No.	**9**	2	2	**8**	8	9

Table 5.15: Sensitivity of time window length given by parameter $WINFAC$. $WINFAC = 4$ was used a priori.

$WINFAC$	1	2	**4**	6	8
CE^+	24%	20%	**6%**	7%	36%
CE^-	23%	30%	**10%**	8%	36%

An explanation is that too short time windows hinder to estimate stable features due to noisy data and lack of frequency resolution. On the other hand, too long time windows make wave phase discrimination difficult.

As introduced in Section 4.1, we use the center frequency of the overall frequency band (center of all employed frequency bands) to compute the same time window length for each feature. However, within the context of Fourier transform computation for short time windows, in order to obtain a stable estimation, it has been suggested to define window length depending on the considered frequency band. Thus, time windows must overlap (low frequencies), or gaps will exist (high frequencies), respectively, to obtain the same number or instances for features computed in different frequency bands. Carrying out such a setting yields $CE^+ = 9\%$ and $CE^- = 9\%$. Considering both the false positive and the false negative errors, the results are quite similar compared with the non-overlapping time windows of constant length ($CE^+ = 6\%$ and $CE^- = 10\%$). Therefore, the way of defining the window length is apparently not important for our approach.

Finally, we investigate the effect of the feature weighting in the SOM training phase. We can state that weighting has a positive effect on performance. Without weights we obtain slightly higher misclassification rates ($CE^+ = 7\%$ and $CE^- = 12.4\%$), which leads to the conclusion that the ranking with respect to Z_{test} is meaningful.

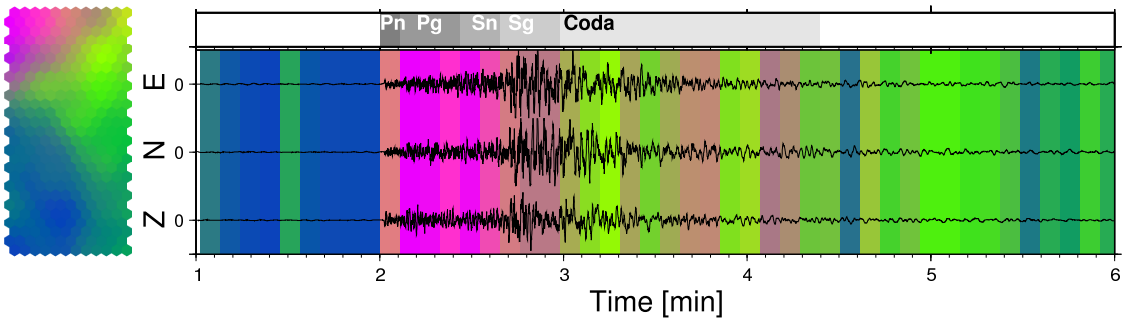

Figure 5.15: Three-component seismogram for an earthquake record. On top different wave phases are indicated. On left hand side the corresponding SOM is shown. Background coloring of seismograms corresponds to SOM coloring which is based on prototype vector similarity.

SOM Visualization

Fig. 5.15 shows the similarity coloring of the SOM, which was trained using an automatically selected feature set. Furthermore, the same color scale is used for the seismogram background colors of one event. This visualization gives an impression about the natural discrimination between seismic phases. The Pn and Pg onsets of the earthquake, as well as the Sg phase, are clearly highlighted. However, obviously the differences between Pn and Sn are not so prominent. Furthermore, we observe that the earthquake coda is lasting longer than visually derived from the signal amplitude. Clear differences between background noise and coda are highlighted in spite of lack of transients.

5.2.2 Data Set Including 44 Events

In this section we investigate the generalization capability of our unsupervised learning procedure using the data set of 44 events from station RDO (Seismograms are shown in the Appendix, Fig. 8.6). For this purpose, again cross-validation experiments are carried out using 44 folds, i.e. by leaving out time slices of one event each fold.

Cross-Validation on Feature Selection

First, CV is performed including feature selection ($Z_{limit} = 1.96$) and final SOM learning. Unlike CV for the synthetics, feature selection Level 1 is included in CV, since no random permutation of the data is done. We use only three ground-truth classes (P wave, S wave and noise), since we are not able to identify all weak phases for all events. The S wave class includes all time windows after the S wave onset until the end of the event. Thus, this class also includes potentially existing surface waves and the event coda. Furthermore, although the transition between coda and background noise is continuous, we introduce a third time pick, which should define the "end" of the event. This definition of classes is of course a strong simplification. However, since for example Love waves are gradually developing from SH waves interactions, similar characteristics compared to S waves are expected.

Fig. 5.16 summarizes the automatically selected features for both the cross-validation experiment and using directly the complete data set. The latter ones are also listed in Table

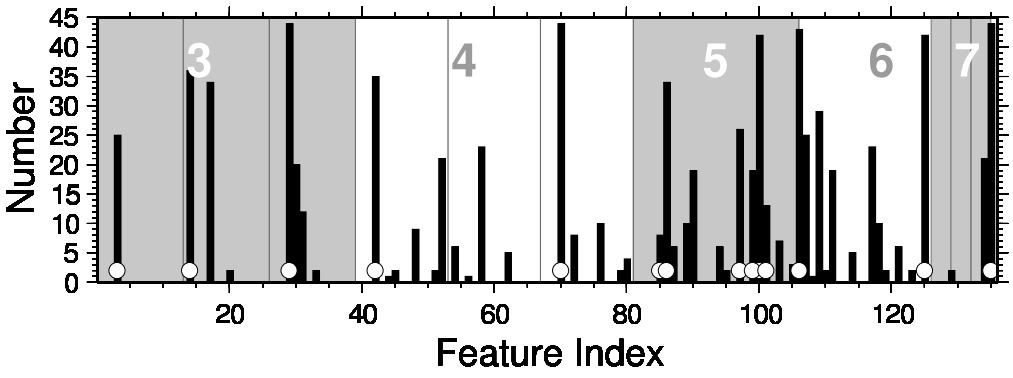

Figure 5.16: Features selected in the 44-fold CV experiment for earthquake data set (RDO). Background coloring distinguishes different feature generation methods (see gray and white numbers). Vertical lines denote frequency bands. White circles correspond to feature set obtained using the complete data (no CV).

Table 5.16: Features automatically selected for earthquake data set (RDO). Short name (see Table 4.2) and runs test statistic Z_{test} are given.

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

5.16 in detail. The histogram in Fig. 5.16 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. Again, 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.

Table 5.17 shows that the cross-validated classification errors, obtained with and without feature selection, are similar within their uncertainties. Hence, the information content is maintained. Furthermore, number of features, and therefore model complexity, is reduced significantly. Again, features of different generation approaches are combined. For comparison, using only features from the most common seismological approach (Method 5), yields classification errors close to the upper bound of uncertainties regarding the complete feature selection (FS). Furthermore, also random feature sets, with dimensions similar to the automatically obtained set, result in higher errors. Thus, our procedure finds a meaningful and suitable combination of features. In order to get a feeling for the

Table 5.17: CV results for earthquake data set (RDO) using features selection (FS), all features (noFS), features from generation Method 5 and random feature sets.

	FS	no FS	Method 5	Random FS
$CVCE^+$	33.8±10.9%	34.6±13.8%	41.7±12.1%	43.5±10.5%
$CVCE^-$	34.8±12.1%	36.1±11.0%	41.3±9.8%	45.3±9.5%
No. of Features	18.6±2.6	136	26	20

goodness of the $CVCE$ estimates, consider again the result of a randomly labeled SOM, which yields $CVCE^+ = 64.9 \pm 12.6\%$. In fact, this value is expected since the number of instances within each class is similar. Thus, there is a chance of 66.6% of drawing a vector not belonging to a particular class.

Comparing our feature selection procedure and application of PCA for dimensionality reduction, similar results are obtained. For 14 principal components ($T_{PCA} = 10\%$) we obtain $CVCE^+ = 34.4 \pm 13.4\%$. Including more components does not improve performance. Hence, we have a further confirmation for the reliability of our procedure. Moreover, the advantage compared to PCA is the improved interpretation ability.

SOM Visualization

Fig. 5.17 presents the SOM U-matrix visualization and data hits on the SOM. The U-matrix shows a clustered, more sparse region at the bottom. For the remaining SOM no clear clusters can be observed. The hits of all P wave time windows (white), and only the first P onset time windows of each event (black), are located within the clustered area (Fig. 5.17b). Most S wave onset windows (black) are also well separated from P wave and noise (Fig. 5.17c). However, the spread of all signal windows after S onset is higher (white). There is a prominent cluster of hits close to the left corner of the SOM, which may indicate surface waves. Furthermore, there is an expected continuous transitions from the S wave class to noise and no distinct cluster boundaries (Fig. 5.17d). In fact, noise hits after the event (white) are mainly located on the righthand area of the SOM. Thus, they are still affected by the event.

Cross-Validation on SOM Learning

In order to quantify seismic phase discrimination on the SOM, a further CV experiment is conducted for two particular feature sets. We employ the automatically selected feature set, which is composed of attributes from different generation methods (Table 5.16), and the spectral features alone (Method 5). Validation is again carried out based on SOM prototype vectors (no hierarchical clustering). We compute median false positive and negative classification errors for the Noise, P wave and S wave class. Using the median and the mean deviation from median (MD), instead of mean and standard deviation, is more appropriate here, since results for individual classes j and folds (CE_j) show non-normal distributions of errors (see Appendix, Fig. 8.5). On the other hand, class averaged

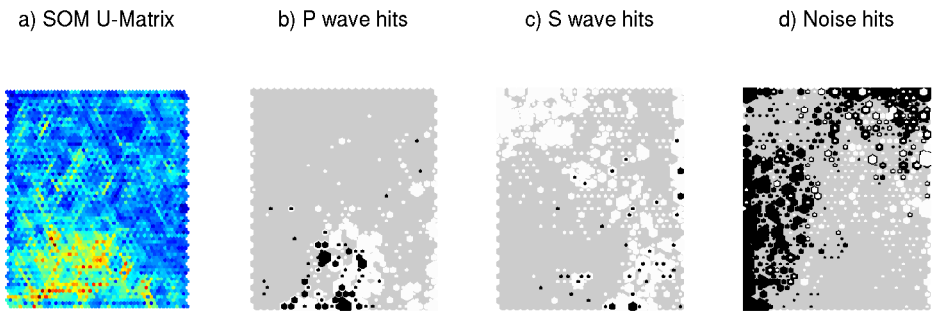

Figure 5.17: SOM visualizations for earthquake data set. a) U-matrix. b) Hits of P wave time windows on SOM for all events (white). Black symbols show first onset window for each event. c) Same for S wave windows. d) Hits of noise class (black). White hits represent noise time windows after event.

errors (Equation 4.22) are approximately normally distributed. Thus, for individual class errors, median CV results are presented ($MedCE_j$), while for the class-averaged errors both the median ($MedCE$) and the mean ($CVCE$) are shown for comparison.

Although the uncertainties of the results are rather high, we can derive some qualitative insights from Table 5.18. The highest misclassification rate is obtained for the S wave time windows, which confirms the observations from Fig. 5.17 (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. As for the data set including three events, the manual labeling may be incorrect 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. However, still the uncertainties are rather high. 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 signal detection. However, for S wave recognition we obviously need additional polarization information to improve phase discrimination.

Using 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. Using 50 seconds 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 (Table 5.19). In spite of a simple vector quantization and nearest neighbor classification without 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 in Table 5.19. Nevertheless, since classification

Table 5.18: CV results for earthquake data set. Median values of false positive and negative classification errors are given for each class and for two feature subsets ($MedCE$). A set obtained by feature selection (FS) and features from generation Method 5 (spectral features) are used. The class "signal" contains both the P and the S wave class. Class averaged errors are also presented as median and, additionally, as mean ($CVCE$) of all CV fold errors.

	FS		**Method 5**	
Class	Percent $MedCE^+$	Percent $MedCE^-$	Percent $MedCE^+$	Percent $MedCE^-$
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 (P and S)	25.4 ± 16.8	19.2 ± 16.0	23.6 ± 16.2	33.3 ± 18.3
Class Averaged	33.4 ± 10.3	35.8 ± 9.9	39.3 ± 9.27	41.7 ± 7.5
	Percent $CVCE^+$	Percent $CVCE^-$	Percent $CVCE^+$	Percent $CVCE^-$
Class Averaged	36.1 ± 13.5	35.6 ± 12.9	41.7 ± 12.1	41.3 ± 9.8

Table 5.19: Same as Table 5.18 but using only 25 seconds of data before and 25 seconds after P wave onset. Medians ($MedCE$) and mean deviations from median MD are given.

	FS				**Method 5**			
Class	$MedCE^+$	MD	$MedCE^-$	MD	$MedCE^+$	MD	$MedCE^-$	MD
P	0.0%	10.4%	0.0%	11.0%	0.0%	13.7%	0.0%	12.3%

errors and spread (MD) are clearly higher for the three-class and complete-event problem (see Table 5.18), our results do not imply that no context-dependency is required for discrimination of more classes (P and S waves) and longer records.

Sensitivity Test for Z_{limit}

In a next step, we explore the sensitivity of Z_{limit} to test the significance of feature selection Level 1 on SOM learning for real data. For this purpose, we first apply feature selection using different values for Z_{limit}. Afterwards, a CV experiment is conducted for each feature set in the same way as in the previous section (only SOM training). The results are again presented as false positive and negative mean class-averaged classification errors ($CVCE$, compare with lowermost row in Table 5.18).

Table 5.20 shows the results for Z_{limit} values between zero and 35. For $1.96 \leq Z_{limit} \leq 15$ classification errors are very similar. Furthermore, also the corresponding feature sets for $1.96 \leq Z_{limit} \leq 10$ have a comparable size. The lowest errors are obtained

Table 5.20: CV results using different feature subsets which correspond to different values for Z_{limit}.

Z_{limit}	0.0	1.96	3.0	5.0	10.0
Percent $CVCE^+$	40.9±14.6	36.1±13.5	37.4±12.9	37.0±11.7	32.4±10.4
Percent $CVCE^-$	40.4±11.9	35.6±12.9	38.0±11.3	38.4±11.5	34.5±11.1
No. of Features	24	14	18	11	13
Z_{limit}	15.0	20.0	25.0	30.0	35.0
Percent $CVCE^+$	39.4±11.7	43.4±14.9	43.1±13.3	41.8±13.3	45.7±12.5
Percent $CVCE^-$	42.2±11.4	43.7±13.3	44.2±11.5	43.0±13.2	45.7±11.6
No. of Features	9	9	4	7	2

for $Z_{limit} = 10$. However, compared to the uncertainties, this difference is not very significant. Therefore, using $Z_{limit} = 1.96$, to obtain non-random feature time series, is again a good choice for a suitable feature set. For $Z_{limit} > 15$, classification errors have a slight tendency to increase towards the upper bound of the uncertainty of the lowest error. An explanation is that features are excluded which represent patterns of shorter periods. In fact, we expect a minimum pattern length of one time window for our problem (see phases in Fig. 5.15). Taking into account the results of the runs test experiments in Section 4.2 and the length of our data set ($N = 2420$), this corresponds to $Z_{test} < 5$. Thus, in order to ensure that we recognize all existing patterns, and under considerations of results in Table 5.20, using $Z_{limit} = 1.96$ is appropriate for the earthquake data set.

Furthermore, we observe that rejecting random features in feature selection Level 1 ($Z_{test} < 1.96$), in fact has a significant effect on the number of features. For feature generation Method 3 (from 39 to 28 features) and 4 (from 42 to 25 features), feature sets are reduced clearly. Without Level 1 ($Z_{limit} = 0.0$) slightly higher errors are obtained (Table 5.20).

SOM Clustering

Note that we conducted no automatic hierarchical SOM clustering for the previous experiments. For classification and exploration of signal discrimination, we directly used the SOM prototypes. Fig. 5.18 summarizes how SOM partitioning effects the classification performance for different number of clusters. First, the relative cluster validity criteria DB and CS are computed between $k = 2$ and $k = 30$ (stars and crosses). While CS shows a local minmum at $k = 4$, and then increases monotonically, DB exhibits a broader minimum (lowest value at $k = 6$), before it finally decreases again due to overfitting for $k > 10$. Since the best clustering in Fig. 5.19b ($k = 6$) is a meaningful fit of the U-matrix structure in Fig. 5.19a, and the best solution regarding CS is $k = 2$, we have again an evidence to favor the DB index as validity criterion, given that we avoid overfitting by using a maximum number of clusters (k_{max}). Fig. 5.19c shows also the SOM similarity coloring. A figure showing all seismograms using this color scale for the seismogram background can be found in the Appendix (Fig. 8.6).

Furthermore, the false positive classification errors (CE^+) are shown in Fig. 5.18 (values for CE^- are similar). In order to avoid overfitting, we present the results for

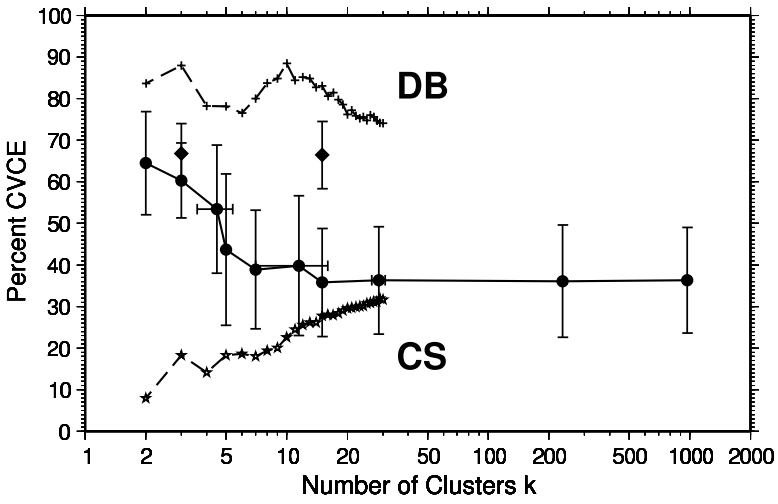

Figure 5.18: False positive classification errors ($CVCE^+$, circles) for earthquake data set (RDO). Diamonds represent $CVCE^+$ obtained after randomly distributing cluster labels using $k = 3$ and $k = 15$. Additionally, relative cluster validity criteria (DB index: crosses, CS: stars) are given. Note that for DB and CS no scale is given. Only the trend of the curve is of interest.

Figure 5.19: U-matrix, best clustering ($k = 6$), and SOM similarity coloring for earthquake data set (RDO).

a CV experiment ($CVCE^+$, circles). At about $k = 7$, comparable with the local DB index minimum, the lowest misclassification rates are reached. In other words, similarly good performance is achieved compared with previous classification results based on SOM prototypes, what would correspond to $k = 972$ in Fig. 5.18 (each prototype is a cluster). Therefore, using a limited number of groups, clustering is suitable for earthquake detection and seismic phase discrimination, although the U-matrix does not show such a distinct grouping as for the synthetics. However, more clusters than theoretically expected seismic phase classes are necessary to obtain the lowest achievable misclassification rate. For comparison, Fig. 5.18 shows again the results for a random distribution of cluster labels using $k = 3$ and $k = 15$ (diamonds).

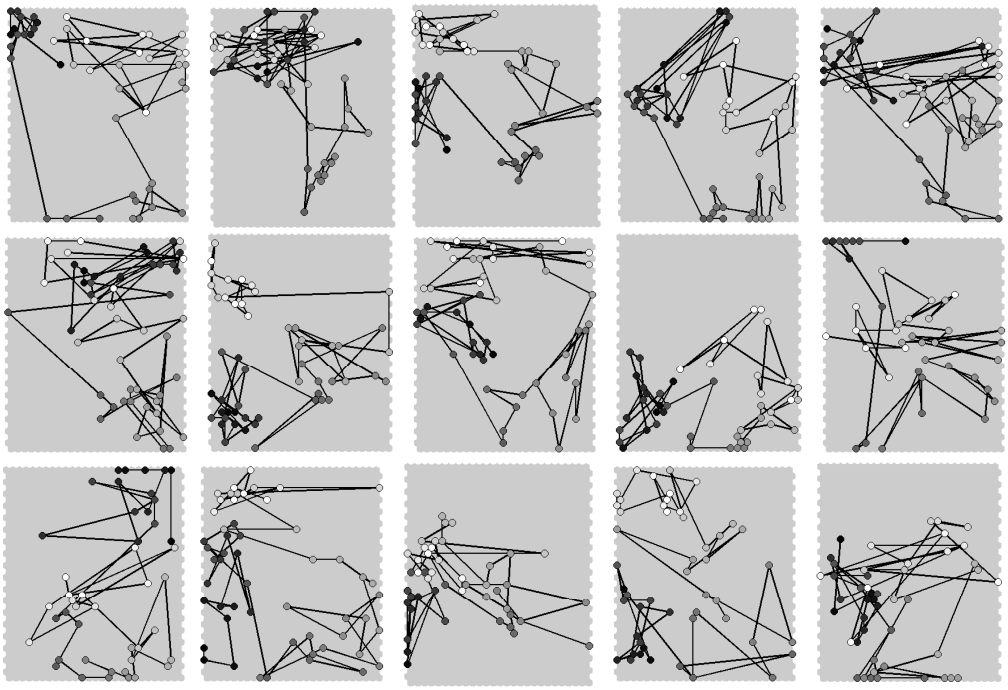

Figure 5.20: Hit histories for a selection of events using a trajectory on the SOM. Black symbols correspond to beginning and white to end of the seismogram. In between a gray scale is used.

Event Discrimination

In a last step we want to investigate the differences between individual event records, including the background noise time slices. For this purpose, the temporal SOM hit history of each record (55 feature vectors, 6 minutes) is parametrized by the 2D-coordinates of corresponding BMUs (110 parameters for each event). In doing so, a representation of the temporal context of time slices is obtained for each event. Fig. 5.20 shows the hit histories on the SOM using a trajectory visualization for a selection of events.

In order to visualize the record similarity with respect to the temporal dependencies between feature vectors, a new SOM is trained using 44 instances (number of events) each with 110 features (single list of x followed by y SOM coordinates of all 55 time windows). We will call this map *Trajectory SOM* (TSOM) in the following. Note that this approach is different to our main objective in this work, which is time window clustering using the corresponding feature vectors. The U-matrix of the TSOM in Fig. 5.21a shows the existence of clusters. In order to find the event characteristics which cause this grouping, we consider exemplary source depth, epicenter distance and temporal occurrence for each hit on the TSOM (Fig. 5.21) (Fig. 8.4 in the Appendix shows the distributions of all three properties for all 44 events). For the depths and distances on the TSOM in Fig. 5.21, we use a continuous gray scale for the symbol colors, where white corresponds to high (150 km, 11°), and black to low values (0 km, 0°). In order to visualize the occurrence, the data set is split into three time periods of one year each. The symbol colors are correlated with the season. The gray scale goes from black (December) to white (June), and back to

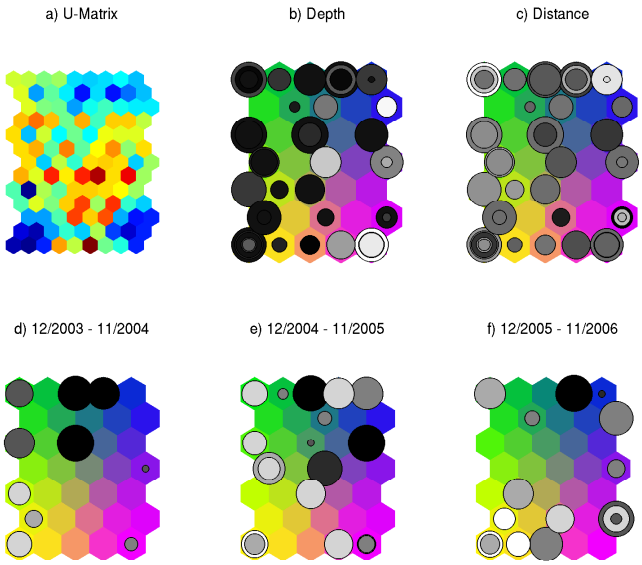

Figure 5.21: a) U-matrix of the TSOM trained using SOM trajectory of all events as input data (see Fig. 5.20). b-f) Event hits using a color scale which corresponds to particular event properties. Background color represents prototype similarity. In b) and c) black stands for low and white for high depth and epi-distance values. Symbol colors in d-f are correlated with season of occurrence. Gray scale runs from black (December) to white (June) and back to black (December). For clarity, symbol sizes are different for each data hit in all panels.

black (December). For clarity symbol sizes are different for each data hit.

The results show that there is a slight seasonal dependency. Events during winter time (black) are mainly located on the top-right TSOM, while during summer (white) the hits dominate on the bottom-left part. A likely reasons for that observation is the increasing power of microseismicity, and thus changing properties of, e.g., the amplitude spectrum, caused by increasing storm activity in the Mediterranean Sea during the autumn and winter months. Furthermore, also hypocenter depths seem to contribute to the event grouping. High source depths (light colors) can be found only on the right side of the TSOM, although the corresponding events occur within all seasons (see Fig. 8.4). There is no clear grouping effect due to the epicenter distances.

Fig. 5.21 shows how particular earthquake characteristics can be visualized on the TSOM. However, we want to emphasize that also many other event properties may be responsible for the observed grouping. In order to allow a further interpretation of the TSOM, the earthquake epicenters are plotted in Fig. 5.22. For the symbol color we use the BMU color obtained from the TSOM similarity coloring in Fig. 5.21b-f. For instance, we observe that the events of the deep Vrancea source region in Romania are located on to the right TSOM (blue and violet colors). Another event group can be roughly associated with the subduction zone of the Hellenic Arc (yellow color). Furthermore, also earthquakes, which occurred on the eastern part of the North Anatolian Fault and on the East Anatolian Fault, are similar (light green colors). However, the visualization in Fig.

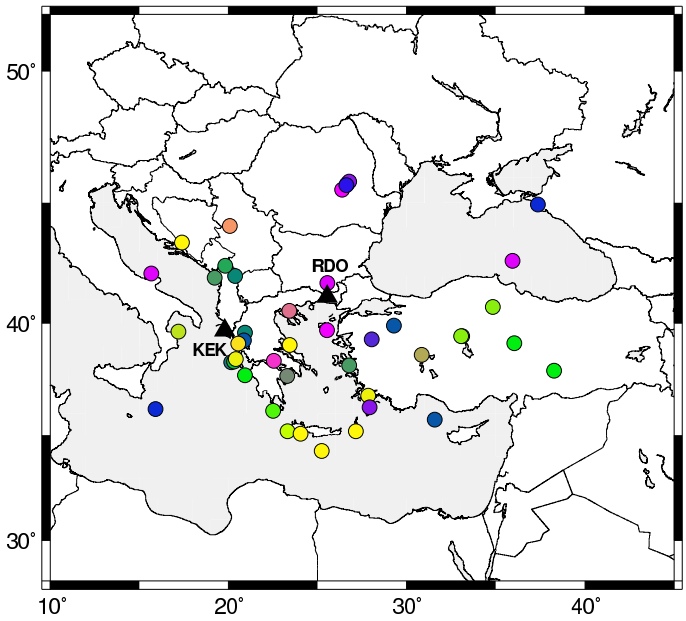

Figure 5.22: Earthquake epicenters of events recorded at RDO (circles). Symbols colors correspond to TSOM similarity coloring in Fig. 5.21b-f. Black Triangle is the location of station RDO. Additionally, location of station KEK is shown.

5.22 does not indicate a clear spatial earthquake clustering. There are events which are similar with respect to the TSOM coloring, however, they not occurred within the same source region.

5.2.3 Including a Second Receiver

So far we have solely considered the characteristics of earthquake recordings at a single receiver. In order to explore a possible site dependency, which may be crucial for generalization and classification, we will now integrate the station KEK into our analysis (See Fig. 5.22 for location). Feature Selection and SOM training is done for a single data set including both the 44 events for RDO and the 57 earthquakes for KEK. Fig. 5.23 shows the CV feature selection statistic and, additionally for comparison using white symbols, the final feature set of station RDO (see Table 5.16). In fact, there are only slight differences between features suitable for both receivers and only for the RDO station (see third feature of Method 3 and last feature of Method 6). Thus, the suitable set of features, which allows for the discrimination of earthquake wave phases, apparently does not depend on the site of the receiver.

Fig. 5.24 presents the hits of P wave, S wave, and noise time windows on the SOM, separately for station RDO (Fig. 5.24a,b,c) and KEK (Fig. 5.24d,e,f). As in Fig. 5.17, black hits indicate the P and S wave onset time windows, and white colors the noise after the event. At first view, the P and S wave hits are roughly located within similar SOM areas for both stations. However, going into detail, there are significant dependencies on the recording site:

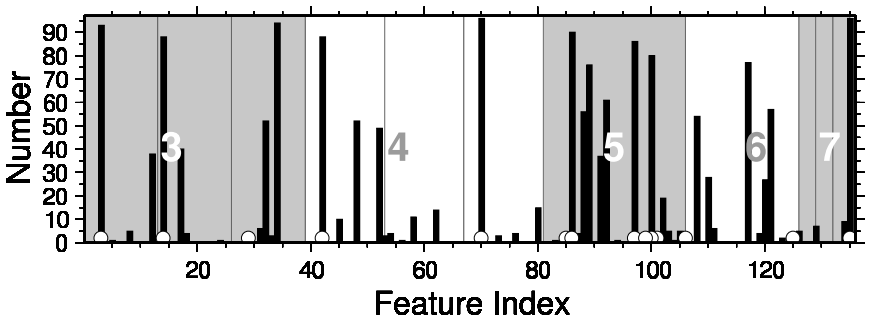

Figure 5.23: CV selection statistic for each feature for earthquake data set including station RDO and KEK. Background coloring distinguishes different feature generation methods (see gray and white numbers). Vertical lines denote frequency bands. White circles correspond to feature set obtained using only station RDO.

- Majority of P wave time windows are separated for KEK and RDO (right, center).
- Majority of S wave onset time windows are separated for KEK and RDO (left, center).
- No clear discrimination between P and S waves for KEK.
- Background Wavefield (Noise) before event is different for KEK and RDO.

Hence, we can conclude that the characteristic feature vector of P and S waves depend on the receiver or site properties. Furthermore, the discrimination between both phases can become more difficult. Since the separation between signal and noise is given (mainly located on top SOM), a simple event detection algorithm is still possible for two stations. However, for phase classification additional information has to be employed.

5.2.4 Discussion

In order to achieve a perfect classification performance, we require that all time windows of the record are classified correctly. Hence, relatively high misclassification rates and uncertainties are obtained for the conducted CV experiments. However, often only the onset of seismic phases is of interest. The visualization of signal hits on the SOM, and the experiment only on P wave detection, showed that improved classification accuracy can be achieved by SOM clustering.

Furthermore, we have only shown results for a limited data set within the previous sections, i.e. only regional earthquakes and only two stations. Therefore, we have to be careful with generalization. Without further investigations, it is not possible to transfer the findings for event and phase discrimination to a different region or to other types of earthquakes. This application example should rather give an impression about how such investigation can work, and which results may be expected. Nevertheless, our results show the potential of unsupervised learning to give important insights about suitable features and expected class discrimination by clustering. This information is an useful assistance to define labeled training data for the development of supervised classification algorithms.

In particular, for our data, we show that the event detection on station network can be achieved by a simple clustering approach. Nevertheless, phase discrimination only

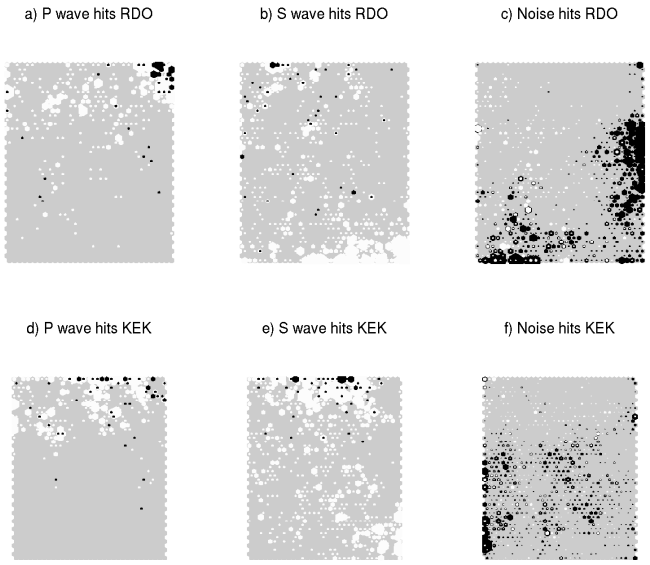

Figure 5.24: Data hits on the SOM which is trained using data from station RDO and KEK. a) to c) show class hits for RDO and d) to f) those of KEK. Black hits are P and S wave onset time windows. Noise after the event is also indicated by white colors.

works partially on individual receivers. Furthermore, our unsupervised feature selection methods is successfully applied to earthquake data. As for the synthetic data, combination of features, not only from one particular approach such as the frequency spectrum, has the potential to improve signal discrimination. Summarizing our insights for a short and simple recommendation, we suggest to take a feature set including the sonogram, planarity, and the H/V ratio to train and apply classifiers. Those classification algorithm should, furthermore, consider also context information between time windows. By integrating context-dependency, it may be possible to extend an automatic phase classification system also to a station network. In particular, we expect that performance and confidence can be further improved by applying an advanced supervised techniques such as Bayesian Networks. As an unsupervised tool to visualize and investigate context-dependency, we suggest the TSOM.

5.3 Volcano-Seismic Wavefield at Mount Merapi

We will now use a data set lasting several days. We do both keeping only selected transients as in Section 5.2 and full analysis of the seismic wavefield over a longer time period. For this purpose, we investigate seismicity recorded close to an active volcano. Volcano-seismic signals are important and mandatory for eruption forecasting and to assess the activity state of a volcano. Consequently, many studies have been carried out to detect and distinguish between event types at volcanoes worldwide (Minakami, 1960). In this section we apply clustering and visualization techniques to array recordings of seismicity at Mount Merapi, which is a high-risk volcano on Java, Indonesia. Volcanic activity within this region is caused by the subduction of the Australian tectonic plate under the Eurasian. The high hazard of Mount Merapi exists due to an observed dome formation and collapse cycle of 2-7 years. Furthermore, the area close to the volcano is densely populated with about one million inhabitants.

Table 5.21: Employed data and manually identified events for array KEN. Local time is given for July 1998.

Day	Hour	Number of Events			Day	Hour	Number of Events		
		Gug.	VTB	MP			Gug.	VTB	MP
2	12 a.m.	2	0	0	4	8 p.m.	0	1	0
2	2 p.m.	3	0	0	5	12 p.m.	0	1	0
2	5 p.m.	2	0	0	5	2 a.m.	0	1	0
2	6 p.m.	2	0	0	5	5 a.m.	0	1	0
2	7 p.m.	1	0	2	5	9 a.m.	0	1	0
2	8 p.m.	1	0	0	5	1 p.m.	0	1	0
2	9 p.m.	2	0	1	5	4 p.m.	0	1	0
2	10 p.m.	3	0	3	5	8 p.m.	0	1	0
3	6 a.m.	0	0	0	5	9 p.m.	0	1	0
4	8 a.m.	0	1	0	6	12 p.m.	0	1	0
4	10 a.m.	0	1	0	6	1 a.m.	0	1	0
4	1 p.m.	0	0	0	6	2 a.m.	0	1	0
4	3 p.m.	0	1	0	6	3 a.m.	0	1	0
4	6 p.m.	0	0	0					

We use seismic data from the beginning of July 1998 during a phase of high volcanic activity. Two arrays are employed (KEN and GRW), each consisting of three broadband three-component stations. The analyzed recordings are spread over five days (July 2 to July 6). For the first array (KEN), we chose 27 hours of data for which the occurrence of volcano-seismic events is known. We use hand-picked events for cluster evaluation for three types of signals, which have also been employed by Ohrnberger (2001) as training data for an automatic classification system. Shallow (<1.5 km) volcano-tectonic events belong to the first signal class (VTB, 5-8 Hz). The second class contains less impulsive multiphase events due to lava dome growth (MP, 3-4 Hz). Finally 1-2 minutes lasting rockfall-induced

Table 5.22: Employed data for array GRW. Local time is given for July 1998.

Day	Hour	Day	Hour	Day	Hour	Day	Hour	Day	Hour
2	8 a.m.	2	6 p.m.	3	4 a.m.	4	noon	4	9 p.m.
2	10 a.m.	2	8 p.m.	3	6 a.m.	4	1 p.m.	4	11 p.m.
2	noon	2	10 p.m.	4	7 a.m.	4	3 p.m.	5	1 a.m.
2	2 p.m.	2	midnight	4	9 a.m.	4	5 p.m.	5	3 a.m.
2	4 p.m.	3	2 a.m.	4	11 a.m.	4	7 p.m.	5	5 a.m.

signals called Guguran events are observed (1-20 Hz). MP events show a rapid amplitude decay with epi-distance. The characteristic property of the VTB event is a clear P but no S wave onset. Table 5.21 gives an overview of employed data and identified events for array KEN. For the second array (GRW), we make use of more regular sampling over two 24-hour cycles (July 2-3 and July 4-5, Table 5.22). Every second hour of the record, a total of 25 hours, is chosen to investigate the background ambient vibration wavefield.

A time window length of 1.7 seconds is used for feature generation in the following ($WINFAC = 7$, $f_{cent} = 4.1$ Hz). Features are computed in frequency bands between 0.8 and 16 Hz (See Appendix, Table 8.6). In the next two sections, data from array KEN is analyzed. First, we only consider time windows of known events in order to investigate signal discrimination. Subsequently, all 27 hours are employed for processing. Finally, the wavefield recording of array GRW is analyzed to explore long-term patterns in ambient vibration wavefields.

5.3.1 Selected Events From Array KEN

We use a set of 38 time sections in all, which divide into 16 VTB, 6 MP, and 16 Guguran events (Table 5.21). The section length is chosen so that background noise before and after the event has the same number of samples (time windows) as the event itself. Merging all events results in a number of 2367 time windows or feature vectors.

Feature Selection

For feature selection we may choose an appropriate value for Z_{limit}, since we know that the length of the expected pattern (the complete event) is longer than a single time window. By making use of the results of the theoretical runs test experiments in Section 4.2, we estimate an approximate value as a guideline. For 2367 time windows (N) and 38 events, where 50% corresponds to noise, we obtain a pattern length (L) of 31 samples on average. This corresponds to a ratio L/N of 0.013. By using these values, we are able to extrapolate estimates for Z_{test} from Fig. 4.2 for different numbers of patterns:

- 10 pattern: Z_{test} between 3 and 7
- 50 pattern: Z_{test} between 20 and 50

Table 5.23: Features selected for KEN recordings ($Z_{limit} = 20$). A selection of 38 time sections, including known signals and noise before and after the events, are used.

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
sonoh10	43.2	*b10*	31.5	*b4*	29.1
bbz	38.8	*ifz1*	31.5	*a_b4*	29.1
b1	36.3	*FQ1zh2*	30.9	*spacimz2*	28.5
a_b8	33.3	*b6*	29.8	*a_b2*	26.5
ratiolf	32.8	*spacimt1*	29.4	*H/V3*	26.0
prt2	31.9	*prz1*	29.2	*sonoh4*	22.8

- 100 pattern: Z_{test} between 10 and 20

For 38 patterns, $Z_{limit} = 20$ seems to be suitable. Using this parameter, we obtain the features listed in Table 5.23. We will now compare these features with those manually selected by Ohrnberger (2001). In that study, the vertical sonogram, features from the real 3c-covariance matrix (*rect*, *plan*, *inc* and azimuth), and vertical f-k analysis (including absolute power (*apz*), slowness and azimuth) have been considered as potential parameters. Finally, making use of a robustness criterion and visual assessment, the following relevant features have been chosen: *prz*, *apz*, *inc*, and *sonoz1* - *sonoz8*. Linearity and planarity have been excluded due to less discriminative power.

By comparing these results with our automatically selected features, similar conclusions can be made. The most relevant features are spectral attributes derived from the sonogram (*sonoh*, *bbz*, *ratiolf*). Furthermore, coherency (*pr*) is contained in both feature sets. We do not use absolute semblance power here because parametrization should be independent of the signal amplitudes. Polarization information is implicitly available from the ellipticity spectrum (*a_b*, *b*) and H/V in our feature set, and from angle of incidence (*inc*) in the set of Ohrnberger (2001). However, no direct polarization measures, such as planarity and linearity, belong to the final feature sets of both studies.

Cross-validation of Clusterings

In order to evaluate our estimate for Z_{limit}, we generate different feature subsets using values between 1.96 and 35. Subsequently, we conduct CV experiments on SOM learning and clustering. Table 5.24 shows class-averaged (VTB, MP, Guguran, Noise) cross-validated misclassification rates (false positive and false negative). The resulting classification errors are rather high. However, let us first only consider relative differences. In the following it will become clear that these high misclassification rates exist due to one particular class (MP). Regarding both the mean and the standard deviation, there is the tendency to favor $Z_{limit} = 20$, which corresponds to 18 features. Using more (decreasing Z_{limit}) or less features (increasing Z_{limit}), slightly increases classification errors. Thus, the estimation based on expected pattern length turns out to be meaningful.

Table 5.24: CV results for clustering volcano-seismic events. Different feature sets are tested which are obtained using different values for Z_{limit}.

Z_{limit}	1.96	3.0	10.0	**20.0**	25.0	30.0	32.0	35.0
$CVCE^+$	50.2	51.2	47.9	**45.7**	47.7	48.6	47.7	65.6
	±9.4%	±9.9%	±7.3%	**±2.6%**	±2.3%	±8.1%	±7.9%	±1.3%
$CVCE^-$	44.3	45.0	42.8	**40.9**	42.9	43.9	42.4	60.6
	±5.4%	±5.8%	±4.2%	**±2.4%**	±3.2%	±6.3%	±5.2%	±1.5%
No. of Feat.	22	22	18	**18**	14	7	6	3

Table 5.25: CV results for clustering volcano-seismic events using feature set corresponding to $Z_{limit} = 20$. Results for each event class are given. Furthermore, class-averaged errors, with and without MP events, are shown.

Class	$CVCE^+$	$CVCE^-$	Class	$CVCE^+$	$CVCE^-$
VTB	41.3±5.9%	12.8±7.3%	Noise	15.8±2.8%	27.9±5.1%
Gu	25.5±4.7%	22.8±5.0%	Av.	45.7±2.6%	40.9±2.4%
MP	100.0±0%	100.0±0%	Av. no MP	26.6±2.2%	20.9±1.9%

As mentioned above, we have to consider the individual, cross-validated class errors for each type of event to understand the clustering performance. In fact, Table 5.25 shows that it is not possible to find automatically the MP events. Excluding MPs from CV, we obtain a clearly improved class-averaged classification rate ($CVCE^- = 20.9 \pm 1.9\%$). Similar good classification rates are obtained for the Guguran events. On the other hand, VTB events exhibit high false positive and very low false negative errors. Hence, time windows are classified as VTB, which have been labeled as other classes, whereas almost all windows manually labeled as VTB are correctly classified. An explanation is that VTBs are lasting longer than suggested by the signal amplitudes. As for the earthquake data set in the previous section, this shows the difficulty to define a meaningful end of a seismic event. In fact, there is a continuous transition to the background wavefield, which also explains the higher false negative compared to the false positive classification errors of the noise class.

For a continuous mode application of the hidden Markov model-based classification system on five days of data, Ohrnberger (2001) obtained an averaged recognition rate of 67%. The best results have been found for VTBs (89%), followed by Gugurans (74%). The worst recognition rate has been observed for MP events (64%) due to ambiguous wavefield properties and weak amplitudes.

Since context-dependent information is included in hidden Markov models and the results correspond to a continuous data set, it is not possible to compare the recognition rates directly with our classification errors based on time window clustering. However, considering the qualitative findings for each class, our observations are in line with those of Ohrnberger (2001). VTB events also show the lowest false negative misclassification

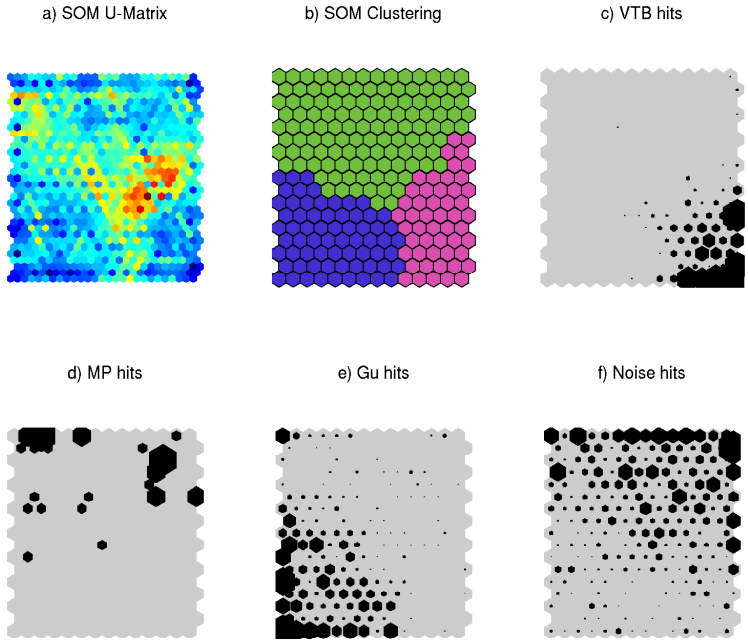

Figure 5.25: SOM visualizations for time sections including volcano-seismic signals recorded by array KEN.

rates, whereas worse results are obtained for Guguran signals, possibly due to their more heterogeneous character. In contrast to Ohrnberger (2001), we are not able to identify automatically even a single MP event. Here, the limits of simple clustering are evident, at least for array KEN.

Visualization

In this section we will interpret the quantitative results of the previous section by considering the SOM visualizations. Fig. 5.25 shows the U-matrix, best clustering, and the class hits on the SOM. The U-matrix clearly suggests a first order partition into three clusters, which is also the most favorable clustering regarding the DB index (Fig. 5.25b). In fact, considering the DB indices between $k_{min} = 2$ and $k_{max} = 15$ (Fig. 5.26), shows that the minimum at $k = 3$ is very prominent. The CS criterion is less suitable to obtain the number of clusters suggested by the U-matrix since no clear minimum is obtained.

Going into more detail, the U-matrix shows that the upper cluster in Fig. 5.25b (green cluster) is less homogeneous and more structured than the others. The data hits in Fig. 5.25f indicate that this cluster corresponds to the background noise. Hence, the heterogeneity can be explained by the character of ambient vibrations (superposition of signals) and also by the observation that the majority of MP time windows are located within this cluster. There is no part of the noise cluster which is clearly dominated by MP hits. Thus, discrimination of MP events is impossible. On the other hand, VTB and Guguran events seem to be compact and well-discriminated clusters. However, Guguran time windows are, as expected, more heterogeneously distributed over the SOM. Particularly, there are hits in the noise cluster. Furthermore, noise hits within the VTB cluster (violet) can be

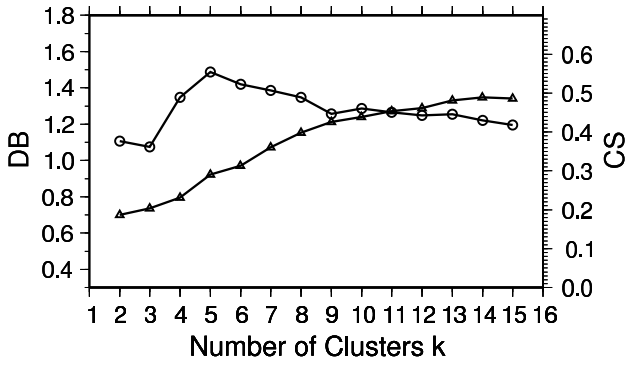

Figure 5.26: Internal cluster validity criteria for different number of clusters (k) regarding SOM clustering of volcano-seismic signals recorded by array KEN. Circles indicate DB and triangles CS criteria.

observed. As mentioned above, these time windows actually belong to the VTB event. Therefore, biased false positive errors for the VTB class and biased false negative errors for noise are produced due to the uncertainties of manual labeling.

From the SOM component planes in Fig. 5.27, we can derive signal properties by considering the features for each cluster. Three different types of features may be distinguished. The first group allows signal detection in general, i.e. discrimination between noise and events. Either features have consistently low values for signals and vary for noise (*sonoh10*, *bbz*, *b1*, *spacimt1*), or show high values for the event clusters and are low or intermediate for noise (*H/V3*, *prt2*, *ifz1*). These observations are in line with the domain knowledge, because it is intuitively expected that seismic signals have, e.g., a lower bandwidth (*bbz*), more energy on horizontal components (*H/V3*), and higher coherency (*prt2*) than the background wavefield. Furthermore, there are features which discriminate between the VTB and Guguran clusters, and not necessarily between noise and the signal clusters. The VTB cluster is mainly determined by features *b4*, *ratiolf*, and *sonoh4* (high values), whereas for Guguran time windows, high values for features *FQ1zh2*, *a_b8*, and *b6* are observed. Thus, higher frequencies for the Guguran events is the main difference compared to VTB events. This is also confirmed by our expectation. While for volcano-tectonic events (VTB), dominant energy between 5 and 8 Hz is observed at Mt. Merapi, for Guguran events frequencies can go up to 20 Hz. Moreover, higher bandwidth is expected for Guguran events. In fact, values for *bbz* are slightly higher than for VTBs. However, the dominant pattern here is a higher bandwidth of noise. Finally, there are features which do not discriminate one particular class, but rather different characteristic within a class (*prz1*, *spacimz2*, *b10*, *a_b4*). They may indicate seismic phases or patterns in the background wavefield.

5.3.2 Complete Data From Array KEN

In the next processing step, 27 hours of data from array KEN are used for feature selection, SOM learning and clustering. Most hours include identified and labeled volcano-seismic events (see previous section). Furthermore, unlabeled events and other transients are expected. Since volcano-seismic signals represent only a small part of the training data, it

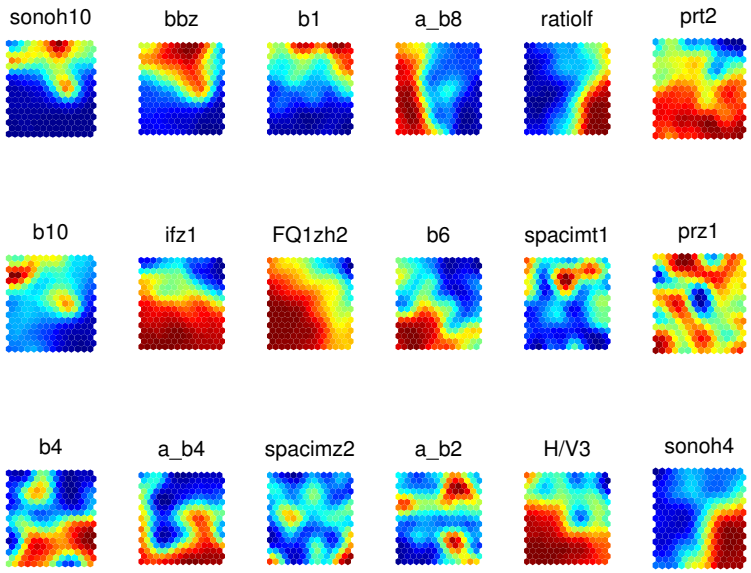

Figure 5.27: Component planes for SOM in Fig. 5.25.

Table 5.26: Features selected from KEN recordings ($Z_{limit} = 20$). 27 hours of data are used.

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
sonoh10	158.2	*prz1*	143.0	*domfz*	119.8
a_b10	149.3	*spacimr1*	138.5	*bbz*	117.1
prr2	146.3	*sonoh1*	135.7	*sonoz1*	100.6
ratiolf	145.9	*a_b6*	135.4	*H/V3*	94.2
prz2	144.7	*a_b4*	128.3	*H/V1*	60.2
prr3	143.9	*a_b1*	125.4	*sonoz5*	52.2
prt1	143.8	*FQ1zh2*	123.7	*ellzn2*	38.4

is necessary to investigate the resulting effect of feature selection and clustering on signal detection.

Table 5.26 shows the obtained feature set using an appropriate Z_{limit}. From the size of the data set (56241 samples) and assuming again a pattern length of 31 samples, we obtain a ratio L/N of 0.0005. Extrapolation from Fig. 4.2 again leads to $Z_{test} \approx 20$. The 21 selected features are comparable with those found for the reduced data set (Table 5.23). The majority of features either occur in both sets (7 features) or are of the same type, but computed for close frequency bands (*a_b*, *sonoh*). The difference is mainly that direct polarization information is also represented (*b*, *ellzn2*). Nevertheless, we have further evidence of the robustness of our feature selection procedure regarding the size of the data set.

Fig. 5.28 presents the SOM visualizations. The U-matrix shows no clear clustering,

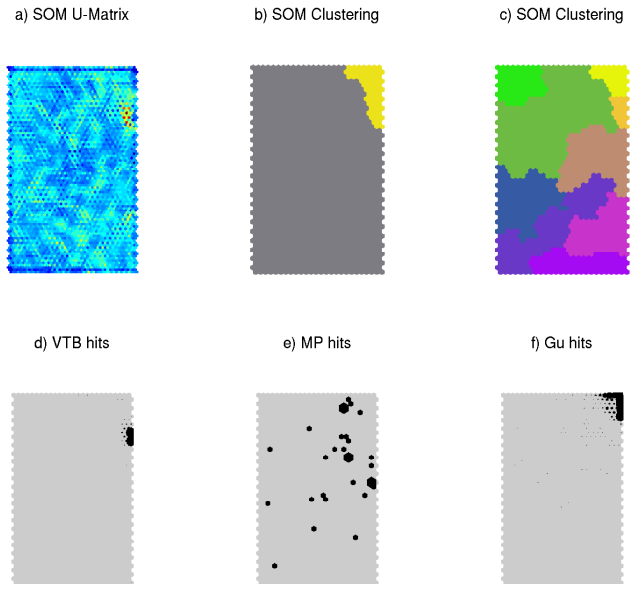

Figure 5.28: SOM visualizations for data recorded by array KEN. In b) $k = 2$ (best clustering) and in c) $k = 9$.

rather a very complex structure with areas of low and high data density. The only striking feature on the SOM is located close to the top righthand corner. Red colors in Fig. 5.28a may indicate a cluster border, at least in one direction. In fact, in accordance with this pattern, the best clustering, with respect to the DB index, yields two clusters (Fig. 5.28b). The SOM hits of labeled time windows in Fig. 5.28d-f (VTB, Guguran, MP) show that the yellow cluster contains VTB and Guguran events, whereas MP is again distributed over the entire map. Moreover, VTB and Guguran events can be discriminated as they cover different areas on the SOM. In fact, Fig. 5.28c and the cluster dendrogram in Fig. 5.29 show that for $k = 9$, both event types are located in different children of the signal cluster ($k = 2$). Hence, it is possible to find at least two volcano-seismic signal classes (VTB and Guguran) by clustering, even when noise time windows dominate the data set. On the other hand, although there is a trend for more frequent MP hits close to the VTB and Guguran groups, no distinct MP cluster is found. The clustering for the remaining SOM is less meaningful for $k = 9$, i.e. cluster borders do not necessarily fit the data structure. However, this partition can be used as a first order grouping which highlights similarities in the data space.

Quantitatively, we obtain $CE^+ = 97.8\%$, $CE^- = 7.4\%$ for VTB and $CE^+ = 97.9\%$, $CE^- = 2.9\%$ for Guguran time windows ($k = 9$). Thus, as already suggested by Fig. 5.28, high recognition rates are obtained for all known event time windows. The high false positive classification errors show that, as expected, events are found, or time windows are classified as signals, which do not belong to the labeled training data set.

Finally, we present eight seismogram plots (vertical component) in Fig. 5.30 and 5.31, each lasting one hour, using the SOM clustering colors of Fig. 5.28c. Hence, we visualize the patterns discussed above within their original temporal context. Besides the clear imaging of VTB (orange) and Guguran (yellow) events in all hours, some other

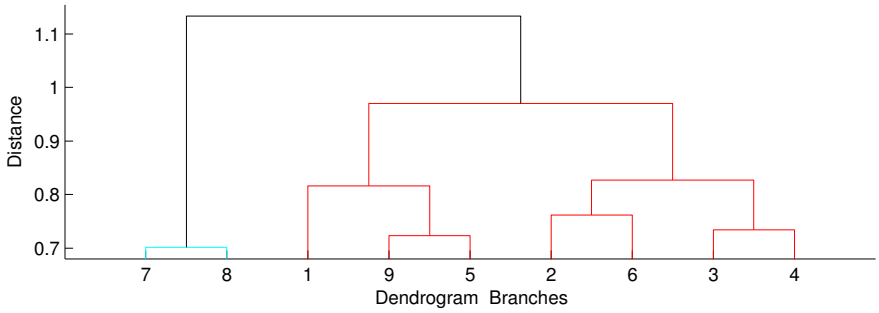

Figure 5.29: Dendrogram of SOM clustering for data recorded by array KEN.

interesting patterns can be observed. In the daytime (12 a.m., 3 p.m., 5 a.m., 9 a.m., local time), transients (light green) occur which are not observed on all three receivers simultaneously. It is difficult to distinguish between VTB, Gugurans, and those signals by considering the raw amplitudes only. However, since the signal properties differ clearly, visual discrimination in the seismogram is made possible. The background wavefield in the daytime is more heterogeneous. However, dark green labeled time windows dominate. Furthermore, during the night, blue and violet colors highlight the background noise. Except for volcano-seismic events, no other transients occur. The obvious interpretation is that man-made noise, originated close to one particular receiver (light green) or all around Mt. Merapi (dark green), dominates in the daytime. In fact, the lowermost panel in Fig. 5.30 (5 a.m.) shows very nicely the successively increasing human activity in the early morning, shortly before sunrise in tropical regions.

5.3.3 Background Wavefield Analysis at Array GRW

In order to investigate the background ambient vibration wavefield in more detail, we will now use 25 hours of unlabeled data from array GRW regularly sampling two days. After feature selection, a SOM is learned, whose U-matrix, best clustering and SOM similarity coloring based on PCA projection (see Section 4.4) are presented in Fig. 5.32. The U-matrix shows a very complex structure. The bottom right part of the SOM is more clustered compared to the rest. In fact, the most favorable clustering divides the data set into two clusters, one covering more or less this area (red cluster in Fig. 5.32b). As we will see later, this part of the SOM corresponds to time windows containing volcano-seismic events. Thus, as for array KEN, signals of interest are recognized by clustering.

For an interpretation of all patterns, we consider Fig. 5.33 and 5.35. Fig. 5.33 shows the daytime-dependent distribution of SOM hits using cumulative data from both days. Although the spread is high, a clear daily trend can be observed. The majority of hits are moving clockwise over the SOM. Starting from the left side of the SOM at night (yellow in Fig. 5.32c), the top SOM corresponds to early morning (green), the right side to noon, the bottom part to evening, and finally the left side is again captured.

For the seismogram background coloring in Fig. 5.35, we combine the red cluster color from Fig. 5.32b and, for all other clusters, the colors given by Fig. 5.32c. The figure clearly shows the daily cycle. A striking exception from this trend is that on the second day, a clear day-like behavior can be observed during midnight (23.00 h - 01.00 h), which does not exist for the first day. Since the daily trend is mainly controlled by human activity in the

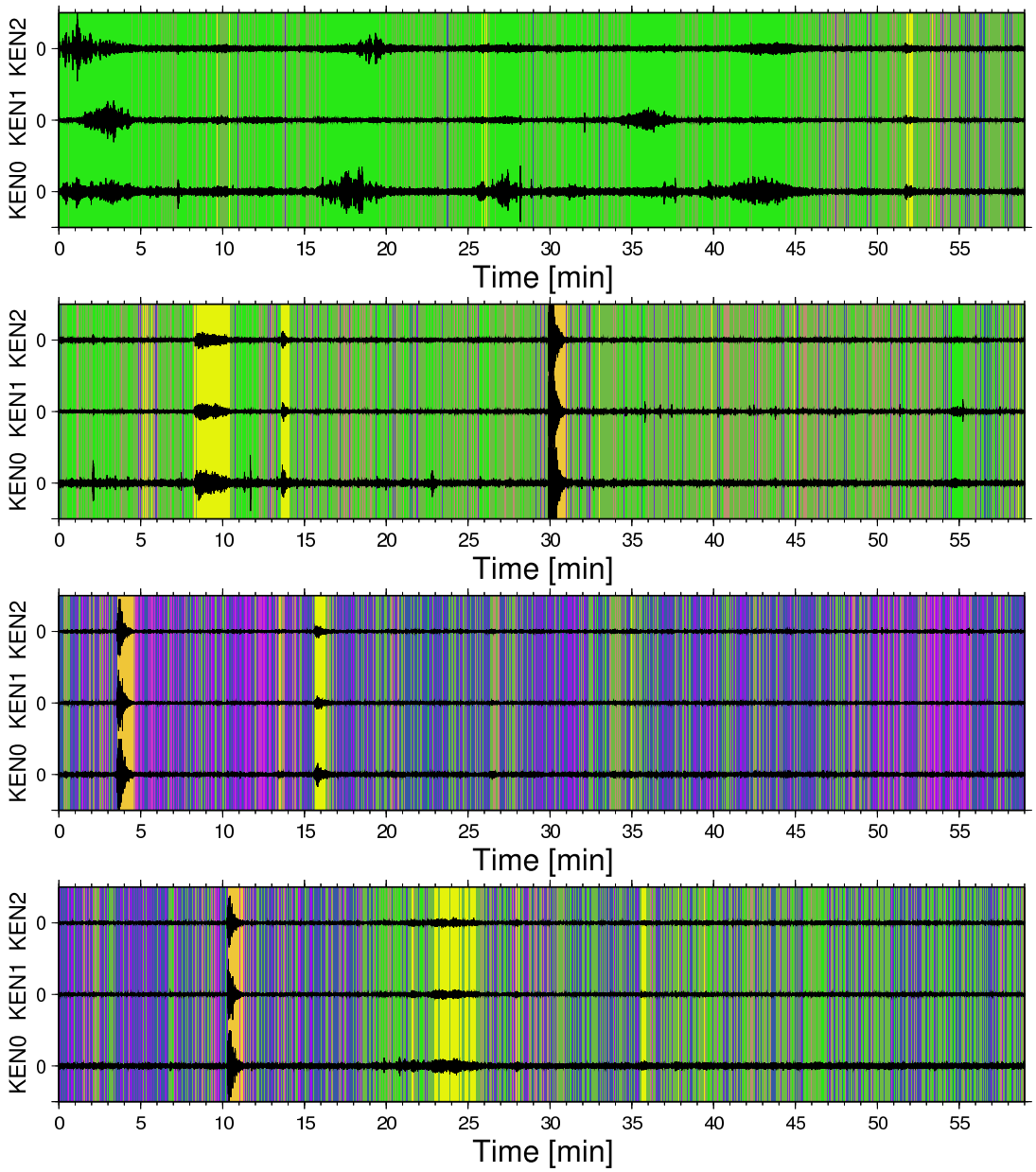

Figure 5.30: Vertical seismograms of array KEN. Background coloring corresponds to SOM clustering in Fig. 5.28c. Starting with the uppermost panel, records for 12 a.m., 3 p.m., 8 p.m., and 5 a.m. are shown, for July 4 and 5. Seismograms are bandpass-filtered in the frequency range where features are computed (0.3 - 19 Hz).

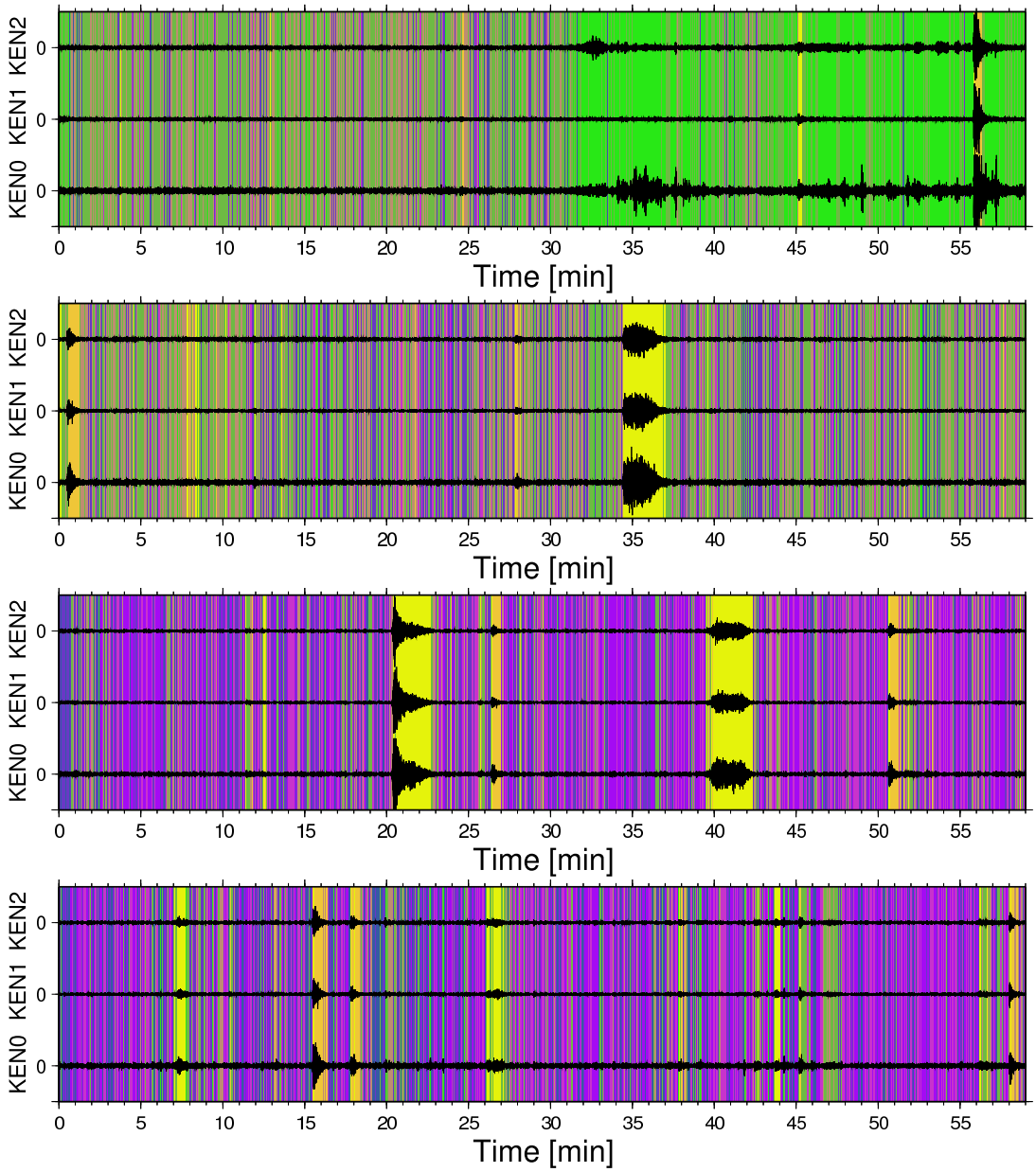

Figure 5.31: Continuation of Fig. 5.29. Starting with the uppermost panel, records for 9 a.m., 4 p.m., 9 p.m., and 3 a.m. are shown, for July 5 and 6.

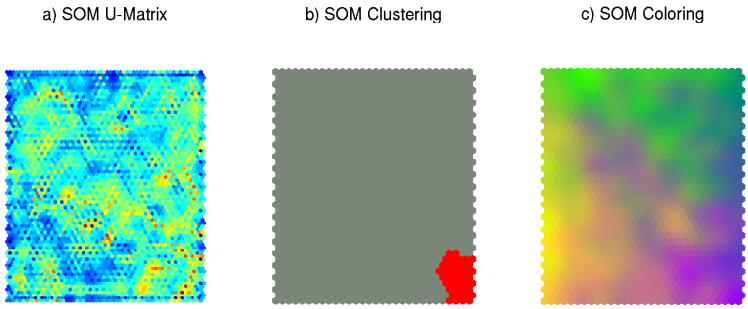

Figure 5.32: SOM visualizations for data recorded by array GRW. In b) $k = 2$ (best clustering).

daytime working hours (6 a.m. - 6 p.m.), which becomes noticeable, e.g., by the increasing amplitudes and more non-volcanic transients (light green colors), a possible explanation is the existence of man-made noise due to activities at night. In particular, close sources may exist due to the activity of researchers observing or climbing the volcano. However, the latter reason is not very likely during midnight and, moreover, during a phase of high volcanic activity. Another explanation could, therefore, be atmospheric phenomena such as wind and rainfall, or other kinds of signals originated by the volcano. However, no noticeable rainfall events are observed at station GRW during these hours (see Appendix Fig. 8.4).

Most events of volcanic origin belong to the red cluster independent of the time of day. Nevertheless, the discrimination is not as clear as for array KEN, particularly during afternoon and evening hours (violet background).

In a last step, let us consider the SOM component planes in Fig. 5.34 in order to determine the wavefield properties which control the intuitive visualization in Fig. 5.35. The daytime ambient vibration wavefield is characterized by higher frequencies compared to the nighttime and the volcanic events (*domfh*, *sonoz8*), whereas for the background wavefield at night, low frequencies dominate (*sonoz1*). Furthermore, polarization attributes (*b1*, *inc1*) suggest higher ellipticities or a less pronounced single direction of polarization at night. This observation also fits with lower linearities (*rect*, *a_b6*) compared to high values for volcanic events (P and S waves), and variable values for the day noise (close as well as distant sources). Hence, nighttime ambient vibrations show an expected Rayleigh wave-like behavior due to the dominant contribution of oceanic microseismicity caused by distant sources.

Besides the high polarization, volcano-seismic events are also discriminated by their high radial coherency (*prr2*) and high to intermediate frequencies on the horizontal components (*FQ1zh3*, *sonoh7*). On the other hand, the vertical coherency (*prz1*) at night corroborate the Rayleigh wave hypothesis. The differences between morning, noon, afternoon, and evening are more difficult to interpret than just comparing day and night. These patterns are controlled by changing frequencies (e.g. *domfh*, *sonoz8*), coherency (*prz1*) and polarization (e.g. *H/V2*, *rect2*) over the day. This may be due to gradual changes in the anthropogenic source distribution or in the meteorological conditions. For instance, rainfall has a clear daily cycle in tropical regions (Fig. 8.4).

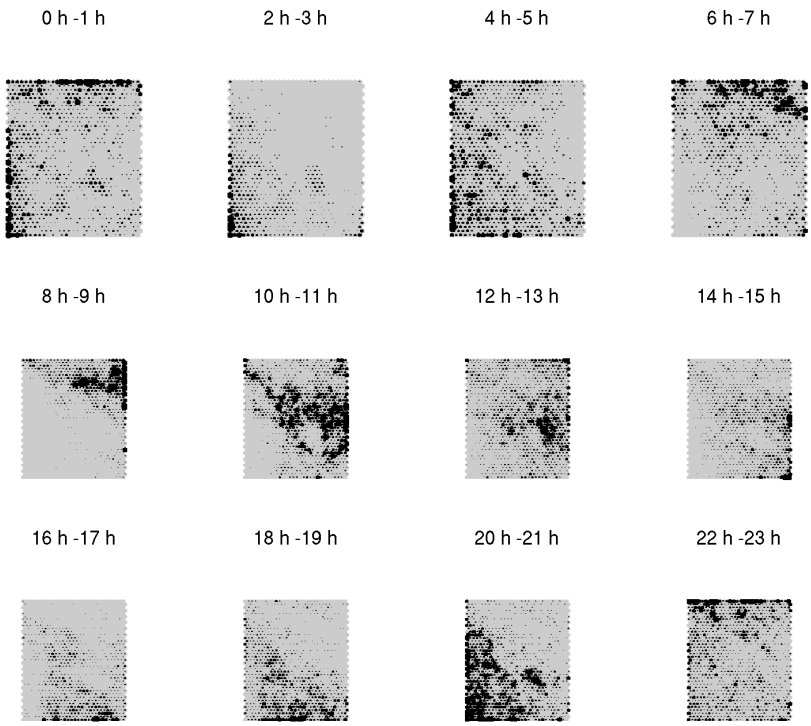

Figure 5.33: Distribution of data hits on the SOM over a day for array GRW. Numbers indicated the hour of day.

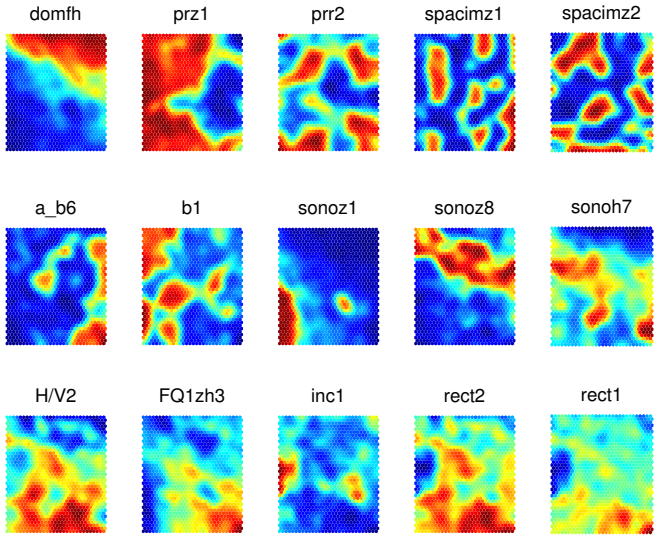

Figure 5.34: SOM Component Planes for data recorded by array GRW.

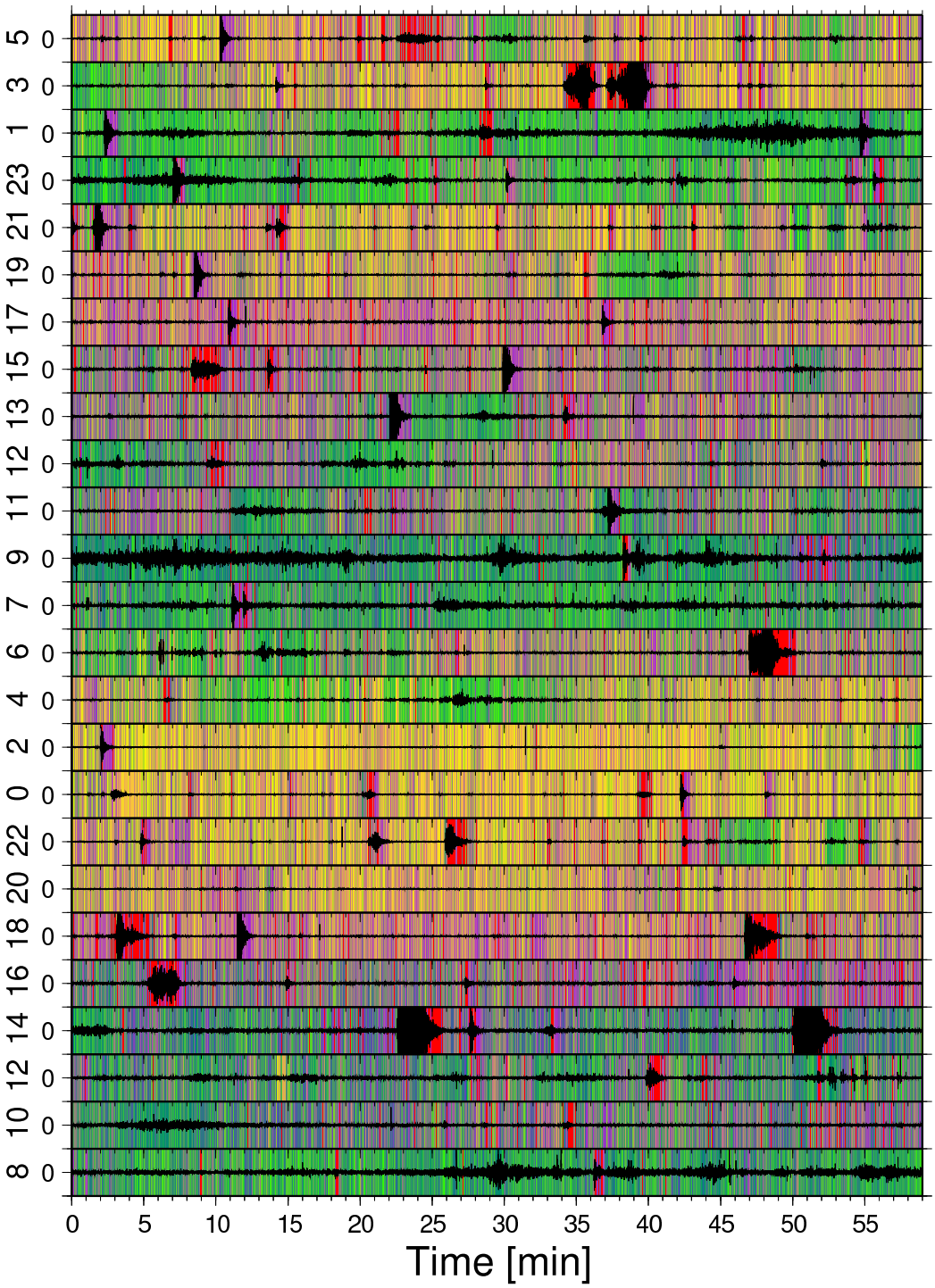

Figure 5.35: Wavefield behavior over two days for array GRW. Seismogram background coloring corresponds to SOM similarity coloring shown in Fig. 5.32c. For the cluster of volcano-seismic signals, the red color from Fig. 5.32b is used. Only vertical records at a single station are shown. Y-axis labels indicate hour of day. Seismograms are bandpass-filtered in the frequency range where features are computed (0.3 - 19 Hz).

5.3.4 Discussion

The results show that identification of typical volcano-seismic signals by means of a simple time window clustering approach is possible, independent of the size of the data set. Even in a large set, where the background wavefield dominates, characteristic signals for VTB and Guguran events can be discriminated. Furthermore, SOM-based mapping of trends within the wavefield can be done intuitively. Therefore, clustering of volcano-seismic records can be an easily realized first processing step to get an overview of the available data, without manual inspection of all seismograms. By including information such as time of day and manually identified events, the learned and clustered SOM can be easily interpreted. For instance, in order to monitor the activity of a volcano, SOMs can be used by continuously projecting new data on the map. However, for automatic classification and warning, supervised classifiers are required, which give a more precise prediction about an occurred event type.

Table 5.27: Characteristics of ambient seismic vibration data sets used in this section.

Site	Analyzed Record(s)		Array	Characteristics
	Short-term	Long-term		
Pulheim	2000/09/06: 2:00 - 2:55 p.m.	2000/09/06: 10 a.m. - 6 p.m., 52 min*	12 stat.	well studied
Lörrach	2002/04/11: 4:40 - 5:30 p.m.	-	9 stat.	valley, close motorway
Hamburg	2006/04/21: 9:04 - 09:50 a.m.	-	8 stat. (KGVB)	close to airport
Lüneburg	2007/09/04: 1:00 - 2:00 a.m.	2007/09/03-04: 6 p.m. - 7 a.m. 9 - 11 a.m. 1 - 3 p.m. & 5 p.m.	5 stat. (East)	day and night records
Colfiorito	-	2002/07/29: 1 - 3 p.m., 55 min* 2002/07/30-31: 5 p.m. - 6 a.m., 59 min*	12 stat. (B&D)	rural, day and night records

* used data from each hour

5.4 Ambient Seismic Vibration Wavefields

In the previous section we have already investigated short- and long-term patterns in a seismic background wavefield. We will now continue with ambient vibration analysis using recordings of array measurements at five different sites. The experiments have been carried out in the context of local subsurface exploration. Table 5.27 summarizes the relevant information about all employed data sets, including site characteristics and lengths of the processed records. As formulated in Section 3.5, our goal is to investigate the effect of long- and short-term patterns on the results of common ASV analysis methods. In particular, we aim to find out whether a wavefield decomposition is possible and reasonable. In contrast to the previous applications, no direct evaluation of results, e.g. by means of cross-validation, can be done since no label information is available for classification. Thus, making use of the findings of the previous sections about algorithm parameters and reliability of processing, we carry out a completely unsupervised investigation. However, the final results, i.e. dispersion curves and H/V ratios for the decomposed wavefield, will be qualitatively evaluated based on domain knowledge. For that it is sufficient to compute dispersion curves for a limited number of frequency samples only (about 10 to 20). Nevertheless, note that usually more points are used in ASV analysis and Vs inversions.

Depending on site character and available data (Table 5.27), different aspects are considered for each data set. For the urban Pulheim site close to the city of Cologne, we know that good and stable results for both the dispersion curves and the H/V ratios have been obtained in various studies (see references in Köhler et al., 2007). Hence, as a first step we aim to explore whether a wavefield decomposition based on short-term patterns

(<1 hour) makes sense to further improve results for a nearly "ideal" ASV data set. Since a longer, continuous record over several hours is available, also long-term patterns in the daytime will be investigated. In contrast, for the Lörrach site close to Basel and the measurements in Hamburg, only records of about one hour length are available for each array. In fact, this is typical for standard ASV experiments. Furthermore, the challenge for the Lörrach site is the presence of strong surface and subsurface topography and a close, busy motorway. Therefore, 2D or 3D effects and a directional source distribution may be expected. On the other hand, a close airport and an entry lane for starting and landing airplanes directly above the array may disturb the results for the Hamburg site. The impact of acoustic aircraft noise and, furthermore, closely passing persons will be explored. For the last two sites (Colfiorito and Lüneburg) one-day or longer lasting records are available for the investigation of long-term wavefield patterns. Furthermore, Colfiorito is a rural and isolated site in contrast to the other locations.

Within the next section we will address short-term patterns for the Pulheim, Lörrach, Hamburg, and Lüneburg sites. Section 5.4.2 will focus on long-term patterns for site Pulheim, Lüneburg, and Colfiorito. All features are generated for 10 frequency bands (410 features in all). Details on bands and finally selected feature sets for each data set are given in the Appendix. Note that we choose the frequency bands for feature generation by taking into account the suitable range of f-k or SPAC analysis, since our goal is improvement of those techniques. Therefore, the array aperture controls the frequency limits. However, as mentioned in Section 4.1, there are also features which parametrize the complete frequency range (e.g. dominate frequency, *domf*).

5.4.1 Short-Term Patterns

After feature generation ($WINFAC = 5$) and selection ($Z_{limit} = 1.96$ due to unknown pattern length), the SOMs are trained and clustered. Results are presented as U-matrix and best clustering with respect to the DB index. Whenever necessary, e.g. due to an unsatisfying fit of the U-matrix, another more meaningful final clustering is chosen. Subsequently, the clusters are interpreted regarding their temporal context in the seismogram and with respect to the SOM component planes. The clustering is then used to estimate dispersion curves and H/V spectra using time windows of each grouping separately.

Pulheim

In Fig. 5.36 the SOM U-matrix, a SOM clustering, and component planes of six representative features are shown. Those features are chosen visually for easy interpretability with respect to the obtained SOM clustering. We choose six clusters as one possible solution, since the U-matrix suggests more clusters than given by the lowest index ($k = 2$). Finally, the cluster memberships of time windows are highlighted in Fig. 5.37.

The seismogram amplitudes in Fig. 5.37 show one-minute transients, visible on all three receivers, which correspond to the greenish clusters (top of the SOM). Considering the component planes in Fig. 5.36, except for the small one in the center which does not contain many time windows, these clusters are characterized by high frequencies on the horizontal and low frequencies on the vertical components (*domfz*, *sonoh10*, *sonoz9*). This is an expected behavior for man-made signals generated by sources close to the array. The described properties are more distinctive for the light green cluster, which, furthermore,

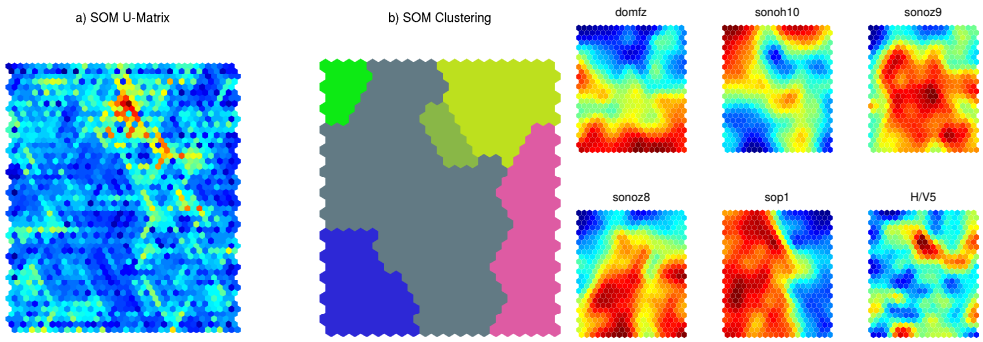

Figure 5.36: SOM visualizations for Pulheim recordings. Component planes for six representative features are shown.

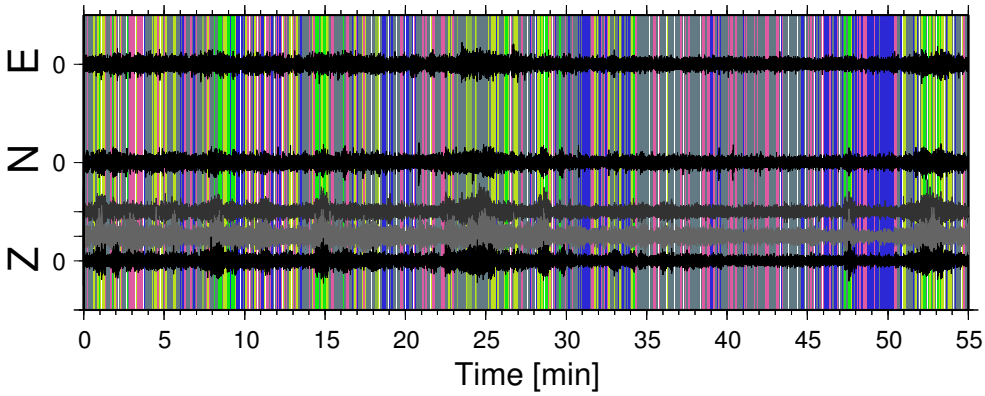

Figure 5.37: Visualization of significant short-term patterns in the Pulheim recordings. Background coloring corresponds to SOM clustering in Fig. 5.36. Seismograms of three receivers, bandpass-filtered between 0.1 and 4 Hz, are given. For one station (black) all three components are shown.

shows higher strength of polarization (*sop1*). For the other patterns, i.e. the transition between blue, gray and violet clusters, no clear interpretation can be given considering the seismograms alone. The component plane for strength of polarization (*sop1*) suggests that the gray and blue colored time windows have higher polarization. This may indicate a higher contribution of distinct, man-made sources and less random noise. Furthermore, this is confirmed by higher horizontal frequencies (*sonoh10*) for the gray cluster compared to the blue and violet ones.

Nevertheless, all recognized patterns are only relevant for practical aspects of ASV analysis when there is a significant effect on estimated dispersion curves and H/V spectra. Fig. 5.38 clearly shows that this is not the case. Considering mean values and standard deviations, slowness estimates do not differ significantly for all clusters, and variances are not reduced on all three spatial components. Hence, both the transients and the background wavefield yield the same results independent of the amplitude. An exception is the dispersion curve for the small green cluster mentioned above. There are probably too few time windows for a stable estimation. Moreover, regarding the observed horizontal dispersion curves, there is no trend to more Rayleigh or Love wave-like clusters. Consid-

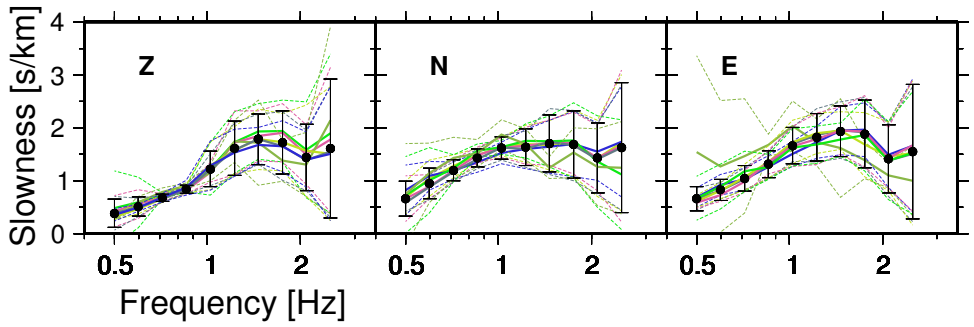

Figure 5.38: Mean dispersion curves and standard deviations (dashed) obtained from averaging slowness values for time window clusters (Pulheim). Colored curves correspond to SOM clusters in Fig. 5.36. Black symbols show the results for all time windows.

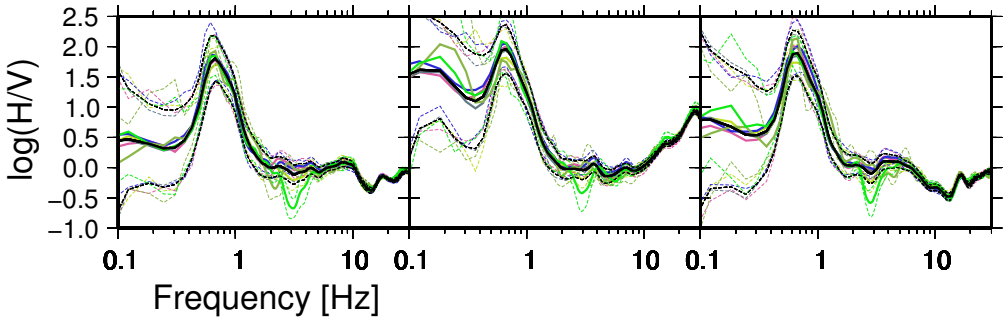

Figure 5.39: H/V spectral ratios and standard deviations (dashed) for SOM clusters (colors from Fig. 5.36) and using all time windows (black). A selection of three receivers is shown for site Pulheim.

eration of the position and amplitude of the main peak in the H/V ratios in Fig. 5.39 leads to the same conclusion. The only striking feature is located close to the upper limit of the frequency band employed for feature generation. Due to a lack of resolution, it is of no relevance for f-k analysis. The light green cluster shows a clear trough in the H/V spectra at 3 Hz. Interestingly, this corresponds very well to the trough in H/V spectra of local earthquakes observed by Ohrnberger et al. (2004). Thus, a few time windows, which contain transients from distinct sources, favor the determination of the H/V structure at higher frequencies, while the superposition of signals or lack of energy leads to a smoothing of the spectrum. This observation may be considered as a link between passive and active experiments.

Lörrach

The dominant pattern in the Lörrach data set is the clear cluster border on the bottom SOM separating the violet and magenta SOM units from the rest (Fig. 5.40, $k = 5$). Considering the north components of all receivers in Fig. 5.41, only a few time windows belong to those clusters. In fact, the violet colored time windows correspond to very short transients which are only observed on one particular station (fifth receiver from the bottom). Furthermore, this cluster is characterized by low frequencies and highly linear

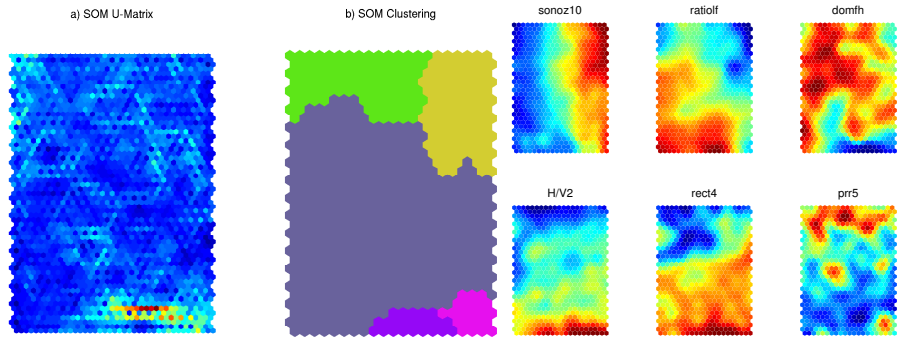

Figure 5.40: SOM visualizations for Lörrach recordings. Component planes for six representative features are shown.

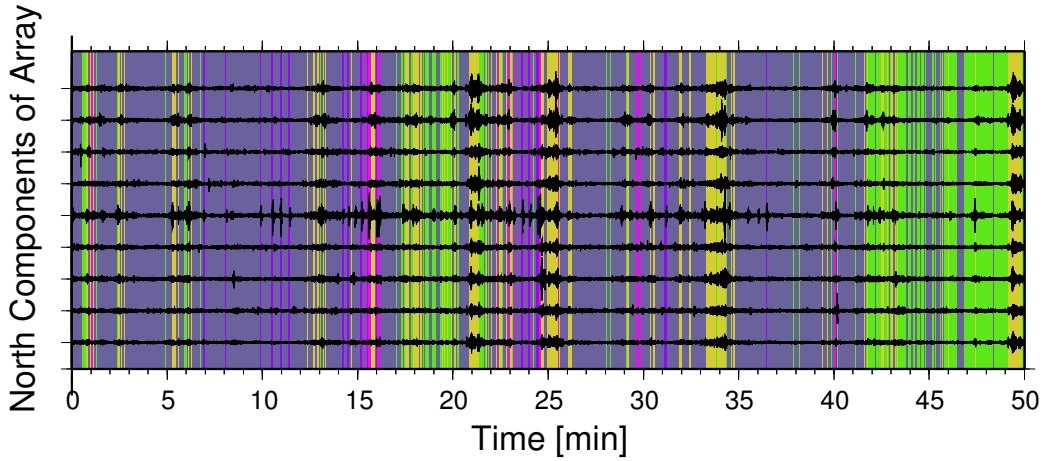

Figure 5.41: Visualization of significant short-term patterns in the Lörrach recordings. Background coloring coresponds to SOM clustering in Fig. 5.40. For all stations the north components are shown. Seismograms are bandpass-filtered between 0.3 and 7 Hz.

horizontal signals. Therefore, those signals seem to contain near-field displacement and tilt caused by very close sources, probably the footsteps of passing people inside the array.

Furthermore, there are one-minute transients which propagate over the entire array (yellow cluster). They show high coherency, intermediate linearity, and high frequencies also on the vertical components. Therefore, those signals are very likely related to strong or close man-made sources outside the array. Finally, an amplitude-independent pattern can be observed which occurs between the 17th and 22nd, and again between the 42nd and 50th minute of the record (green cluster). As the previously mentioned class, those signals seem to be coherent. In contrast to the remaining wavefield (blue and gray clusters), a contribution of high frequencies on horizontal components exists (*domfh*), vertical energy dominates compared to horizontal components for lower frequencies (*H/V2*), and signals are non-linearly polarized (*rect4*).

Fig. 5.42 shows the dispersion curves for each cluster. While vertical dispersion due to propagating Rayleigh waves does not differ significantly, a clear effect on the horizontal components can be observed. As expected, the violet cluster shows the most unstable

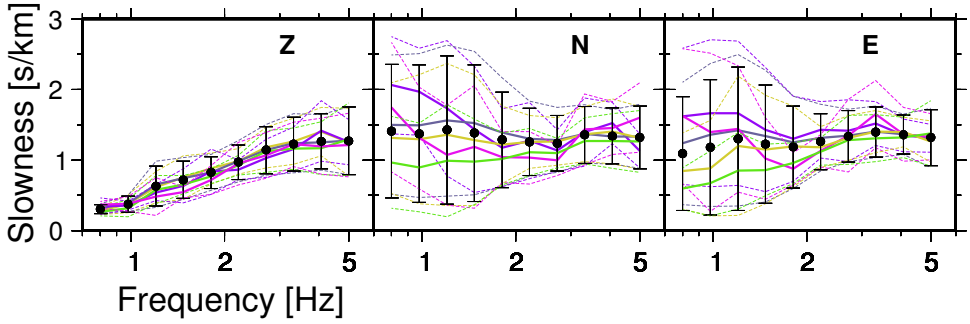

Figure 5.42: Mean dispersion curves and standard deviations (dashed) obtained from averaging slowness values for time window clusters (Lörrach). Colored curves correspond to SOM clusters in Fig. 5.40. Black symbols show the results for all time windows.

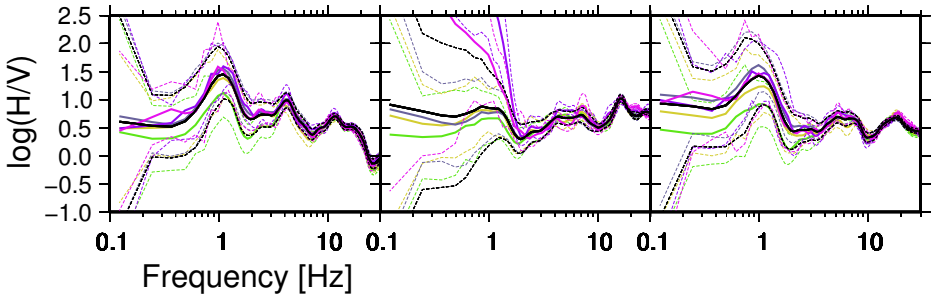

Figure 5.43: H/V spectral ratios and standard deviations (dashed) for SOM clusters (colors from Fig. 5.40) and using all time windows (black). A selection of three receivers is shown for site (Lörrach).

results due to too few time windows and a strong source close to one instrument. Except for the green cluster, the distribution of slowness estimates of all remaining clusters also shows no typical and stable dispersion curves below about 2 Hz. Only the green curve can be interpreted as a realistic mixture of Rayleigh and Love waves.

Except for transients at the fifth station, H/V spectral ratios in Fig. 5.43 show stable peaks for all receivers and clusters. However, the peak amplitude of the green cluster is consistently shifted towards lower values, as already suggested by feature *H/V2*.

Hence, the question of what is happening on the horizontal components arises. Since we observe a stable dispersion curve on the vertical components for all clusters, we can assume that the H/V peak exists due to Rayleigh wave ellipticity. Increased H/V peak amplitudes would then correspond to higher horizontal or, respectively, lower vertical amplitudes at frequencies around 1 Hz due to the changing proportion of Rayleigh waves and other signals. Thus, the lack of those signals or increased contribution of Rayleigh waves would explain the more Rayleigh wave-like dispersion curve and lower H/V amplitude for the green cluster. This is also confirmed by the observed feature properties (*H/V2*, *rect4*, *prr5*). For body waves one would expect lower slowness values. Thus, they cannot be responsible for the shift towards higher values for the remaining clusters. Furthermore, dispersion is not in line with a realistic curve expected for Love waves (decreasing velocities with frequency), and acoustic or electromagnetic signals have different velocities. However,

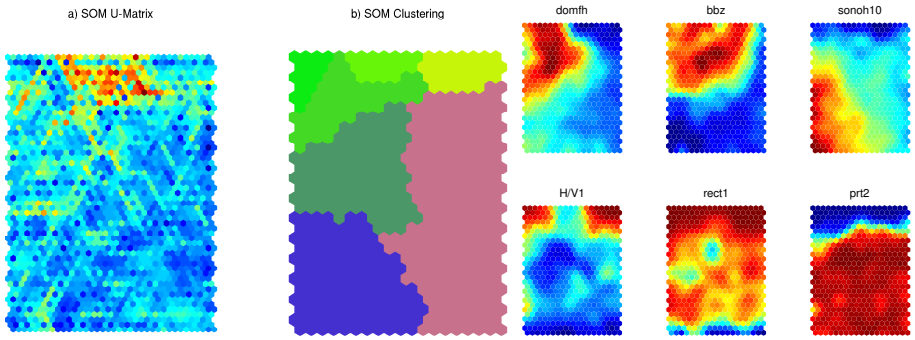

Figure 5.44: SOM visualizations for Hamburg recordings. Component planes for six representative features are shown.

there must be a horizontal signal at frequencies around 1 Hz, otherwise why high linearities and no Rayleigh wave dispersion on the horizontal components are observed cannot be explained. Therefore, we have to assume other horizontal, linear polarized signals which do not fulfill the model assumption for f-k analysis (1D velocity structure, planarity of wavefronts). For instance, consider the effects related to the surface and subsurface topography at this site, which might cause phenomena such as standing waves. Furthermore, a changing, complex source distribution due to the close motorway, industrial activities, and meteorological phenomena may contribute to non-1D wave propagation. Nevertheless, no clear, closing interpretation can be given, except that clustering allows the selection of signals more suitable for dispersion analysis on horizontal components and estimation of ellipticities from H/V amplitudes.

Hamburg

We choose seven SOM clusters for the Hamburg recordings (Fig. 5.44 and 5.45). The interpretation of the results is very similar compared with the Lörrach site. While light green clusters correspond to short transients, probably due to footsteps close to receivers (see, e.g., low coherency *prt2*), the one-minute patterns (blue) show increased amplitudes at all stations. Regarding the features, short transients are characterized by high and linearly polarized amplitudes at low frequencies (0.3 Hz) and a high bandwidth (*rect1*, *H/V1*, *bbz*). Both can be a result of a short impulse which causes tilt and near-field displacement. On the other hand, the blue cluster seems to correspond to close sources but outside the array (higher frequencies on the horizontal components and lower bandwidth compared to the background wavefield).

Considering the length and the occurrence of this pattern, it is very likely that the blue colored time windows contain signals due to the starting or landing of airplanes, either directly generated after touch-down and before lift-off, or indirectly by the ground coupling of infrasonic acoustic emissions and air turbulence. Whatever the sources may be, there is no negative effect on the dispersion curves for all components in Fig. 5.46. As expected, only the green clusters cannot be used to estimate dispersion below 2 Hz. In fact, leaving out those time windows slightly decreases the mean and clearly reduces variance compared to employing all time windows. Since the H/V peak frequency is located at relatively high frequencies around 5 Hz and the disturbing effect of transients is mainly observed at lower

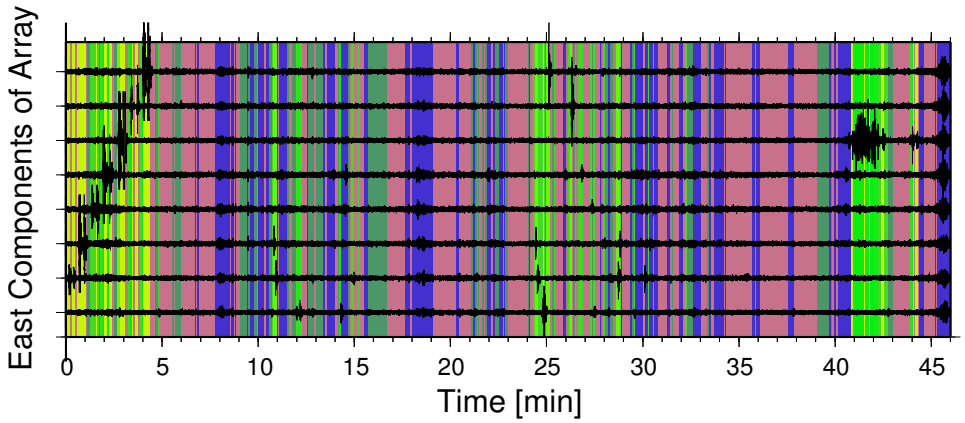

Figure 5.45: Visualization of significant short-term patterns in the Hamburg recordings. Background coloring coresponds to SOM clustering in Fig. 5.44. For all stations the east components are shown. Seismograms are bandpass-filtered between 0.2 and 10 Hz.

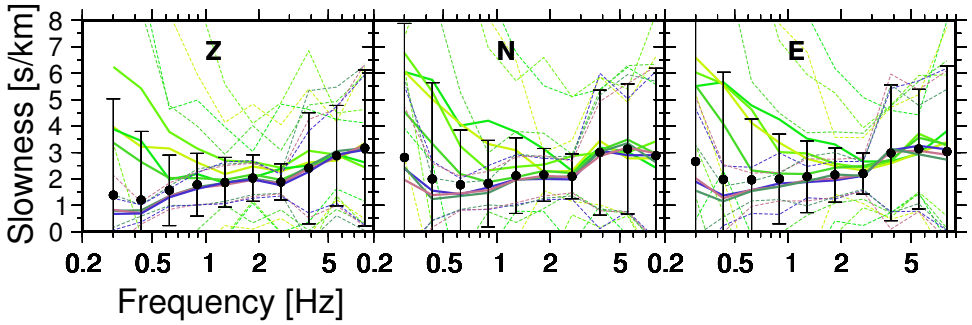

Figure 5.46: Mean dispersion curves and standard deviations (dashed) obtained from averaging slowness values for time window clusters. Colored curves correspond to SOM clusters in Fig. 5.44. Black symbols show the results for all time windows.

frequencies, no significant effect on the H/V spectral ratios can be identified (see Fig. 8.9 in the Appendix).

Lüneburg

Since we are interested in short-term patterns occurring during the night, we use data between 1 and 2 a.m. for the Lüneburg site. Based on the U-matrix (Fig. 5.47), a meaningful clustering ($k = 6$) is chosen for further investigation. Considering the seismograms in Fig. 5.48, we can distinguish between signals of higher amplitudes (violet and light blue clusters, bottom SOM), slightly increased amplitudes (orange and red clusters, top right SOM), and the background wavefield (green). No short transients due to sources inside the array are observed during the night, as expected. The clustering is mainly controlled by the changing frequency content on the vertical and horizontal components (see CPs in Fig. 5.47).

The vertical dispersion curves in Fig. 5.49 show that the choice of time windows has a clear effect on the results between 6 and 10 Hz. Note that we are already close to the

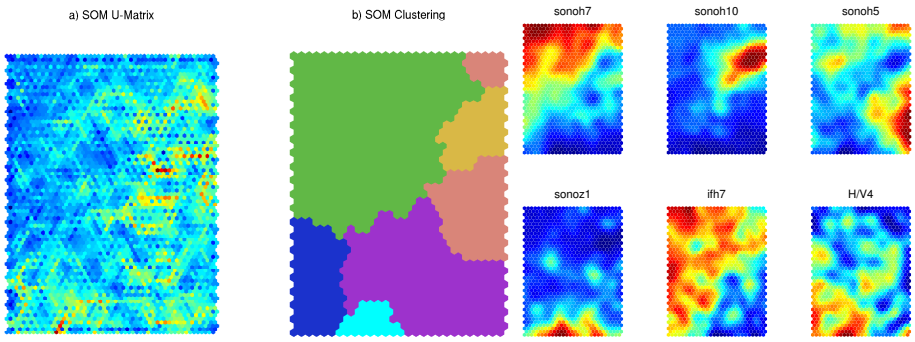

Figure 5.47: SOM visualizations for Lüneburg recordings. Component planes for six representative features are shown.

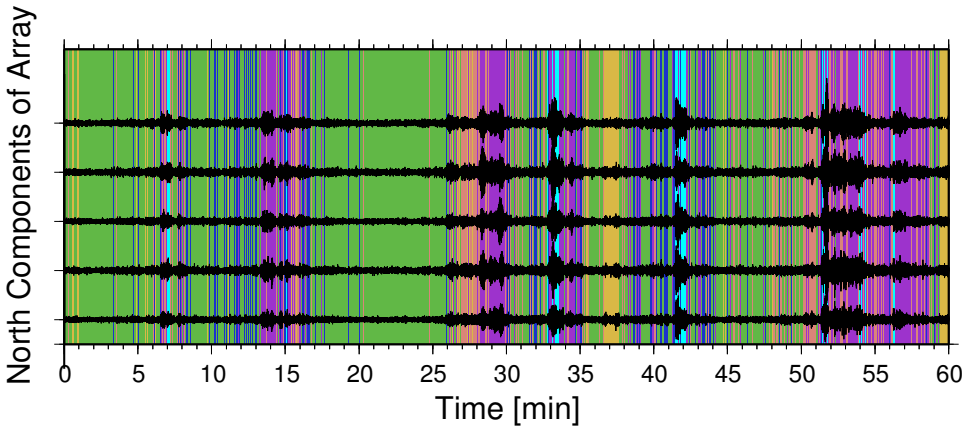

Figure 5.48: Visualization of significant short-term patterns in the Lüneburg recordings. Background coloring coresponds to SOM clustering in Fig. 5.47. For all stations the east components are shown. Seismograms are bandpass-filtered between 4 and 15 Hz.

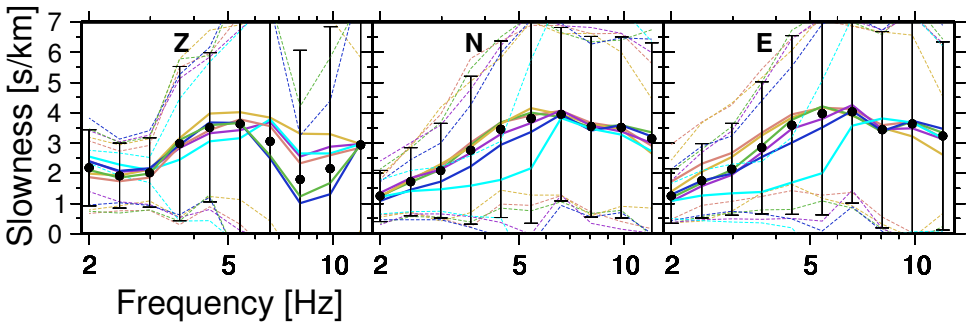

Figure 5.49: Mean dispersion curves and standard deviations (dashed) obtained from averaging slowness values for time window clusters. Colored curves correspond to SOM clusters in Fig. 5.47. Black symbols show the results for all time windows.

upper resolution limit of f-k analysis for those frequencies. Thus, in practice, arrays with smaller aperture are required for the estimation of reasonable slowness values. However, it is possible to interpret the observed patterns qualitatively. Without clustering, a clear trough is observed at 8 Hz, which indicates the lack of surface waves or the dominance of other signals. This phenomenon is less pronounced for the clusters corresponding to higher signal amplitudes, in particular for the yellow category. Hence, those time windows probably contain more surface wave-like signals and, therefore, compensate the wavefield pattern which causes the trough in the dispersion curve. Since the trough is limited to 8 Hz (neighborhood frequency bands are affected due to smoothing), we may assume a continuous, narrow-band signal which is apparently not a planar surface wave. Furthermore, no trough can be observed for the horizontal dispersion curves. Therefore, this signal is dominant on the vertical components only (see also feature *H/V4*). Most likely, a local source is responsible for this phenomenon. A low slowness (high apparent velocity) would explain waves originated or reflected within the subsurface. This signal seems to be strong enough to dominate the normally surface wave-dominated ASV wavefield. Moreover, the responsible local source may be located inside the array. One possible explanation would be a pump located within a well, for which evidence has been found at this site.

5.4.2 Long-Term Patterns

In this section, feature selection, SOM learning and clustering is again conducted for several ASV data sets. However, larger time scales are considered. We are no longer interested in short-term decomposition of the wavefield to select suitable signals for common analysis methods, but aim to investigate the long-term behavior of the wavefield. In particular, since the lengths of our data sets are restricted, mainly daily and hourly variations are considered. Results may be useful for planning measurements, i.e. they allow to assess whether the quality of estimates depends on the time of day. We compute features within the same 10 frequency bands used for the short-term decomposition, but using longer time windows (increasing $WINFAC$). Furthermore, we use different values for Z_{limit} depending on the expected pattern length of one hour and the total number of samples. However, we also test $Z_{limit} = 1.96$ to be sure that we visualize all temporal patterns. All other parameters for feature selection and SOM processing remain unchanged.

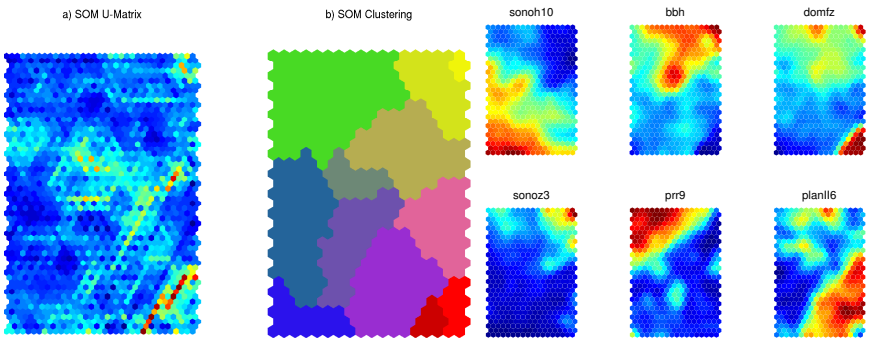

Figure 5.50: SOM visualizations for nine hours of Pulheim recordings. Component planes for six representative features are shown.

Pulheim

For Pulheim, nine hours of recordings between 10 a.m. and 6 p.m. local time are available. We process only 52 minutes of each hour due to missing data. We use a window length of 39.1 seconds ($WINFAC = 40$) and $Z_{limit} = 5$ for feature selection. Fig. 5.50 shows the existence of a clear clustering ($k = 12$), which is defined by changing frequency content, polarization (planarity), and coherency (*prr9*). Fig. 5.51 shows that, aside from close transients during all hours (blue and gray clusters), a gradual change in the background wavefield from bluish (top SOM) to greenish colors (bottom SOM) can be observed. The most likely reason is the increasing human activity (i.e. traffic) during afternoon hours. This would explain the higher bandwidth (*bbh*) and increased coherency at higher frequencies (*prr9*) due to more distinct and close sources.

Furthermore, the U-matrix shows that the two red clusters are more clearly discriminated compared to the other groups. Both clusters are mainly defined by high dominant frequencies (*domfz*). In fact, looking at the raw data in more detail, reveals the existence of a continuous, harmonic and monochromatic signal at 25 Hz on several stations (not visible on the receiver in Fig. 5.51). The Appendix Fig. 8.8 shows this signal for one station using different zoom factors. Since there is an amplitude decay and time delay between receivers (of about 0.2 seconds), this signal is rather seismic than electromagnetic. Furthermore, an instrument failure is also not very likely since data contains the normal seismic wavefield after lowpass filtering. Therefore, we assume a close, local source, e.g. an engine, which generates monochromatic, seismic energy by ground coupling. However, this is of no significance for subsurface investigations since we are outside the frequency band suitable for array methods and clearly above the H/V peak frequency. Remember that the feature *domfz* also takes into account the frequency spectrum outside this band. Therefore, we are able to identify this pattern. Furthermore, we compute dispersion curves (f-k and 3c-MSPAC) and H/V spectra for all nine hours (not shown). As for the short-term patterns, the variation in the ambient vibration wavefield does not affect the results.

Lüneburg

Fig. 5.52 and 5.53 show the results for 21 hours of data ($WINFAC = 100$, 22.6 s window length, $Z_{test} = 5$, $k = 9$). We observe an expected behavior of ASV wavefields

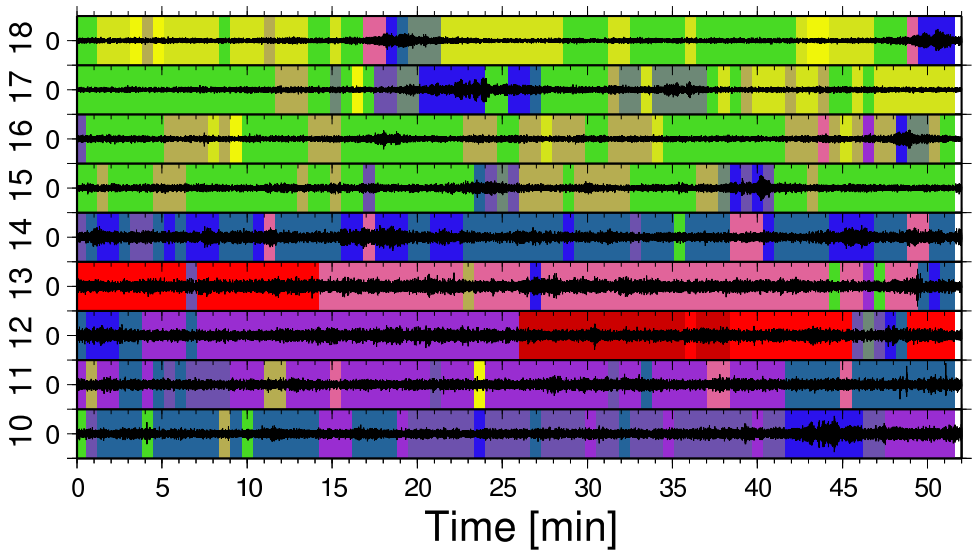

Figure 5.51: Visualization of significant long-term patterns in 9 hours of Pulheim recordings. Y-axis labels indicate hour of day. Background coloring coresponds to SOM clustering in Fig. 5.50. The vertical component is shown for one station.

in urban areas. During the night, the amplitudes of the background wavefield decrease (violet clusters). Furthermore, distinct transients (blue clusters) exist at all times with a changing frequency of occurrence. As for the previously analyzed data for the same site (Section 5.4.1, 1 a.m. to 2 a.m.), again the pattern causing the trough in the dispersion curves around 8 Hz can be identified at night (see *rect8*, *sonoz8* and right panels of Fig. 5.53). In the daytime, this effect is compensated. In fact, the amplitude spectra in Fig. 5.54 show that man-made noise clearly decreases at night for frequencies higher than 1 Hz, and that a peak at 8 Hz becomes visible. Thus, the signal does not disappear by day, but is rather dominated by other signals.

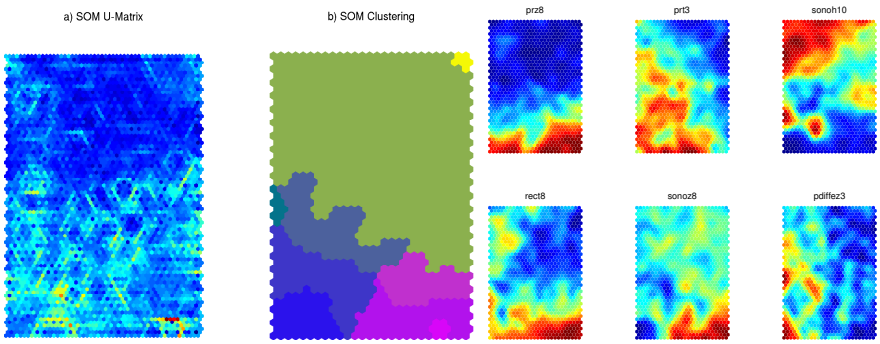

Figure 5.52: SOM visualizations for 21 hours of Lüneburg recordings. Component planes for six representative features are shown.

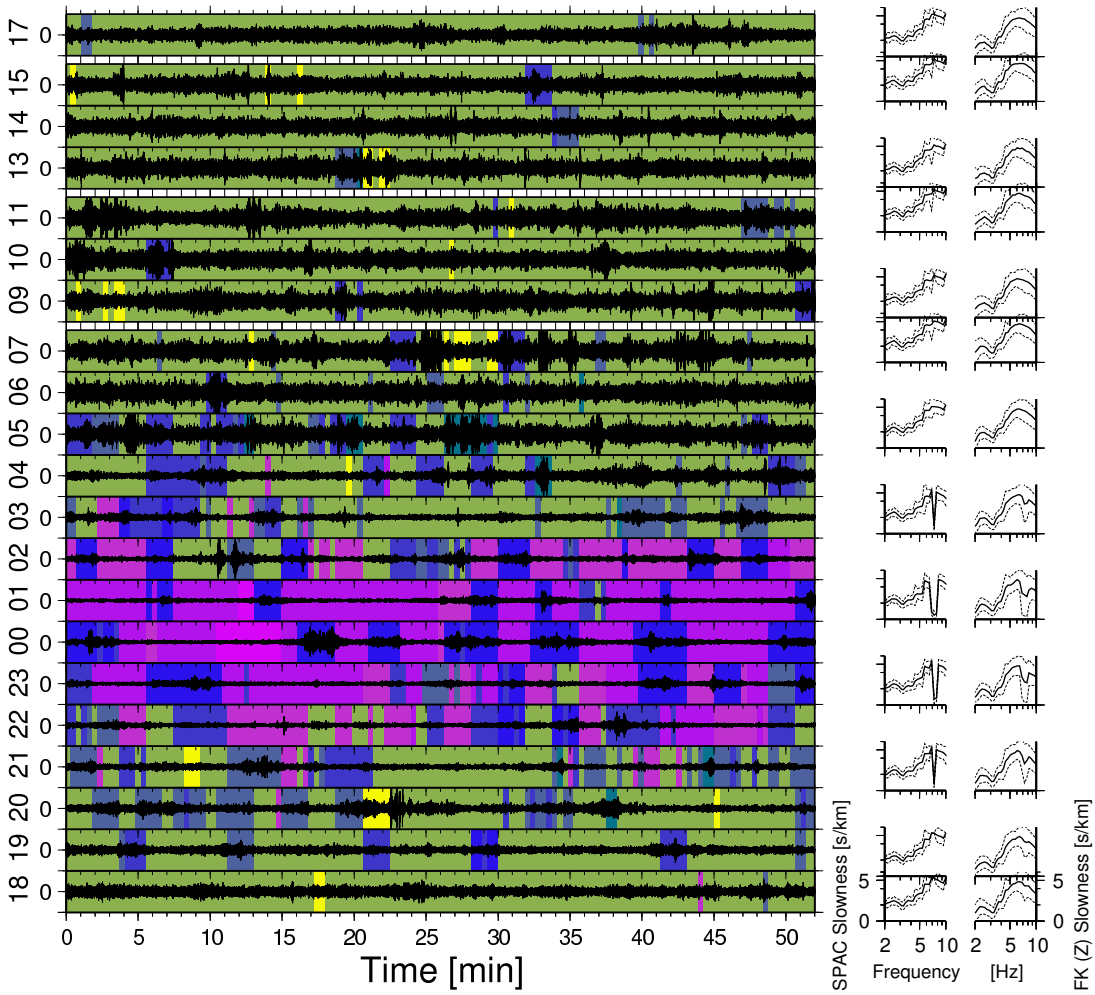

Figure 5.53: Visualization of significant long-term patterns in 21 hours of Lüneburg recordings. Y-axis labels indicate hour of day. Background coloring coresponds to SOM clustering in Fig. 5.52. For one station the vertical component is shown. The right panel shows Rayleigh wave dispersion curves obtained from 3c-MSPAC and f-k, averaged over the corresponding hour.

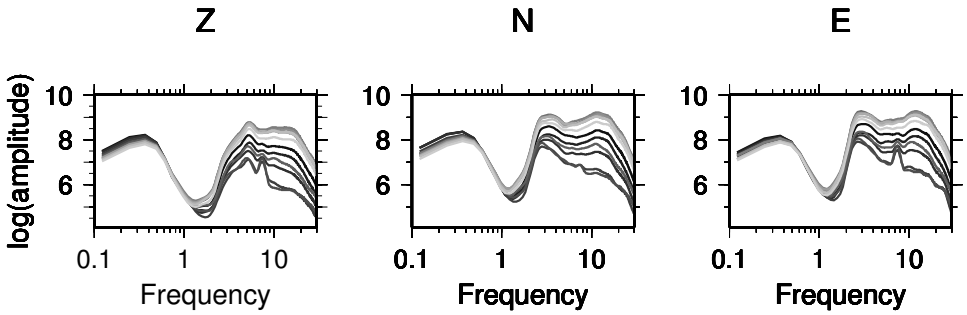

Figure 5.54: Spectra of all three components for each second hour of the Lüneburg recordings. Gray scale goes from black (night) to white (day).

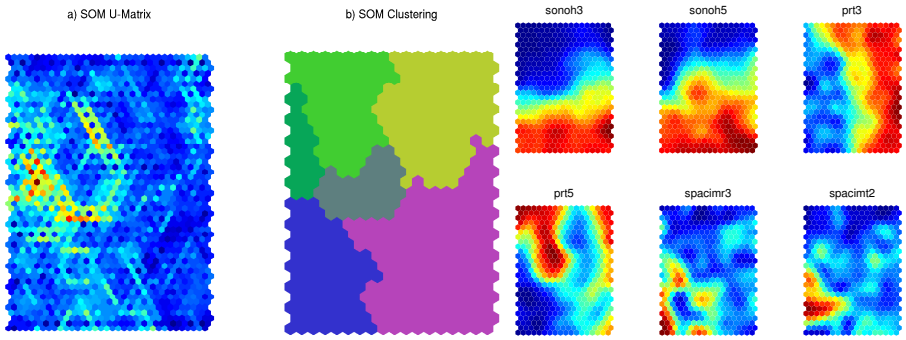

Figure 5.55: SOM visualizations for 17 hours of Colfiorito recordings using a reduced set of potential features (see text). Component planes for six representative features are shown.

Colfiorito

Using a window length of 97.8 seconds ($WINFAC = 100$), we process 17 hours of data for the Colfiorito site. The first three hours (13 - 15) have been recorded a day before the remaining hours (see Table 5.27). Therefore, we have a gap of 24 hours.

Since we observe a striking pattern in the 3c-MSPAC Love dispersion curves at low frequencies during nighttime (no effect on Rayleigh wave dispersion), we reduce the set of potential features to the ones responsible for that variation in a second processing run for optimal visualization. In particular, we choose features from Methods 1, 2, and 5 within the six lowermost frequency bands and apply the feature selection algorithm ($Z_{test} = 1.96$). Results in Fig. 5.55 and 5.56 ($k = 6$) show again the daily cycle due to human activity. The violet cluster (night) is mainly defined by the increased energy contribution at lower frequencies or decreased power at high frequencies compared to the yellow and green clusters (daytime). Moreover, between 9 p.m. and 1 a.m. (blue cluster), the estimated Love wave dispersion curves become unstable (high slowness). Low coherency and high imaginary SPAC coefficients on horizontal components indicate that the assumption of planar surface waves is not fulfilled anymore. An explanation could be the lack of energy at night. However, since the transition between low and high SPAC coefficients is not continuous, and because there is not such a clear indication for this pattern in the sonogram (*sono*), this phenomenon may also be related to a local source.

Furthermore, without considering the hours of unstable results, we observe that the fraction of Rayleigh waves in Fig. 5.56 (Alpha) is slightly higher at night.

5.4.3 Discussion

The results of all analyzed data sets have shown the existence of different types of short-term patterns. In most instances they can be related to time windows with increased amplitudes (transients). Furthermore, patterns in the background wavefield are observed which are not visible from the raw seismograms (see Lörrach). However, the relevance of those patterns for estimation of dispersion curves and H/V spectral ratios has to be evaluated as the case arises. The majority of found clusters for all data sets have no significant impact on vertical dispersion curves. Therefore, the existence of more or less Rayleigh wave-like time windows cannot be shown. Even a comparably small number of

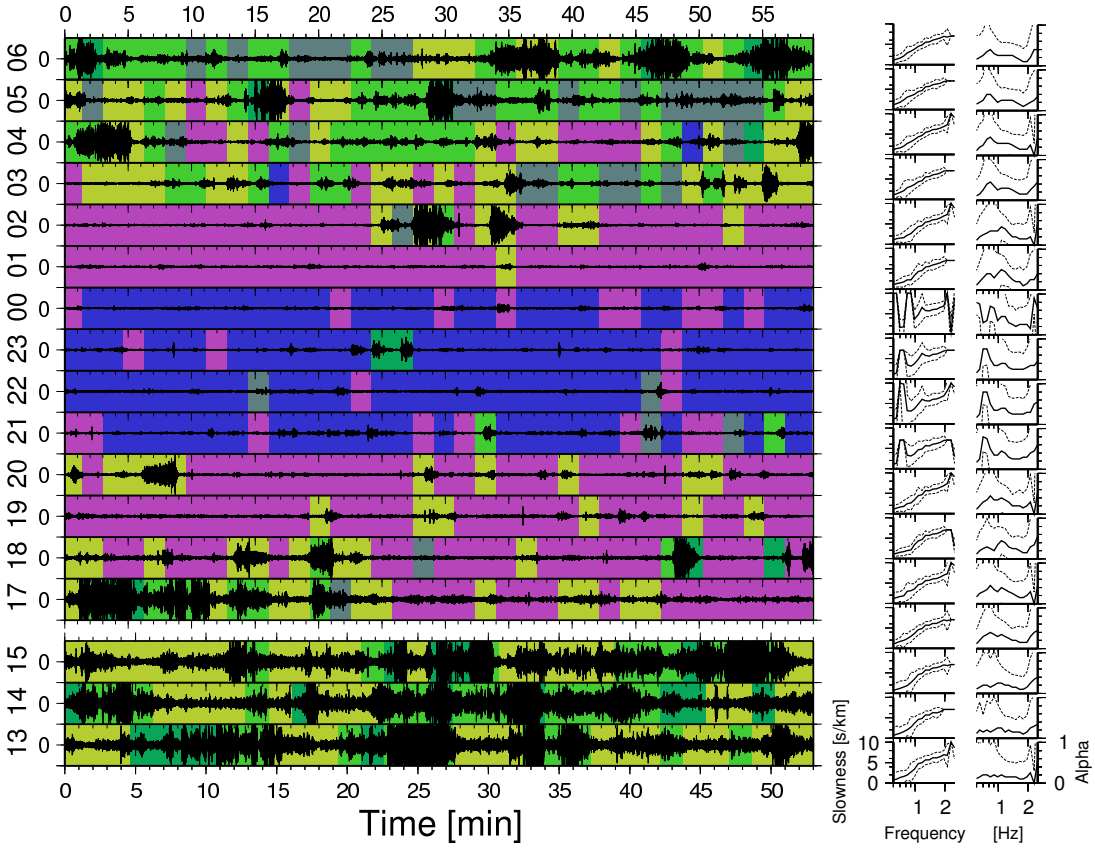

Figure 5.56: Visualization of significant long-term patterns in 17 hours of Colfiorito recordings. Y-axis labels indicate hour of day. Background coloring coresponds to SOM clustering in Fig. 5.55. For one station the vertical component is shown. The right panel shows the 3c-MSPAC Love wave dispersion curves and fraction of Rayleigh waves (Alpha) averaged over each hour.

windows yields similar good results compared to using the complete recordings.

Furthermore, a short-term decomposition into pure Rayleigh and Love wave time windows cannot be achieved by clustering. Although we observe effects of patterns on the horizontal dispersion characteristics, no clear evidence for two different dispersion branches can be observed. Here, application of the 3c-MSPAC method to the complete record is a more suitable alternative for estimating Love wave dispersion curves. In summary, except for special cases, which we will discuss in the following, all time window clusters contain a mixture or superposition of surface waves and not pure Rayleigh, Love or body waves.

However, some important insights into the nature of the ambient vibrations wavefield and practical aspects for data processing can be gained by short-term clustering. A decomposition can be helpful as it allows for discriminating between more and less suitable signals. Within this context, the impact of local sources can become significant, especially when all three wavefield components are analyzed (3c-f-k and H/V). In fact, dispersion curve and H/V estimates can be improved as well as impaired by local sources.

In particular, it could be shown that signals which exhibit increased amplitudes compared to the background wavefield can improve results. If generated outside the receiver

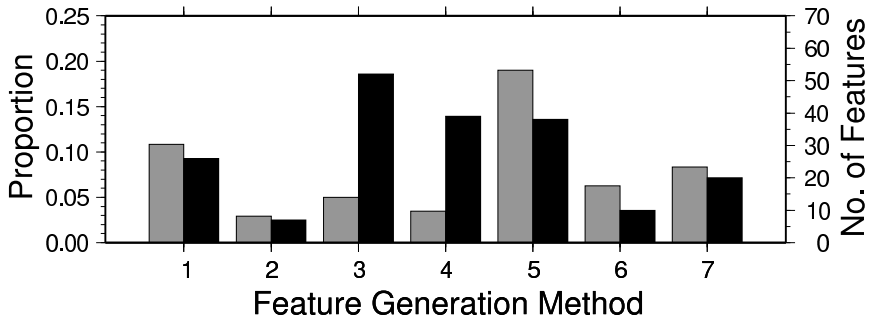

Figure 5.57: Cumulative feature statistic for all processed ambient vibration data sets (4 sites for short-term and 3 for long-term patterns). Black bars indicate the absolute number of features selected from each feature generation method. Gray bars show the relative proportion of selected features with respect to the total number of each method.

array, those signals show clear surface wave characteristics. Therefore, they could even be used alone for subsurface investigations. For instance, transients could be employed to increase resolution of H/V amplitudes at higher frequencies, since similarities have been observed compared with earthquake H/V ratios (see Pulheim). Furthermore, if there are signals in the background wavefield showing no surface wave properties, the transients may be the only valid signals (see Lüneburg).

Contrary observations are made for transients generated by very close sources inside the array. Leaving out those time windows has the potential to decrease the variances for slowness and H/V amplitude estimates (see Hamburg). One could argue that those signals could very easily be rejected using a simple STA/LTA anti-trigger. However, this would also reject the previously mentioned transients originated outside the array. Hence, other, amplitude-independent wavefield properties, mainly the changing spectral content in different frequency bands and on spatial components, have to be considered.

Parolai & Galiana-Merino (2006) did not distinguish between both types of transients since H/V at a single station have been considered. However, similar to our results, using only transients has shown to deliver good results for the estimation of H/V peak frequency. Furthermore, the findings of Roberts & Asten (2008) that wave front curvature from close sources outside the array is minimal, fit with our observations.

In contrast to the vertical component, non-transient wavefield patterns can affect horizontal dispersion. The Basel data set shows the existence of more suitable signals, which lead to more realistic dispersion curves. However, the more reasonable curve is probably still affected by waves which do not fit with the assumptions for f-k analysis (planar surface waves). Hence, our short-term decomposition is a rather more qualitative analysis tool than an automatic preprocessing method for f-k or other techniques.

The long-term patterns are mainly controlled by variation between day and night due to human activity. For both analyzed 24h-data sets, we observe an impact on estimated dispersion curves. However, these effects are limited to particular frequencies close to the resolution limits of the array methods. For both data sets the cause of those effects, i.e. the trough in Rayleigh wave dispersion curve for Lüneburg and unstable Love wave dispersion for Colfiorito, can be related to the lack of energy at night. Therefore, other sources may become dominant producing signals unsuitable for array analysis (e.g. non-planar surface

waves, body waves). On the other hand, there is no preferable time for ambient vibration measurements in the daytime, since the observed temporal patterns have no significant effect on dispersion curve estimates.

Fig. 5.57 presents the cumulative feature statistic for all processed ambient vibration data sets. In particular, which feature generation method a feature belongs to is evaluated. Followed by Method 1 (coherency from f-k) and Method 7 (amplitude ratios), Method 5 (frequency spectrum) has the highest ratio between selected and total number of features (gray bars). Hence, those features have the highest relevance (with respect to Z_{test}). Furthermore, there is less redundancy in the original feature sets. Otherwise, fewer features from those generation methods would "survive". The absolute number of features (black bars) shows that Methods 3 (polarization) and 4 (complex trace) have a significant contribution to the final feature sets. In spite of using two-level correlation hunting (see Section 4.7), this may also be due to the fact that the largest, individual feature sets have been generated for these methods (see Table 4.1). However, since the relative frequency of occurrence is low, selection is limited to specific features (e.g. instantaneous frequency). Methods 2 (imaginary SPAC) and 6 (polarization spectrum) are less important.

Chapter 6

Conclusions

The goal of this work has been the development of an unsupervised pattern discovery approach for different types of 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 2D visualization method, we have successfully employed the multi-purpose and easily applicable Self-Organizing Map (SOM) method. Furthermore, clustering based on a trained SOM, in order to group existing patterns, has been performed. Using newly generated SOM features such as the component planes (CPs) or the trajectory of data hits on the SOM, we have also trained and employed higher-level representations, i.e. SOMs of SOMs (CP-SOM, TSOM).

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. In particular, features are ranked using the runs test statistic and grouped by means of clustering the CP-SOM. Using the automatically selected feature set, training and clustering of the SOM is performed as a last step of our learning approach. While the temporal context has been considered during feature selection by the runs test, a simple context-independent clustering using hierarchical techniques has been used. Based on different SOM visualizations such as CPs and similarity colorings, interpretation of temporal patterns and obtained signal grouping has been carried out.

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., earthquakes. 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 time windows.

We have applied our processing scheme to different types of seismic data sets. As a first step, we extensively tested performance and algorithm parameters by computing classification rates using synthetic and real-world data for which theoretical or manually defined labels have been available. Regarding the time window length for feature generation, it was found that for the discovery of short-term patterns (transients) as well as long-term variations in the background wavefield, windows including at least four to five cycles of the

signal of interest produce stable results. We showed the reliability of the Wald-Wolfowitz runs test for ranking features according to their relevancy to show significant temporal patterns. Furthermore, it was found that the average linkage rule for hierarchical clustering and working with large SOMs produces the most favorable and meaningful groupings for SOM-CPs as well as for the seismogram time windows. We confirm the need of data normalization for SOM learning. Furthermore, results of the tests have shown similar or slightly improved classification rates using our filter feature selection method compared to the complete set of all available features and feature selection by means of Principal Component Analysis (PCA). Moreover, in contrast to employing PCA as feature extraction method, results remain easily interpretable since no new features are generated. Besides its high computational cost, a wrapper method for feature selection has shown not to be suitable even for a simple toy data set. Our feature selection method has produced a lower model complexity over a decreased feature set whose size 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 compared to using only the common and most suitable seismological approach, which we found has been the frequency spectrum. Furthermore, beside spectral features, results for all data sets have particularly shown the suitability of coherency from Frequency Wavenumber Analysis, polarization attributes such as linearity from covariance matrix, and amplitude ratios of wavefield components.

The final SOM itself facilitated intuitive mapping of trends and imaging of patterns in seismic wavefields for different types of recordings. We found that clustering can be carried out as a first order grouping even when transitions between signals are continuous. This is typical for seismic data where often no clear "end" of a signal can be defined. Furthermore, we found that the hierarchical structure of clusterings allows a meaningful interpretation of results. The combination of quantitative (cluster validity criteria) and qualitative considerations (SOM visualization) allowed us to define the best clustering, i.e. to find the most meaningful number of clusters. Particularly, we confirm that the Davies-Bouldin and Cluster Stability criterion are well suitable for the quantitative evaluation. While the first has shown to be more robust for automatic approaches, computing the latter is more time consuming. Furthermore, we found that Cluster Stability has the tendency to prefer a solution of only two clusters. However, we approved the ability of Cluster Stability to avoid overfitting. Furthermore, it was found that a meaningful number of clusters does not necessarily have to fit with the expected number of signals classes.

For the real data, we have shown that the natural seismic phase and event discrimination can be intuitively visualized using SOMs. Furthermore, simple SOM clustering allowed the detection of events. In particular, we have applied our techniques to a data set of regional earthquakes using preselected time windows including the background wavefield before onset. Projection of the manually picked seismic phase labels on the SOM allowed the visualization of phase discrimination for single events. Good classification rates or discrimination have been found for P wave time windows for arbitrary earthquakes. On the other hand, it has not been possible to define distinct P and S wave classes for all events recorded on two different receivers. Therefore, it turned out to be a challenge to design simple phase classifiers for station networks. However, improvements are expected for supervised classification by including the context-dependency of feature vectors. Furthermore, we have analyzed recordings of seismicity at volcano Mt. Merapi (Indonesia).

In summary, we were able to identify or detect two characteristic high-polarized volcano-seismic signals which are crucial for eruption forecasting. We found that they are assigned to distinct clusters over the entire day independent of the event amplitudes what has not been possible for a third type (MP events). Furthermore, the two event classes (VTB and Guguran) could be distinguished. Moreover, we have investigated the long-term wavefield behavior based on SOM colorings, which have been obtained from similarities of cluster and SOM prototype vectors. These visualizations allowed intuitive recognition and confirmation of the clear 24-hour cycle on the SOM which is known to be related to the human activity.

Finally, in order to test wavefield decomposition, we applied SOM clustering to ambient seismic vibration data sets. Clusters have been found for different types of transients and for patterns in the background wavefield. Clusters of transients originated inside and outside the receiver array have been distinguished. Long-term patterns have been recognized for recordings lasting more than 24 hours which could be attributed to local sources and day-night variations due to anthropogenic activity. We have found that most long- and short-term patterns in ambient vibrations have no big impact on standard analysis methods, i.e. Rayleigh wave dispersion curves and H/V spectral ratios. Nevertheless, negative effects such as increased variances and unstable dispersion curve segments have been observed at particular sites due to day-night variations and transients generated inside the receiver array. No decomposition into pure Rayleigh and Love waves has been possible. However, since we found clusters of time windows less suitable for analysis, there is the potential to decrease the variance of estimates or improve results in general. For instance, this can be achieved by including expert knowledge about how a realistic dispersion curves should look like (e.g. no troughs).

Considering the effort, useability, and applicability of techniques, we can conclude that feature selection, SOM learning, and clustering has been working rapidly and robustly for all analyzed data sets. Nevertheless, high computation cost has been spent on feature generation, most on f-k analysis. For instance, computing all features for one hour of the Merapi array recordings (three stations, 40 Hz sampling rate) takes about 37 minutes on a Intel® Pentium® 4 1.5 GHz machine with 1024 MB RAM. Computing only features that are known to have been suitable for similar, previously processed data sets or restricting feature generation and selection to parts of the data can decrease computing time, e.g. when real-time processing is required. Almost all algorithm parameters for feature selection and SOM learning need not to be adjusted for new data. Nevertheless, we found that an important parameter for feature selection is the runs test statistic Z_{limit}, which was used as a threshold. It can be tuned empirically using a desired significance level or the shortest expected duration of a pattern.

Our approach can be easily applied to other problems in seismology, geophysics, and beyond. Only the first step, feature generation, has to be adapted to the given time series data. In doing so, our unsupervised feature selection procedure and SOM learning could be applied, e.g., to magnetic or meteorological data sets. For seismological recordings, we suggest particular modifications and applications. For instance, as a semi-supervised learning approach one could use a simple STA/LTA trigger to find transients and subsequently do clustering for a more distinct signal grouping. Another possible application is data quality control. SOM visualization and clustering can be used to find time windows characterized by tilted instrument or clipped signals for instance. Moreover, the occurence

of communication problems between seismometer, digitizer or data storage system may be recognized. In order to improve data grouping in general, ensemble clustering might be used since we found that there is no single best solution for complex data sets such as seismograms. Furthermore, features could be computed for specific problems. For instance, in the context of ambient vibration analysis, new features can be generated which account for increasing slowness with frequency in order to allow the recognition of time windows with realistic dispersion properties.

Chapter 7

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

Arai, H. & Tokimatsu, K., 2004, S-Wave Velocity Profiling by Inversion of Microtremor H/V Spectrum, *Bulletin of the Seismological Society of America*, **94**(1), 53–63

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

Asten, M. & Henstridge, J., 1984, Array estimators and the use of microseisms for reconnaissance of sedimentary basins, *Geophysics*, **49**(11), 1828–1837

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

Bard, P., 1999, Microtremor measurements: a tool for site effect estimation, *The Effects of Surface Geology on Seismic Motion*, **3**, 1251–1279

Bard, P., 2004, The SESAME project: an overview and main results, *13th World Conference on Earthquake Engineering, Vancouver, August 2004*, **paper No. 2207**

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.A. Pérez-Uribe, A., 2007, Improving the Correlation Hunting in a Large Quantity of SOM Component Planes, *Lecture Notes in Computer Science: Artificial Neural-Networks–ICANN 2007*, **4669**, 379–288

Basak, J., De, R. & Pal, S., 1998, Unsupervised feature selection using a neuro-fuzzy approach, *Pattern Recognition Letters*, **19**(11), 997–1006

Bellman, R., 1961. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ

Bettig, B., Bard, P., Scherbaum, F., Riepl, J., Cotton, F., Cornou, C. & Hatzfeld, D., 2001, Analysis of dense array noise measurements using the modified spatial auto-correlation method (SPAC). Application to the Grenoble area, *Bollettino di Geofisica Teorica ed Applicata*, **42**(3-4), 281–304

Bishop, C., 2006. *Pattern recognition and machine learning*. Springer

Bonnefoy-Claudet, S., Cotton, F. & Bard, P., 2006, The nature of noise wavefield and its applications for site effects studies A literature review, *Earth Science Reviews*, **79**(3-4), 205–227

Brenguier, F., Shapiro, N., Campillo, M., Ferrazzini, V., Duputel, Z., Coutant, O. & Nercessian, A., 2008, Towards forecasting volcanic eruptions using seismic noise, *Nature Geoscience*, **1**(2), 126

Cho, I., Tada, T. & Shinozaki, Y., 2006, A generic formulation for microtremor exploration methods using three-component records from a circular array, *Geophysical Journal International*, **165**(1), 236–258

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

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

Dash, M. & Liu, H., 1997, Feature selection for classification, *Intelligent Data Analysis*, **1**(3), 131–156

Davies, D. & Bouldin, D., 1979, A cluster separation measure, *IEEE 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

Dempster, A., Laird, N., Rubin, D. et al., 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, **39**(1), 1–38

Dowla, F., Taylor, S. & Anderson, R., 1990, Seismic discrimination with artificial neural networks: Preliminary results with regional spectral data, *Bulletin of the Seismological Society of America*, **80**(5), 1346–1373

Duvall, T., Scherrer, P., Bogart, R., Bush, R., De forest, C., Hoeksema, J., Schou, J., Saba, J., Tarbell, T., Title, A. et al., 1997, Time-distance helioseismology with the MDI instrument: Initial results, *Solar Physics*, **170**(1), 63–73

Dy, J. & Brodley, C., 2004, Feature Selection for Unsupervised Learning, *The Journal of Machine Learning Research*, **5**, 845–889

Ertöz, L., Steinbach, M. & Kumar, V., 2003, Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach, *Clustering and Information Retrieval,* 83–104

Esposito, A., Giudicepietro, F., D'Auria, L., Scarpetta, S., Martini, M., Coltelli, 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

Ester, M., Kriegel, H., Sander, J. & Xu, X., 1996, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press,* 226–231

Fäh, D., Kind, F. & Giardini, D., 2001, A theoretical investigation of average H/V ratios, *Geophysical Journal International*, **145**(2), 535–549

Fogelman-Soulié, F., 2008. Industrializing Data Mining, Challenges and Perspectives. In W. Daelemans, B. Goethals & K. Morik editors, *ECML/PKDD (1)*, volume 5211 of *Lecture Notes in Computer Science* 1. Springer

Fukunga, K., 1990. *Statistical Pattern Recognition.* Academic Press

Guérif, S., 2008, Unsupervised Variable Selection: when random rankings sound as irrelevancy, *Journal of Machine Learning Research Workshop and Conference Proceedings: New challenges for feature selection in data mining and knowledge discovery*, **4**, 163–177

Guérif, S., Bennani, Y., France, V., Janvier, E., France, N. & France, B., 2005, μ-SOM: Weighting features during clustering, *Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM 05),* 397–404

Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2002, Cluster validity methods: Part I and II, *SIGMOD 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

Herrmann, R. B., 2001, Computer programs in seismology, Version 3.1, *St. Louis University*

Horike, M., 1985, Inversion of phase velocity of long-period microtremors to the S-wave-velocity structure down to the basement in urbanized areas, *Journal of Physics of the Earth*, **33**(2), 59–96

Jain, A., 2008. Data Clustering: 50 Years Beyond K-means. In *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases-Part I,* 3–4

Jain, A., Murty, M. & Flynn, P., 1999, Data Clustering: A Review, *ACM Computing Surveys*, **31**(3), 264–323

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

Kleinberg, J., 2002, An impossibiilty theorem for clustering, *Proceeding of Advances in Neural Information Processing Systems*, **15**

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

Köhler, A., Ohrnberger, M. & Scherbaum, F., 2006. The relative fraction of Rayleigh and Love waves in ambient vibration wavefields at different European sites. In *Proceedings of the third International Symposium on the Effects of Surface Geology on Seismic Motion, Grenoble, France, Paper Number 83*

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*. Springer

Kuehn, D., Ohrnberger, M., Vollmer, D., Dahm, T., Scherbaum, F. & Dehghani, A., 2006. Ambient Vibration Array Measurements Using a Wireless Array Analysis System. In *American Geophysical Union, Fall Meeting 2006, abstract S23A-0140*

Kvaerna, T. & Ringdahl, F., 1986, Stability of various FK estimation techniques, *Semi-anual technical summary, 1 October 1985 - 31 March 1986, NORSAR Scientific Report, Kjeller, Norway*, **1-86/87**, 29–40

Lacoss, R., Kelly, E. & Toksöz, M., 1969, Estimation of seismic noise structure using arrays, *Geophysics*, **34**(1), 21–38

Lange, T., Roth, V., Braun, M. & Buhmann, J., 2004, Stability-Based Validation of Clustering Solutions, *Neural Computation*, **16**(6), 1299–1323

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

Malischewsky, P. & Scherbaum, F., 2004, Loves formula and H/V-ratio (ellipticity) of Rayleigh waves, *Wave Motion*, **40**(1), 57–67

Maurer, W., Dowla, F. & Jarpe, S., 1992, Seismic event interpretation using self-organizing neural networks, *Proceedings of the SPIE - The International Society for Optical Engineering*, **1709**, 950–958

McQueen, J., 1967, Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* 281–297

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, *Bulletin of the Seismological Society of America*, **86**(2), 320–328

Minakami, T., 1960, Fundamental research for predicting volcanic eruptions (Part 1). Earthquakes and crustal deformations originating from volcanic activities, *Bulletin of the Earthquake Research Institute*, **38**, 497–544

Mitra, P., Murthy, C. & Pal, S., 2002, Unsupervised Feature Selection Using Feature Similarity, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(3), 301–312

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

Mucciarelli, M., Gallipoli, M. & Arcieri, M., 2003, The Stability of the Horizontal-to-Vertical Spectral Ratio of Triggered Noise and Earthquake Recordings, *Bulletin of the Seismological Society of America*, **93**(3), 1407–1412

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

Nakamura, Y., 1989, A Method for Dynamic Characteristics Estimation of Subsurface using Microtremor Ground Surface, *Quarterly Report of Railway Technical Research Institute (RTRI)*, **30**(1), 25–33

Nattkemper, T., Twellmann, T., Ritter, H. & Schubert, W., 2003, Human vs. machine: evaluation of fluorescence micrographs, *Computers in Biology and Medicine*, **33**(1), 31–43

Ohmachi, T. & Umezono, T., 1998, Rate of Rayleigh waves in microtremors, *Proceeding of the Second International Symposium on he Effects of Surface Geology on Seismic Motion,* 587–592

Ohrnberger, M., 2001. *Continuous Automatic Classification of Seismic Signals of Volcanic Origin at Mt. Merapi, Java, Indonesia.* Dissertation, University of Potsdam

Ohrnberger, M. & the SESAME Team WP05/WP06, 2004, Continuous array processing toolkit for ambient vibration array analysis, *Deliverable D18.06*

Ohrnberger, M., Scherbaum, F., Krüger, F., Pelzing, R. & Reamer, S., 2004, How good are shear wave velocity models obtained from inversion of ambient vibrations in the Lower Rhine Embayment (NW Germany)?, *Bollettino di Geofisica Teorica ed Applicata*, **45**(3), 215–232

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

Parolai, S. & Galiana-Merino, J., 2006, Effect of Transient Seismic Noise on Estimates of h/v Spectral Ratios, *Bulletin of the Seismological Society of America*, **96**(1), 228–236

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, *Geophysical Research Letters*, **32**(1), L01303 doi:10.1029/2004GL021115

Parsons, L., Haque, E. & Liu, H., 2004, Subspace clustering for high dimensional data: a review, *ACM SIGKDD Explorations Newsletter*, **6**(1), 90–105

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., JL, F., 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 in Computer Science*, **4702**, 565–572

Roberts, J. & Asten, M., 2008, A study of near source effects in array-based (SPAC) microtremor surveys, *Geophysical Journal International*, **174**(1), 159–177

Roberts, R., Christoffersson, A. & Cassidy, F., 1989, Real-Time Event Detection, Phase Identification and Source Location Estimation Using Single Station Three-Component Seismic Data, *Geophysical Journal International*, **97**(3), 471–480

Roux, P. & Kuperman, W., 2004, Extracting coherent wave fronts from acoustic ambient noise in the ocean, *The Journal of the Acoustical Society of America*, **116**, 1995

Roux, P., Sabra, K., Kuperman, W. & Roux, A., 2005, Ambient noise cross correlation in free space: Theoretical approach, *The Journal of the Acoustical Society of America*, **117**, 79

Roweis, S. & Saul, L., 2000, Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, **290**(5500), 2323–2326

Saeys, Y., Abeel, T. & Van de Peer, Y., 2008, Robust Feature Selection Using Ensemble Feature Selection Techniques, *Lecture Notes in Computer Science*, **5212**, 313–325

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

Scales, J. & Snieder, R., 1998, What is noise, *Geophysics*, **63**(4), 1122–1124

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

Schimmel, M. & Gallart, J., 2003, The use of instantaneous polarization attributes for seismic signal detection and image enhancement, *Geophysical Journal International*, **155**(2), 653–668

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

Schoelkopf, B., Smola, A. & Mueller, K., 1997, Kernel Principal Component Analysis, *Lecture Notes in Computer Science*, **1327**, 583–588

Shapiro, N., Campillo, M., Stehly, L. & Ritzwoller, M., 2005, High-Resolution Surface-Wave Tomography from Ambient Seismic Noise, *Science*, **307**(5715), 1615–1618

Shneiderman, B., 2002, Inventing discovery tools: combining information visualization with data mining, *Information Visualization*, **1**(1), 5–12

Strehl, A., 2002. *Relationship-based clustering and cluster ensembles for high-dimensional data mining.* Dissertation, PhD thesis, The University of Texas at Austin

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

Tenenbaum, J., Silva, V. & Langford, J., 2000, A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, **290**(5500), 2319–2323

Theodoridis, S. & Koutroumbas, K., 1998. *Pattern Recognition.* Academic Press, San Diego

Tokimatsu, K., 1997, Geotechnical site characterization using surface waves, *Proc. 1st Intl. Conf. Earthquake Geotechnical Engineering*, **3**, 1333–1368

Toksöz, M. & Lacoss, R., 1968, Microseisms: mode structure and sources, *Science*, **159**, 872–873

Tryon, R., 1939. *Cluster Analysis; Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality.* Edwards brother, inc., lithoprinters and publishers

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

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

Vesanto, J. & Sulkava, M., 2002, Distance Matrix Based Clustering of the Self-Organizing Map, *Proc. International Conference on Artificial Neural Networks–ICANN 2002,* 951–956

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

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

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

Wessel, P. & Smith, W., 1995, The Generic Mapping Tools (GMT) version 3.0, *Technical Reference & Coockbook, SOEST/NOAA. University of Hawaii, Mauoa*

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

Zhang, T., Ramakrishnan, R. & Livny, M., 1996, BIRCH: An efficient data clustering method for very large databases, *International Conference on Management of Data: Proceedings of the 1996 ACM SIGMOD international conference on Management of data: Montreal, Quebec, Canada*, **4**(6), 103–114

Chapter 8

Appendices

8.1 Implementation of Algorithms and Used Software

The implementation of our unsupervised pattern recognition approach for seismic wavefield recordings is divided into different modules, which are all started and controlled from the Unix Shell (Fig. 8.1). Feature generation uses the *Continuous Array Processing Toolkit* (CAP) as basic framework which has been developed for ambient vibration array analysis by Ohrnberger & the SESAME Team WP05/WP06 (2004). Existing functions, such as f-k and estimation of the real covariance matrix, have been integrated in our modified and adapted version of the program. Data for CAP is provided by the GEOPSY data base. GEOPSY is part of the SESARRAY software package for processing ambient seismic vibrations and was implemented by Marc Wathelet (IRD-LGIT, Grenoble, France, www.geopsy.org).

From an analyzed record, features for each time window are extracted from the CAP output files and stored into row-column ASCII files. Feature selection is again implemented as a shell script, which calls the MATLAB® programs for the Wald-Wolfowitz runs test (Level 1) and CP-SOM clustering (Level 2 and 3). For the runs test we use the implementation of Wei Li (University of Chicago Graduate School of Business, Jan 29, 2004). SOM processing, including CP-SOM learning and clustering, is done using the MATLAB® SOM toolbox of Vesanto et al. (2000). Furthermore, SOM toolbox functions have been adapted and newly implemented.

Except of illustrations showing SOM visualization such as U-Matrix, CPs, and SOM Coloring, all figures have been created with the Generic Mapping Toolbox (GMT) written by Wessel & Smith (1995). For the generation of the synthetic data, we used the software of Herrmann (2001).

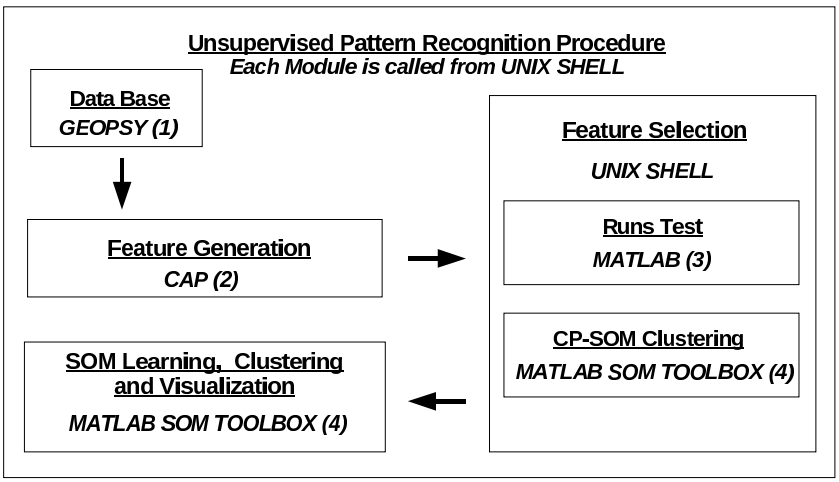

Figure 8.1: Implementation scheme of our unsupervised processing appraoch. Italic text state the employed software. The numbers correspond to the following references: (1) www.geopsy.org, (2) Ohrnberger & the SESAME Team WP05/WP06 (2004), (3) Wei Li (2004), (4) Vesanto et al. (2000)

8.2 Synthetic Data Sets

Table 8.1: Subsurface velocity model used for generation of synthetic data sets.

Thickness [m]	Vp [m/s]	Vs [m/s]	Density [$\frac{g}{cm^3}$]
35	542	209	1.762
178	851-1131*	393-522*	1.736-1.85*
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
* gradient			

Table 8.2: Frequency bands used for feature generation (synthetic data sets).

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
Methods 3,4,7			**Methods 1,2**		
1	0.36 Hz	0.84 Hz	1	0.49 Hz	0.91 Hz
2	0.81 Hz	1.88 Hz	2	0.79 Hz	1.46 Hz
3	1.80 Hz	4.20 Hz	3	1.26 Hz	2.34 Hz
Methods 5,6					
1	0.24 Hz	0.36 Hz	6	2.11 Hz	3.16 Hz
2	0.37 Hz	0.56 Hz	7	3.26 Hz	4.89 Hz
3	0.57 Hz	0.86 Hz	8	5.03 Hz	7.55 Hz
4	0.88 Hz	1.33 Hz	9	7.77 Hz	11.66 Hz
5	1.37 Hz	2.05 Hz	10	12.00 Hz	18.00 Hz

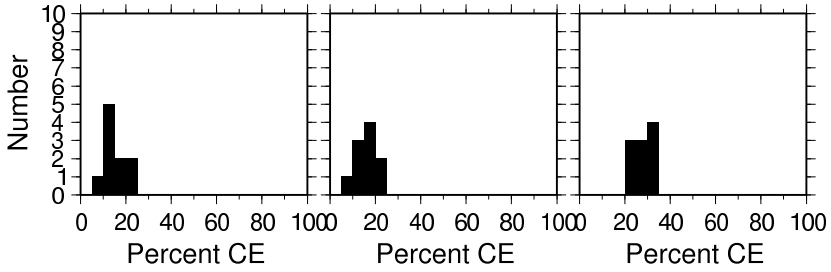

Figure 8.2: Statistic for classification errors (CE) computed for each cross-validation fold (Synthetic data). Three CV experiments are shown: CV for SOM learning and clustering including feature selection (data set 1) and CV without feature selection (data set 1 and data set 2).

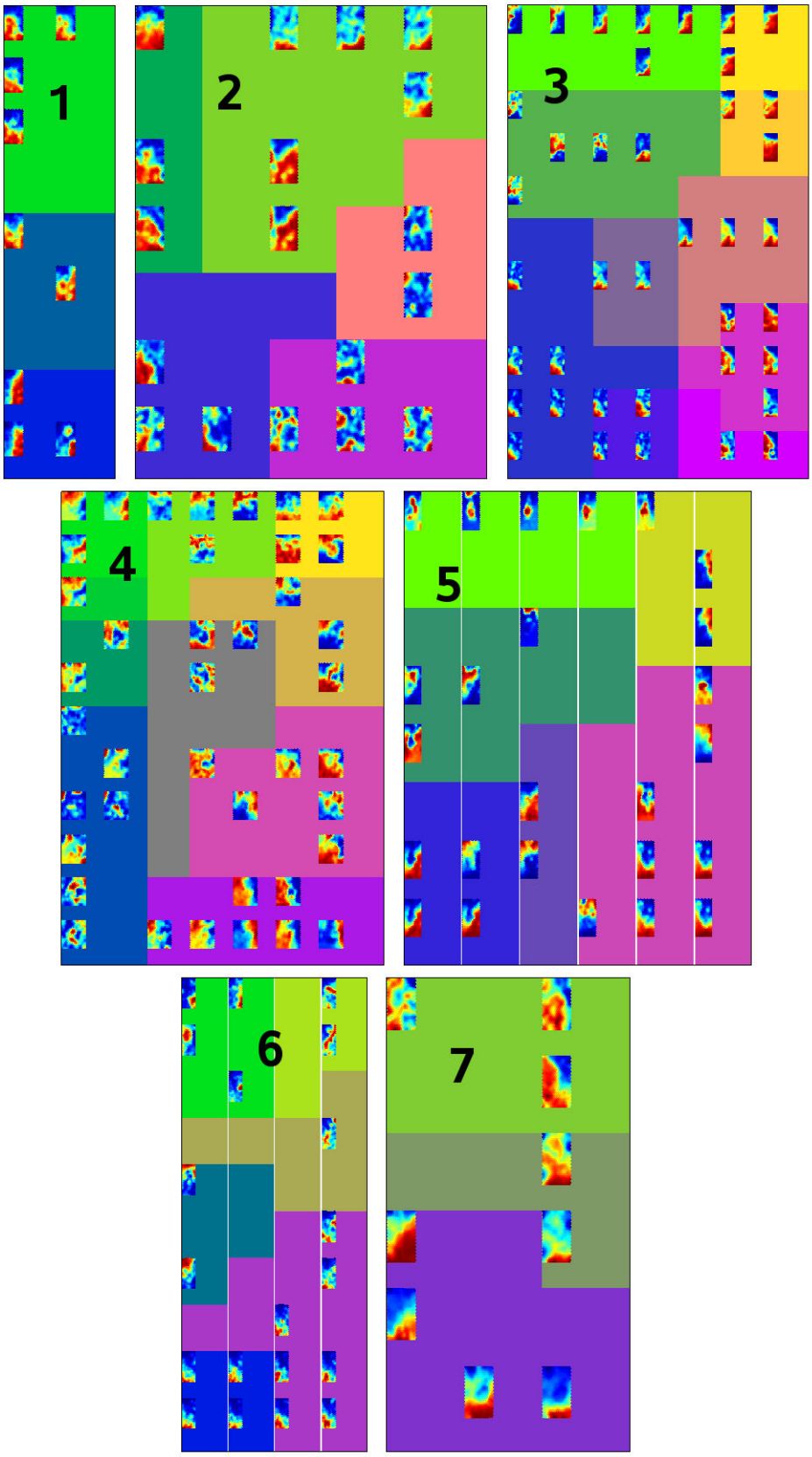

Figure 8.3: Component planes on CP-SOMs and obtained clustering (background coloring) for each feature generation method (feature selection Level 2) using data set 1. Method index is given by black numbers.

8.3 Earthquake Data Set

Table 8.3: Frequency bands used for feature generation (earthquake data set).

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
	Methods 3,4,7			**Methods 1,2**	
1	0.09 Hz	0.21 Hz	1	-	-
2	0.37 Hz	0.86 Hz	2	-	-
3	1.50 Hz	3.50 Hz	3	-	-
			Methods 5,6		
1	0.14 Hz	0.17 Hz	6	0.64 Hz	0.78 Hz
2	0.19 Hz	0.23 Hz	7	0.88 Hz	1.08 Hz
3	0.25 Hz	0.31 Hz	8	1.20 Hz	1.47 Hz
4	0.35 Hz	0.42 Hz	9	1.65 Hz	2.01 Hz
5	0.47 Hz	0.58 Hz	10	2.25 Hz	2.75 Hz

Table 8.4: Source times of earthquakes employed from station RDO.

Day	Time	Day	Time	Day	Time
2003/12/06	00:18:57	2005/05/14	23:46:47	2006/03/06	10:40:45
2003/12/07	15:40:16	2005/05/15	10:54:27	2006/03/22	11:26:16
2003/12/23	12:23:37	2005/05/29	08:55:35	2006/04/12	16:51:59
2004/02/09	03:48:12	2005/06/14	07:31:47	2006/04/17	02:44:04
2004/02/26	04:13:59	2005/07/04	21:33:06	2006/05/10	07:01:42
2004/05/23	15:19:09	2005/07/30	21:45:01	2006/05/19	23:01:53
2004/07/24	19:00:54	2005/07/31	23:41:34	2006/06/17	12:13:59
2004/08/04	03:49:34	2005/08/04	05:47:40	2006/06/21	15:54:44
2004/09/27	09:16:22	2005/08/24	03:06:19	2006/06/24	02:49:26
2004/10/27	20:34:32	2005/09/08	16:35:50	2006/08/22	09:23:20
2004/12/02	15:51:48	2005/09/16	08:08:51	2006/09/08	22:39:09
2004/12/09	18:35:18	2005/09/19	22:17:21	2006/09/26	23:51:05
2005/01/03	21:44:28	2005/10/17	09:55:31	2006/10/04	17:34:18
2005/03/13	01:31:13	2005/12/19	17:44:47	2006/11/24	04:37:37
2005/04/29	22:28:06	2006/02/20	17:20:09		

Table 8.5: Source times of earthquakes employed from station KEK.

Day	Time	Day	Time	Day	Time
2003/11/18	18:36:20	2005/05/29	08:55:35	2006/05/15	04:42:50
2003/11/24	15:51:07	2005/05/30	09:20:04	2006/05/29	02:20:04
2003/12/17	23:15:16	2005/06/30	19:44:50	2006/06/13	14:15:41
2004/02/09	03:48:12	2005/07/02	17:35:19	2006/06/17	12:13:59
2004/03/18	15:14:23	2005/07/03	22:39:22	2006/06/24	02:49:26
2004/03/28	14:54:35	2005/07/07	11:45:12	2006/07/13	15:13:43
2004/05/25	05:34:24	2005/07/23	13:09:23	2006/08/16	18:56:39
2004/07/24	19:00:54	2005/08/01	13:34:58	2006/08/16	19:17:48
2004/09/21	14:15:02	2005/09/19	22:17:21	2006/08/27	12:49:32
2004/10/07	07:16:51	2005/10/18	15:36:31	2006/09/08	22:39:09
2004/12/25	20:22:17	2005/11/19	17:26:33	2006/09/26	23:51:05
2005/01/03	21:44:28	2005/11/27	06:38:11	2006/10/04	17:34:18
2005/01/23	22:36:05	2005/12/19	17:44:47	2006/10/12	13:31:00
2005/01/28	00:17:44	2005/12/09	14:33:20	2006/10/19	09:34:51
2005/02/08	16:38:23	2006/02/20	17:20:09	2006/10/24	03:39:36
2005/02/12	12:13:47	2006/02/27	04:34:01	2006/11/03	10:50:57
2005/03/11	11:05:21	2006/04/13	23:25:31	2006/11/07	11:13:36
2005/03/20	00:51:03	2006/05/10	07:01:42	2006/11/24	04:37:37
2005/05/14	23:46:47	2006/05/11	16:55:34	2006/11/26	17:58:09

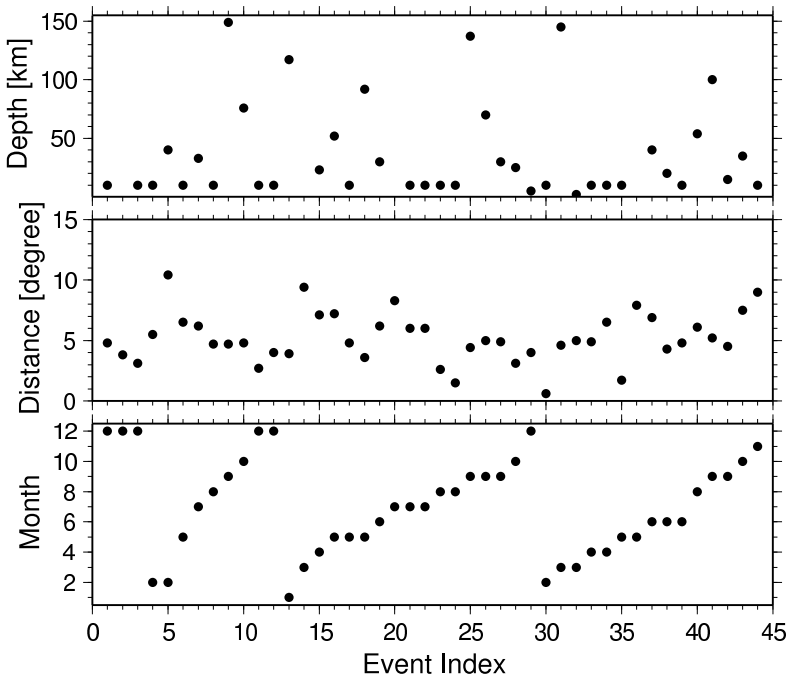

Figure 8.4: Event properties for earthquakes recorded at station RDO.

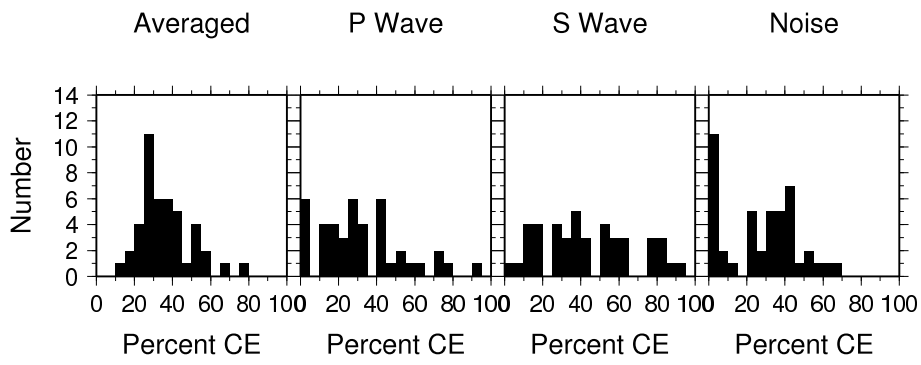

Figure 8.5: Statistic for classification errors (CE) for class P Wave, S Wave and Noise computed for each cross-validation fold (earthquake data set, RDO).

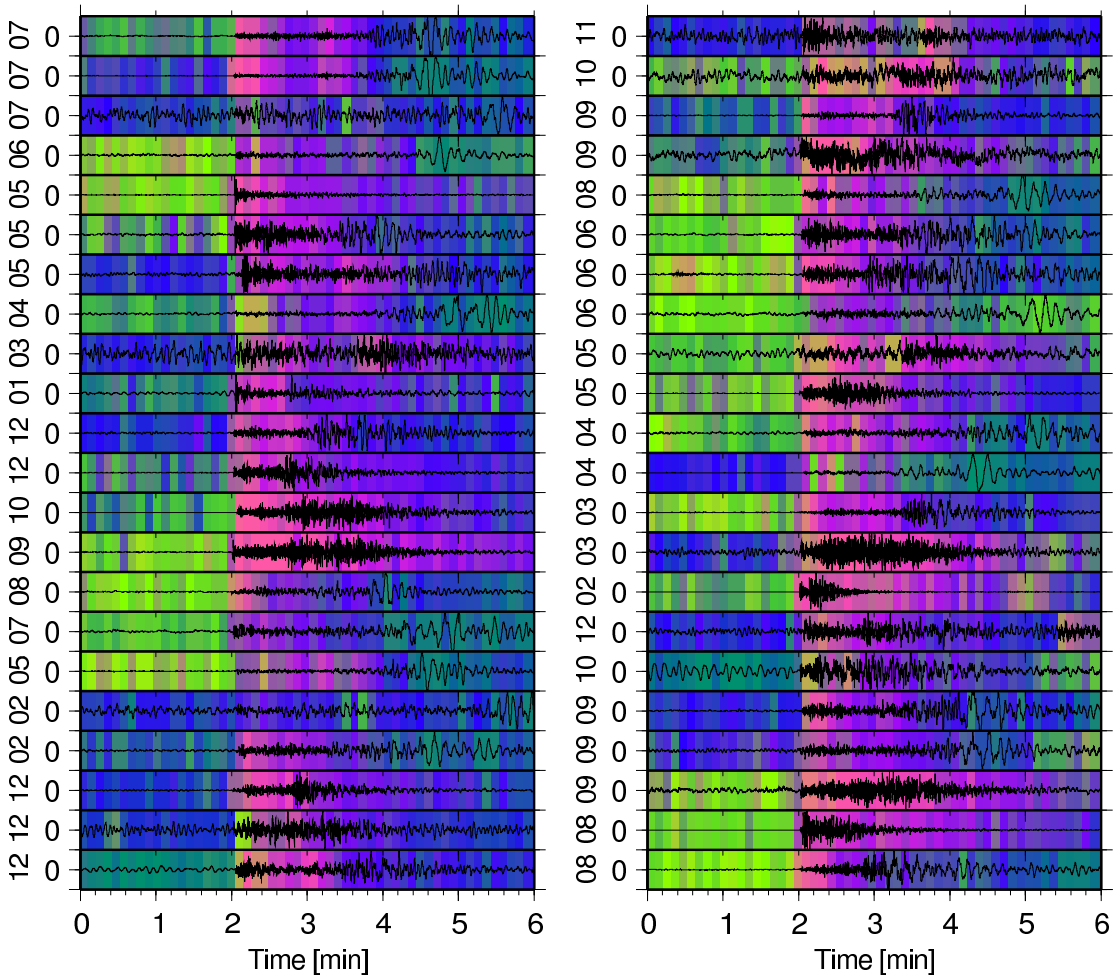

Figure 8.6: Vertical component seismograms of 44 events recorded between 2003 and 2006 at station RDO. Y-axis labels indicate the months in which the events occurred. The amplitudes are normalized by the maximum of each trace. SOM similarity coloring in Fig 5.19c is used for the background (each SOM prototype has a different color).

8.4 Merapi Data Set

Table 8.6: Frequency bands used for feature generation (Merapi data set).

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
	Methods 3,4,7			**Methods 1,2**	
1	0.51 Hz	2.89 Hz	1	0.39 Hz	2.21 Hz
2	1.24 Hz	7.01 Hz	2	0.75 Hz	4.25 Hz
3	3.00 Hz	17.00 Hz	3	1.44 Hz	8.16 Hz
			Methods 5,6		
1	0.67 Hz	0.93 Hz	6	3.55 Hz	4.90 Hz
2	0.95 Hz	1.30 Hz	7	4.95 Hz	6.84 Hz
3	1.31 Hz	1.81 Hz	8	6.91 Hz	9.54 Hz
4	1.82 Hz	2.52 Hz	9	9.64 Hz	13.31 Hz
5	2.55 Hz	3.51 Hz	10	13.44 Hz	18.56 Hz

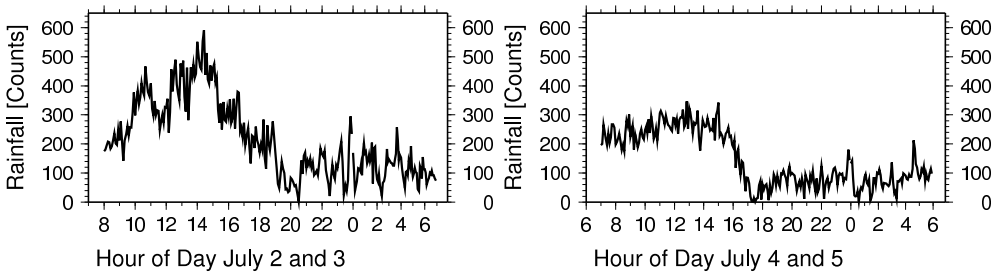

Figure 8.7: Rainfall distribution for the time period which is used for unsupervised wavefield analysis at array GRW.

8.5 Ambient Vibration Data Sets

8.5.1 Pulheim

Table 8.7: Frequency bands used for feature generation: Pulheim

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
1	0.40 Hz	0.60 Hz	6	0.98 Hz	1.47 Hz
2	0.48 Hz	0.72 Hz	7	1.17 Hz	1.75 Hz
3	0.57 Hz	0.86 Hz	8	1.40 Hz	2.10 Hz
4	0.68 Hz	1.03 Hz	9	1.67 Hz	2.51 Hz
5	0.82 Hz	1.23 Hz	10	2.00 Hz	3.00 Hz

Table 8.8: Automatically selected features: Pulheim - short-term patterns

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
domfz	16.2	*PQ1*	6.5	*H/V5*	3.7
bbz	14.0	*prr1*	6.0	*dop1*	3.4
sonoh10	10.8	*sop1*	5.0	*pdiffez2*	2.6
sonoz9	8.5	*inc10*	4.4	*a_b2*	2.5
rect1	8.5	*dopIII7*	4.1	*spacimr6*	2.1
sonoh3	7.5	*ell1*	3.8	*plan5*	2.1
sonoz8	6.8	*ifz2*	3.7		

Table 8.9: Automatically selected features: Pulheim - long-term patterns

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
sonoh10	18.0	*dopIII9*	13.3	*domfh*	10.6
bbh	17.5	*ell9*	12.9	*spacimt5*	10.4
ifh10	16.6	*ifh9*	12.7	*sonoz9*	9.7
domfz	16.4	*ifh1*	12.0	*H/E1*	9.2
sonoz3	15.9	*ifh6*	11.6	*dopIII7*	7.9
prr9	13.8	*H/V8*	11.2	*spacimr2*	6.4
planII6	13.8	*ifh5*	11.1	*elip4*	6.0

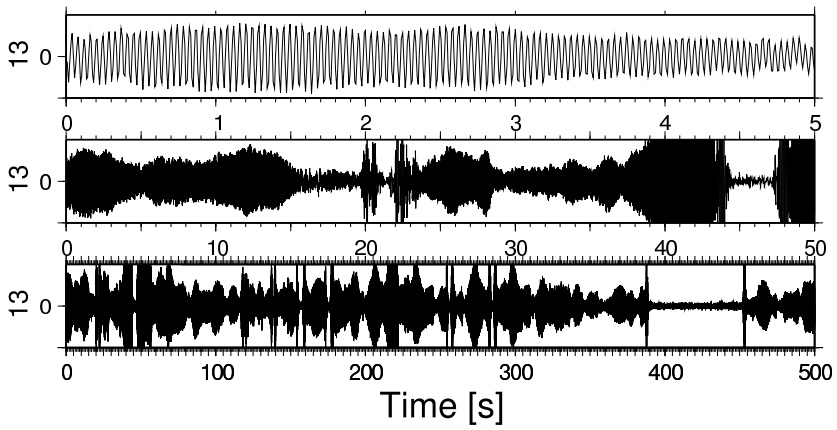

Figure 8.8: Monochromatic signal recorded by several stations of the Pulheim array. Upper traces are zooms of the one below starting at the beginning of each trace.

8.5.2 Lörrach

Table 8.10: Frequency bands used for feature generation: Lörrach

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
1	0.64 Hz	0.96 Hz	6	1.77 Hz	2.66 Hz
2	0.79 Hz	1.18 Hz	7	2.17 Hz	3.26 Hz
3	0.96 Hz	1.44 Hz	8	2.66 Hz	3.99 Hz
4	1.18 Hz	1.77 Hz	9	3.26 Hz	4.90 Hz
5	1.45 Hz	2.17 Hz	10	4.00 Hz	6.00 Hz

Table 8.11: Automatically selected features: Lörrach - short-term patterns

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
sonoz10	19.3	*elip1*	9.1	*dopIII8*	4.0
ratiolf	18.9	*rect3*	9.0	*dopIII7*	4.0
sonoh10	18.1	*dopIII1*	7.5	*dopIII5*	3.6
domfh	18.0	*prr5*	6.8	*prz1*	3.5
bbh	17.4	*H/V6*	6.1	*linII8*	3.5
sonoh1	15.5	*prt1*	5.5	*prr3*	3.5
sonoh8	15.5	*inc3*	5.4	*a_b10*	3.3
sonoh9	13.4	*prr6*	5.2	*FQ1zh9*	2.7
H/V2	12.8	*ifh7*	4.9	*aroz*	2.4
H/E2	12.8	*inc1*	4.9	*sdc3*	2.4
rect4	11.7	*dopIII6*	4.5		
elip4	9.3	*P/Q2*	4.4		

8.5.3 Hamburg

Table 8.12: Frequency bands used for feature generation: Hamburg

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
1	0.24 Hz	0.36 Hz	6	1.49 Hz	2.23 Hz
2	0.35 Hz	0.52 Hz	7	2.14 Hz	3.21 Hz
3	0.50 Hz	0.75 Hz	8	3.09 Hz	4.63 Hz
4	0.72 Hz	1.08 Hz	9	4.44 Hz	6.67 Hz
5	1.03 Hz	1.55 Hz	10	6.40 Hz	9.60 Hz

Table 8.13: Automatically selected features: Hamburg - short-term patterns

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
domfh	19.6	*sonoz8*	11.5	*ifz1*	7.0
bbz	18.6	*H/V3*	11.5	*ellez3*	5.6
sonoh10	17.1	*prt2*	10.3	*P/Q3*	5.2
sonoz10	15.9	*sonoh5*	9.6	*dop1*	5.1
H/V1	15.4	*ifh1*	9.2	*ell3*	4.1
sonoh7	13.9	*inc2*	9.1	*dopIII10*	3.9
rect1	13.6	*elip2*	8.8	*planII9*	3.5
ellzn1	13.4	*pdiffez1*	8.0	*inc6*	3.2
aroz	11.7	*tiltzn1*	7.9	*prz5*	3.1

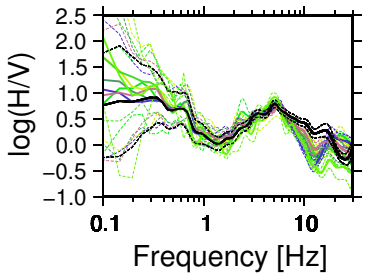

Figure 8.9: H/V spectral ratio and standard deviation for SOM clusters (colored) and using all time windows (black) for one station of Hamburg recordings.

8.5.4 Lüneburg

Table 8.14: Frequency bands used for feature generation: Lüneburg

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
1	1.60 Hz	2.40 Hz	6	4.33 Hz	6.49 Hz
2	1.92 Hz	2.93 Hz	7	5.28 Hz	7.92 Hz
3	2.38 Hz	3.57 Hz	8	6.45 Hz	9.67 Hz
4	2.91 Hz	4.36 Hz	9	7.87 Hz	11.80 Hz
5	3.55 Hz	5.32 Hz	10	9.60 Hz	14.40 Hz

Table 8.15: Automatically selected features: Lüneburg - short-term patterns

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
sonoh7	38.2	*prz8*	13.3	*prz*	16.5
bbh	36.5	*ellez3*	12.5	*prt*	16.2
ratiolf	35.0	*ell1*	11.5	*planII4*	6.1
sonoh10	33.2	*H/V9*	9.7	*dop4*	5.7
sonoh5	23.7	*inc7*	9.7	*pdiffez1*	5.4
dopII1	22.0	*a_b6*	8.8	*ellzn4*	5.3
sonoz1	21.3	*rect7*	8.6	*a_b2*	5.1
H/V1	19.6	*H/V5*	8.6	*planII5*	4.1
ifh9	16.5	*elip3*	8.5	*prt9*	4.1
ifh7	15.8	*prr1*	7.8	*pdiffzn3*	3.4
elip1	15.1	*ifz4*	7.3	*ell10*	2.9
ifh5	13.9	*pdiffzn8*	6.7	*sop4*	2.4
H/V4	13.5	*linIII7*	6.5		

Table 8.16: Automatically selected features: Lüneburg - long-term patterns

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
prr8	51.1	*sonoh10*	25.6	*rect1*	10.6
prz8	51.0	*rect8*	20.9	*H/V9*	10.2
prt3	48.8	*ifh7*	20.3	*vifz4*	9.5
prz3	48.6	*sonoz7*	17.9	*dopII8*	9.1
prr4	48.3	*sonoz8*	16.2	*ell7*	9.1
prt9	48.3	*H/V4*	14.6	*3cell9*	8.9
prz6	47.9	*H/E5*	14.1	*spacimz4*	7.5
prt2	47.9	*pdiffzn9*	14.1	*ell10*	7.3
prz1	47.7	*sonoh6*	12.2	*inc5*	6.6
prr6	47.5	*elip4*	12.2	*ell8*	6.5
ifh8	31.9	*pdiffez3*	12.1		
sonoz1	26.8	*rect6*	10.9		

8.5.5 Colfiorito

Table 8.17: Frequency bands used for feature generation: Colfiorito

Band No.	Frequency		Band No.	Frequency	
	Lower	Upper		Lower	Upper
1	0.40 Hz	0.60 Hz	6	0.98 Hz	1.47 Hz
2	0.48 Hz	0.72 Hz	7	1.17 Hz	1.75 Hz
3	0.57 Hz	0.86 Hz	8	1.40 Hz	2.10 Hz
4	0.68 Hz	1.03 Hz	9	1.68 Hz	2.51 Hz
5	0.82 Hz	1.23 Hz	10	2.00 Hz	3.00 Hz

Table 8.18: Automatically selected features: Colfiorito - long-term patterns

Feature	Z_{test}	Feature	Z_{test}	Feature	Z_{test}
sonoh3	16.9	*prr1*	11.1	*spacimr3*	4.1
sonoh4	16.1	*prt5*	9.5	*spacimt2*	3.9
sonoh5	13.3	*prr4*	8.5		
prt3	11.2	*prz4*	6.0		

Acknowledgments

This work would not have been possible without material, assistance and help provided by many people. Regarding the implementation of software, I have been in the favorable situation that I was able to start with existing packages such as CAP (Ohrnberger & the SESAME Team WP05/WP06, 2004) and the SOM toolbox (Vesanto et al., 2000). Furthermore, the availability of such a large database of ambient seismic vibration recordings has been a really comfortable situation. In particular, I employed data acquired within the SESAME, HADU and NERIES research projects. Within this context, I am grateful that I had the possibility to participate in a couple of field measurements, which allowed me to pause my, sometimes mentally exhaustive, work in front of the computer. Moreover, in the famework of the European NERIES project (contract no. 026130), data from the European broadband network was kindly provided to me. Finally, I am glad for the availability of the Merapi data set.

I am much obliged for any ideas, comments and suggestions which I have received within the last three years. I want to thank all my colleagues from the seismology as well as the applied geophysics working group. Special thanks go to Carsten Riggelsen for his perspective from computer science and the consistency check for the notation in Chapter 4, to Gudrun Richter for providing me rainfall data from Mount Merapi, and, finally, to the "Fryday Beer Group".

Most of my gratitude goes of course to Matthias Ohrnberger and my supervisor Frank Scherbaum. During the years at this institute, including my Diploma thesis and the years before, they have been significantly contributing to finding my intrests in the field of pattern recognition, array seismology and ambient vibration analysis. I am grateful for any inspirations and the opportunities they offered me. I am much obliged to Matthias for his support during this work.

The study has been kindly funded by a postgraduate scholarship of the University of Potsdam (Graduiertenförderung der Universität Potsdam, Land Brandenburg). Moreover, parts of the funding have been made possible under the before-mentioned research projects (e.g. NERIES, JRA5, data mining tools).

Last but not least I want to say thank you to all my friends (SMD, FEG, ...) and to my family. In particular, I am very grateful to my parents for their unconditional and continuous support, not only during and regarding this work. Thanks God for being the wind in my sails.