UNIVERSITÀ DEGLI STUDI DI PALERMO
DIPARTIMENTO DI MATEMATICA E INFORMATICA
Dottorato di Ricerca in Matematica e Informatica

Metrics, clustering and simulations to evaluate seismic signals

Settore Scientifico Disciplinare
INF/01

TESI DI
Francesco Benvegna

COORDINATORE DEL DOTTORATO
Prof.ssa L. Di Piazza

TUTOR
Prof. Domenico Tegolo

CO-TUTOR
Dott. Giosué Lo Bosco

- XXIII Ciclo -

Abstract

This thesis presents an overview on seismic signals analysis and its related activities to clustering. The real applications require the use of metrics, algorithms and data to test hypothesis or to infer them. Hypocenter and focal mechanism of an earthquake can be determined by the analysis of signals, named waveforms, related to the wave field produced by earthquakes and recorded by a seismic network. Assuming that waveform similarity implies the similarity of focal parameters, the analysis of those signals characterized by very similar shapes can be used to give important details about the physical phenomena which have generated an earthquake. Recent works have shown the effectiveness of cross-correlation and/or cross-spectral dissimilarities to identify clusters of seismic events. In this thesis we propose a new dissimilarity measure between seismic signals whose reliability has been tested on real seismic data by computing external and internal validation indices on the obtained clustering. Results show its superior quality in terms of cluster homogeneity and computational time with respect to the largely adopted cross correlation dissimilarity.

Acknowledgments

I owe a great deal of thanks to many people for making this thesis possible. First, I would like to express my gratitude for my advisor Prof. Domenico Tegolo, who has been leading and supporting me and my research to be fruitful in his patience.

A huge thanks goes to Giosuè Lo Bosco for his fundamental and precious collaboration in all aspects of my work. Without his skillful and infinite support my projects would not have been possible.

Thanks to Prof. Vito Di Gesù passed away too early to support me in the last challenge. His colleagues and his students have helped me at all times.

Thanks to my fellow PhD friends, Luca Pinello and Alessio Langiu for our broad-ranging discussions and for sharing the joys and worries of the academic research.

Furthermore, I am deeply indebted to my colleagues at Department of Mathematics and Computer Science that have provided the environment for sharing their experiences about the problem issues involved as well as participated in stimulating team exercises developing solutions to the identified problems.

Finally, I wish to express my gratitude to my wife Katia and my daughters Rebecca and Alice who provided continuous understanding, patience, love and energy.

Thanks to all of you.

Originality Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the test. I give consent to this copy of my thesis, when deposited in the University Library, begin available for loan and photocopying.

Signed..... February 2013

Contents

List of Figures

List of Tables

Chapter 1

Seismograms Analysis

1.1 Introduction

In the last years computer science and data processing have contributed in the analysis of huge amount of data in many fields. A lot of works have been done in computer vision, biology and signal processing. In the past many activities, involving small data, have been performed by people but now thousands or millions of multidimensional signals are collected, and computing system becomes an essential tool for scientific research.

Also sismology, statistics and computer science help researchers to process a great number of signals collected everywhere in the world. Each state has a network of monitoring stations which record all detected seismic events. So far, we have a lot of data to be stored, processed and analyzed. Seismic data are used to solve many problem about classification of the seismic signals and the causes that have generated them.

The seismograms are recordings of ground motion which record how the movements have taken place and how have been transmitted through the ground. Here we focus on earthquakes although we'll work with natural and artificial data. The earthquake seismograms are used to investigate on common characteristics of a group of events so that any type of correlation can be verified between them. Any relationship among events may be very important for the description of the earthquakes and for their common properties.

The source of an earthquake can be described as the release of stored energy due to motion of faults. The geometry of the movements of the faults may lead to quite different seismic events. For this reason the study of the clustering of seismic events can be useful in the characterization of the events that caused them.

The events are catched by devices called seismographs. These devices transform the ground motion into recording data on the paper or archive. The records are used by skilled researchers to identify and characterize the detected seismic event. Today, almost all are digital seismographs and the records are analyzed by using signal processing techniques.

1.2 Seismic Waves

Seismic waves are spread of energy generated by an earthquake, explosion, or a volcano with low-frequency acoustic energy. They travel through the different level of the Earth's underground where they can be deviated and reflexed by each layer of the earth's crust. Seismic waves are studied by seismologists, abd the tools used on recording are seismometers, hydrophones (in water) and accelerometers.

Once the records are collected in large archive, one of the first and important activity for the study of the signal is the phases picking. As described in [DKB] the *phase picking* is the a set of activities executed to define the main phases of a seismic signal: phase timing, phase identification and first motion polarity. By the picking the sismologists discover a lot information about the waves propagated through the Earth.

Seismic signals are elastic waves that propagate in solid or fluid materials. During a seismic event it is possible to identify different groups of waves characterized by several amplitudes and frequencies. These groups are called phases and may be of different types depending from the waves and their propagation.

The main types of waves that are identified are:

- *body waves*

  P-waves: **p**rimary waves are compressional waves. These are the first waves to arrive and to be detected by the instruments. Ground motion is perpendicular to wave direction (see fig. 1.1)

  S-waves: **s**econdary waves are shear waves that arrive after the p-waves. Ground motion is parallel to wave direction (see fig. 1.1)

- *surface waves*: are the combination of the p-waves and s-waves reflections traveling along the surface. Most important are:

  Love waves: produce entirely horizontal motion

  Rayleigh waves. moves the ground up and down, and side-to-side in the same direction that the wave is moving

P and S-waves are called body waves since they travel in the interior of the Earth as opposed to surface waves. The differences between them are the transmission geometry and velocity. For example in the upper crust the typical p-waves velocity is about 6 km/s while s-waves go at 3.5 km/s. The amplitude of the body waves through an homogeneous elastic medium decrease with the distance and at the same distance the power of the s-waves is greater than p-waves.

In a transmission medium with multiple laterally unlimited layers there are many discontinuities between one layer and another. Here the surface waves are generated. Travelling only through the crust, surface waves are of a lower frequency than body waves. The name *surface* is due to the fast decay of the wave amplitude away from the surface.

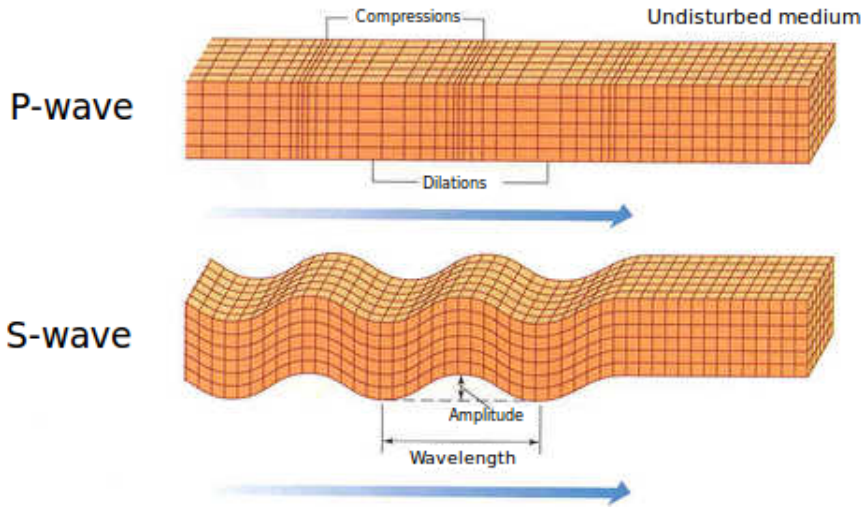


Figure 1.1. P-waves and S-waves propagation

The waves velocity relies on the structure and on the type of the rock along the propagation path. The velocity's parameters of each means are a function of several factors such as mineralogy, density, temperature and lithostatic pressure.

For example in a material like petroleum, water or mud the p-waves have a velocity around 5 km/s while on dolomite or gypsum this value maybe up to $20 - 22$ km/s. The path from the source of the seismic event to the point of detection must be analyzed during the study of seismic events.

Hauskov in [HO10], defined the velocity ratio between P and S velocities, vp/vs equal to $\sqrt{3} = 1.73$. In practice it is often closer to 1.78

The complexity of the medium and source of the seismic event make a seismic wave with uniques amplitude and shape. Altogether the main features of one wave are usually detectable. So it's not difficult to identify p-waves, s-waves and surface waves but they can not be immediately visible. Sometimes it is necessary to cut and filter the signal to identify the waves arrival.

Figure 1.2 depicts an example of seismogram with well identified waves type. For the wave on figure is quite easy to find the p and s arrivals. The signal is well filtered so that any noise cannot disturb the analysis of waves type.

The seismogram is stored as single signal but it is an overlap of multiple waves. The complexity in shape and frequency as shown in figure 1.3 is the results of these overlappings. Although there are a lot of information on these two features, we'll introduce in the next chapters, that a course grained proximity measure able to discriminate signals with equals or better results respect to the most used distance based on cross-correlation.

The origin of the seismic events are of two different type: along *plate edges* and along *faults*. The events along plate edges are caused by movement of the lower layer so that near plates are involved in compression, dilation and slip. These edge movements are origin of earthquakes. A fault is a planar fracture or discontinuity in a volume of rock, usually inside a plate, caused by plate tectonic forces. The movements of the fault surfaces cause an earthquake.

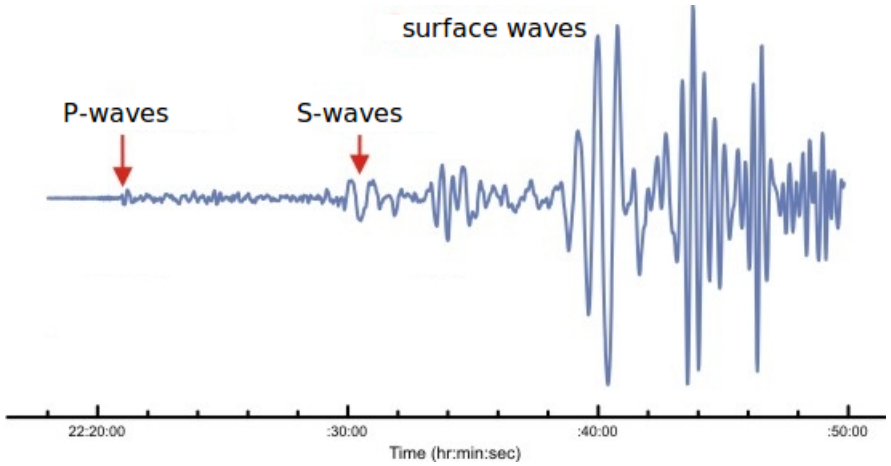
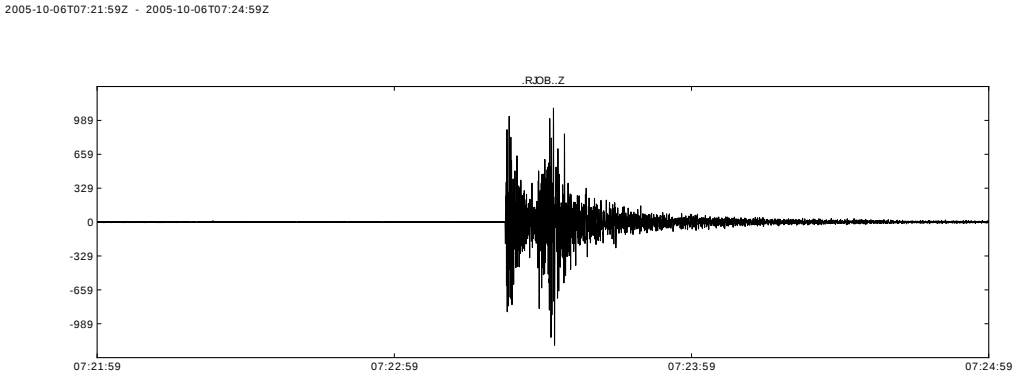
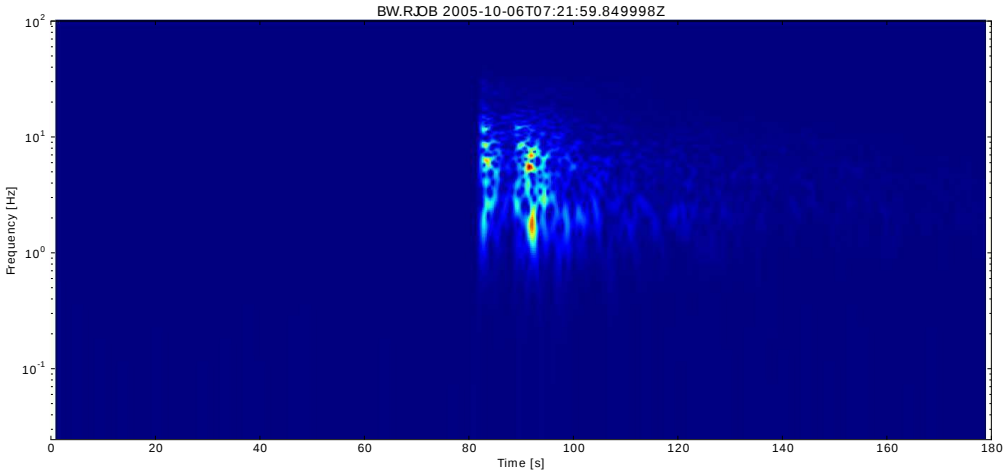


Figure 1.2. Seismogram example



(a) Seismogram of a real signal



(b) Spectrogram

Figure 1.3. A real seismic signal and its relative spectrogram

Figure 1.4 shows three types of faults: thrust, normal and strike-slip. Each type is caused by different forces.

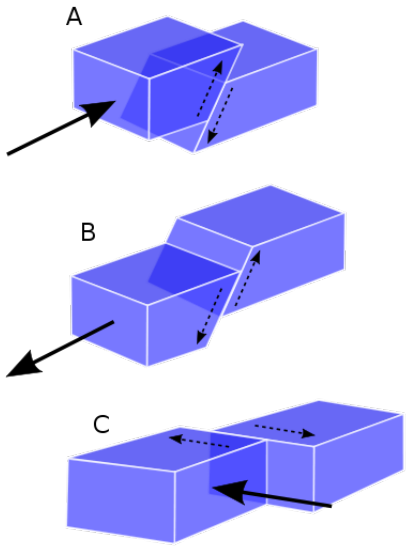


Figure 1.4. Fault type: A) thrust fault B) Normal fault C) Strike-slip fault

Understand the origin of a seismic events is very important to know the ground structure and the dynamics involving the faults. To infer the characteristics of any future event, the seismologists need to know the mechanisms that cause earthquakes with sufficient precision.

1.3 Working with seismograms

Many research activities are related to seismic waves but two of them are very interesting and related to data analysis: *locating hypocenters* and *investigating on focal mechanisms*. The first problem tries to give a suitable location to an unlocated event by the comparison to a well located master event, while the second one is related to the physical and mechanical hypothesis inferred by results of clustered seismic events. Of course data analysis make possible to automate many computations and retrieve solution which maybe investigated by geologists.

1.3.1 Locating hypocenters

The earthquake location is one of most frequent tasks everywhen a seismic event is occurred. The site where the event occured is called **hypocenter** and it is expressed as three values (x, y, z) which are longitude, latitude and depth respectively. Moreover the *origin time* is the time when the event is occurred at hypocenter.

The first step on locating hypocenter is the computation of the distance and the magnitude. As we can see from *"old"* figure 1.5 it is possible by simple measurements of the trace plotted by analog seismometer. The scales, reported below in the figure,

help the analyst to find the distance and the magnitude by fixing two values and tracing one line.

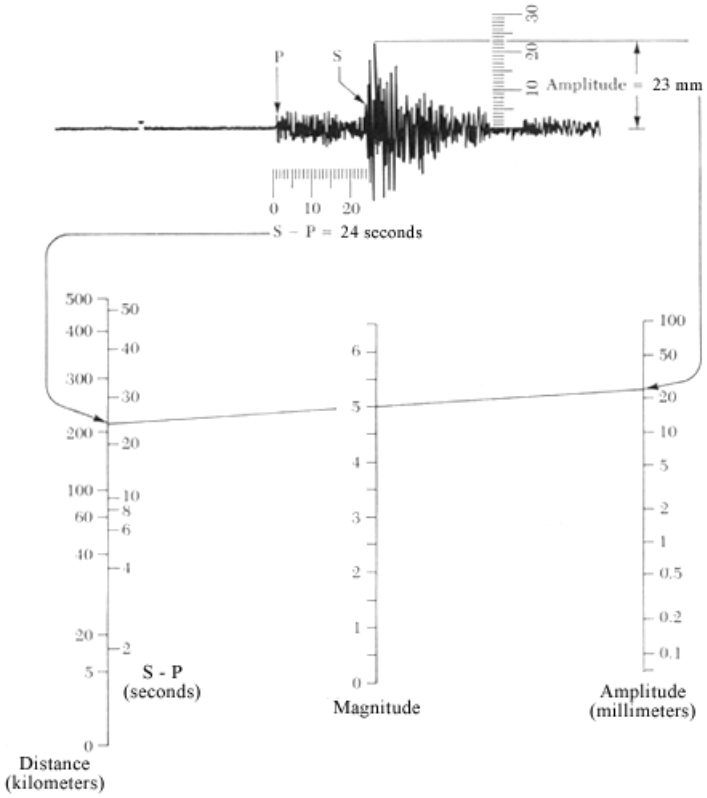


Figure 1.5. Old tecnique schema used to compute distance and magnitude from seismogram

Indeed, today is easy to compute the hypocenter by a mathematical model that requires at least three stations but also not so easy by one. The most diffused method used a set of stations to triangulate data is shown in figure 1.6.

When three station are not available it is possible to compute the hypocenters with only one if and only if it is available 3-component station, thus we use the polarization to compute the direction and the phases to compute the distance. It is clear the p-phase and s-phase identification need to fix assumptions on velocity model. Although this method is available, in practice is quite difficult because the mandatory preprocessing required to find the phases and moreover, the computation of correlation among components may be hard in many case.

Another used technique is the *relative location* by a master event. This method may compute on many cases an absolute location with an accurancy greater than of methods above. Of course, a well-known event called *master event* must be localized with high precision. Through a similarity measure it is possibile to verify if the event to localize is near the master. If the two signal shape and polarization are quite similar it is possible to infer that two events are generated from two sources in the same location. As reported in [HO10], the latter method run very well when applied after a coarse localization.

The last cited method is the *double difference earthquake location* where the

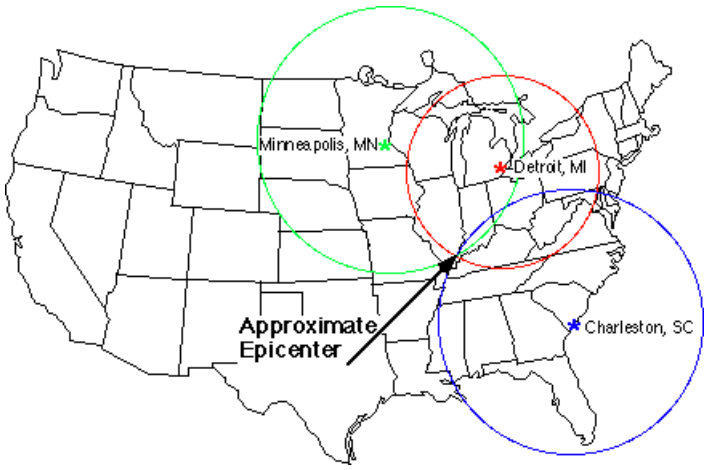


Figure 1.6. Epicenter triangulation by three station located on different places

difference between observed and calculated travel time differences of event pairs (called double difference) is minimized, rather than the difference between observed and calculated travel times for single event. This method is used in practice on large dataset.

1.3.2 Investigating Focal mechanisms

The focal mechanisms define the properties of the source of seismic events. In fault-related event the focal mechanism are called *fault-plane solution* because it describes the orientation of the fault and the slip of the ground. In order to compute focal mechanisms, the seismogram must be analyzed on amplitudes and polarities.

While the computing of fault-plane solution is a work of the seismology area, the grouping of similar events that infer the same solution is a work near data analysis. Grouping the events in some automatic way could help the seimologists to retrieve in less time many properties of the earthquake source.

The unsupervised techniques like clustering are useful to find a groups on related data. The elements are grouped by the application of algorithms and measures suitable on seismograms. In the past years many similarity measures were developed by researchers enrolled in the fields of statistics or signal processing. But overall, the real problem are the seismic signals which have distinctive features from other types of signal. First of all we must consider that a seismic signal is the overlapping of many waves with different characteristics. For example the first part of the seismogram has p-waves while the next has s-waves and so on.

Most of techniques use some measures to compare signals. In seismology the most used measure on signals comparison is the cross-correlation. It is a well known function by statisticians but its use and application is very large on signal processing. On seismogram analysis the cross-correlation is used to compute the delay between two signals or the similarity between them. In the latter case the maximum will be taken in consideration.

Of course the cross-correlation is not only the measure but it is a standard the facto standard. In this thesis we propose a new, and simpler, measure that applied

as a distance outperforms in many cases the max cross-correlation.

1.3.3 Dataset: from analog to digital

The availibility of digital seismometer has made possible the collection of a large amount of data. Today it is possible to access to many site on Internet that have a lot of free data. A detection station with continuously recording instruments can collect about 30-50MBytes of data per day at a 3-component station with 100 Hz sampling rate.

The data may be available on different format and the switch from analog system to digital system enables a simple recording of the collected data. The advantages of this change are a simple storing of a large amount of data, less error due to less mechanisms and the availability of digital analysis tools.

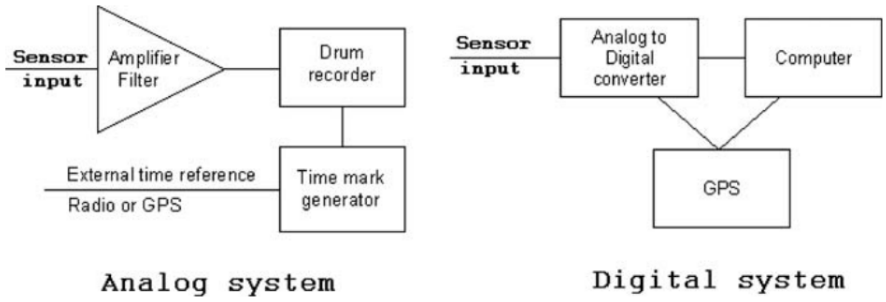


Figure 1.7. Analog versus Digital

Of course, the digital revolution brought also a large effort on developing several tools but luckily the seismologists have made many standard so that everyone can read and view dataset by use of standard program. The most diffused format are SEISAN, GSE2, SEED and SAC. Each format borned in research laboratory so it is simple to find open documentation and software implementation to use them. The SEISAN binary format is used in the seismic analysis program SEISAN. This format proposed by the Group of Scientific Experts (GSE format) has been extensively used by projects on disarmament and is also used in the seismological community. The Standard for the Exchange of Earthquake Data (SEED) format was developed by the Federation of Digital Seismographic Networks (FDSN) and it was adopted as its standard in 1987. Seismic Analysis Code (SAC) is a general-purpose interactive program designed for the study of time sequential signals. A SAC data file contains a single data component recorded at a single seismic station.

All of the formats have an internal structure to contain waveforms and metadata. The waveform is a sequence of all values registered by seismometer while the metadata store the configuration used on recording. The most important metadata are the observed time, the station location and the sampling frequency. Other informations, like phases or event location, may be added on processing of the waveforms. Some data formats contain only one trace while others more.

The use of these data format is very useful for large dataset because each file may be packaged waveforms with the computed data as the hypocenter, the magnitude, p-phase and s-phase. These will be available on large evaluation by techniques of

pattern recognition. Of course, the creation of such dataset requires a great effort in resources and time. Due to these reasons it is not available a free dataset with seismic events analyzed and classified. Only inside of laboratories these dataset are available but not publicly accessible: everyone carefully preserve their investments. This is a great problem for data analysts because the lack of controlled or verified data doesn't provide the right resources to analyze them and infer a model.

To solve the problem of lack of dataset some researchers have developed several simulation tools. The simulation may use many complex model to generate seismograms similar to real recording of digital seismometer. The advantage of these tools is the controlled generation of the waveform: a researcher knows the model before seeing the signal. Then it simple to check the behaviour of the inferred model over the generator of the model. We will use some generated seismograms to test and proof the behaviour of the proposed similarity measure.

Chapter 2

The Process Model

2.1 Problem statement

The care was focused on discrimination of seismic events to make hypothesis on characterization of the waveform by focal mechanism. This contributio analyzes the aspect of similarity measures applied to seismic events. In the previous chapter we saw that many applications require some measure to compare signals of large amount of data. A similarity measure is often required by clustering methods. Which are a field of machine learning used to find some relations among data. This is an unsupervised technique because refers problems for finding hidden structure in unlabeled dataset. An human expert is able to cluster a little number of data through a deep observation of its features but when the data are thousands or millions it is mandatory to automate the process by machine learning techniques. Cluster analysis applied to seismic events helps scientists to do a lot of tasks in less time and in automated way.

In this thesis, we focused on the automatic grouping of seismic signals to infer some properties of the sources that generated them. In order, to perform this task we have used the commons tools of signal processing related to seismic signals.

In order to cluster seismic events we must collect, filter and cut the signals before the application of the clustering algorithms. Each activity of the aboves is very important because the result is used for next step on analysis process. The need to execute them is due to the nature of the signal: often a background noise is present and the event start sometimes is not clear.

The data collection is the activity related to search a set of data useful to execute test and experiments. As we will see below, there are many data types and several sources. Sometimes some geologists are interested on a specific set while others on a different because their studies refers different mechanisms.

In every process related to signal processing a data preprocessing is an essential requirement that include some operations useful to work with data. A coarse data must be prepared before applying every clustering algorithm. The seismic events inside a seismograms must be filtered and detected before to compare them with each other. So old but useful tools of signal processing are used to prepare the dataset for computation.

Indeed the data preprocessing may include some activity as feature selection

and data transformation. These activities are related to the target of the data analysis and they are useful to speed up the learning process. An example of them is the dimension reduction used when a sampled signal can be processed with the same results and an acceptable error than original one. Of course a reduction of the required space to store data or a speed computation must be justified by the application target. Then we have described the phases prior to clustering of the signals.

The clustering can be executed by application of several techniques and similarity measures. Although a similarity measure is not mandatory requirement it is used on the most diffused hierarchical and partitional methods.

The research about the similarity measures is the aim of this thesis. We will see on the next chapters history and results of the most diffused measures and we propose a new measure that run as well as the known cross-correlation. Although there are some clustering techniques that don't use similarity measures the most widely used in the scientific community are these based on algorithms which require them. A research on similarity measure or distance is very important because the result of clustering application is very dependent on it. The measure must be choosen to be applied on fixed type of a signal. It is common to see a measure that with some signal have good results but bad with others. This is because the nature, the amplitude, the frequency are few of a lot elements which require to be evaluated each of them in one way. The literature is full of specialized techniques on a subset of the available data in the world. The most known Euclidean measure is general-purpose and run very well on metric space but enough bad on data where some attributes have to be weighted than others or when the data have an hidden structure which must be compared with others in the right way.

But how are evaluated the results? Are they acceptable? We use the term acceptable because the validation of the results is very difficult and the relative activity in the field of Earth sciences is overloaded by the dimension of the problem and the inaccessibility of the location where events happened. The sources of seismic events source are often located over hundred kilometers of depth, and these sizes are not easy to reach for detecting what are the physics mechanisms of the earthquake source. The dimension is also a problem because a fault segment which contributed in a event may be very long, from hundreds to thousands of meters, and the informations about it are very difficult to retrieve.

The full process model is shown on figure 2.1. Five main step are identified in the analysis process of seismograms.

Figure 2.1. Overview of the process model

Last phase of process regards the validation activity is a last activity to perform but is not less important than others because its aim is to get results from executed experiments. A strong validation on a measure must ensure under some conditions and situations a set of results comparable with other techniques so that everyone can do an objective evaluation of the proposed tools. Of course all the validation

indexes reported on this thesis are used on machine learning and cluster analysis.

Defining a proximity measure involves a validation activity that includes other measures or the state of the art to assess the features of the new. We have choose to compare the new measure with the most diffused measures used by scientific community.

2.2 Data collection

The data collection is the first module of the above process model. As first activity it is oriented on capturing all the necessary resources to run a process of data analysis. As described in the previous chapter it is very difficult to find a free classification by experts of a set of controlled data.

Here, we give an overview on real and simulated data which everyone should use in seismic data analysis. Each type of data can be used to test several aspects of measures and algorithms. As described in [Bor12] the seismic source can be divided into two main groups:

- **natural events**: they are generated by natural factor such as tectonic earthquakes, volcanic earthquakes and storm microseisms.
- **artificial events**: they are generated by man and artificial causes such as explosions, rock bursts or cultural noise [Bor12].

The natural events are stored when occur and the reasearchers study their causes and behaviour. Due to differents sources an event may be localized from few kilometers of depth up to 700 kilometers. Most earthquakes occur along the main plate boundaries. The tectonic earthquakes can be very destructive with a magnitude greater than 6. Volcanic earthquakes have a small energy and duration of tremor type. Some instruments have difficult in recordings this type of event. Miscroseisms are generated by storms over oceans or large water basins. They are not well localized nor fixed to a origin time.

The artificial events such as explosions or rock bursts are generated by human acitivity focusing on scientific aim. These experiments are often a controlled sequence of bursts used to test a detection grid and the transmission medium.

As all the digital signals recorded by an instrument the seismic signals contain noise, too. The noise can be instrumental or real. The instrumental noise is usually less than real but it is depending on the recording station. The noise spectra for a particular station changes over days, months and years due to changes in cultural noise, weather and wear of the tools. For example two very similar events may have two different waveform due to different noise rather than seismic waves. To evaluate the seismic noise it is necessary to record many signals for a long time. Of course it is a great problem in data analysis because each signal may have a different noise which depends on several factors not correlated with itself.

A similarity measure for seismic events have be take into account the noise variations and must be less sensible to them. Two seismic events may be generated by the same source but a different noise changes the waveforms so that their look quite different. The signal filtering is a mandatory activity included in the data preprocessing. Indeed, the filtering must be executed with care because a light filter

does not clear the signal in the right way otherwise an hard filter may remove a meaningful part of the signal.

2.3 Data preprocessing

The data processing phase prepares the data to be analyzed. The seismograms have often a noise and a not well identified start and end of the event. The noise can make complex the use of a similarity measure between signals because a fine grained measure can look the noise as a significant component of the seismic event. The two main activity of the preprocessing are filtering and picking. While the filtering is a required step to clean the signal, the picking has to find, a not trivial work, the event inside a long seismogram. Find an event means finding the p-phase and the tail of the event.

2.3.1 Filtering

As described in [Bor12] by the Fourier theorem any arbitrary transient function $f(t)$ in the time domain can be represented by an equivalent function $F(\omega)$ in the frequency domain, the Fourier transform of $f(t)$. These relations are:

$$f(t) = \int_{-\infty}^{\infty} F(\omega) e^{2\pi i \omega t} d\omega \tag{2.1}$$

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \omega t} dt \tag{2.2}$$

where $|F(\omega)|$ is the amplitude spectral density with the unit m/Hz, $\omega \in \Re$ the angular frequency (with f - frequency in unit Hz).

The filtering is the most well known technique used in signal processing. A filter remove a part of the signal, without information, which stay in a fixed frequency range. The range must be limited, bottom unlimited or top unlimited. We call a filter band pass, low pass or high pass respectively. The filters are shown on figure 2.2.

A band pass filter remove the frequencies outside the fixed range f_1 and f_2. A low pass filter remove the frequencies up to the fixed frequency f_2 while an high pass remove the frequencies down to the fixed frequency f_1.

The filter does not attenuate all frequencies outside the desired frequency range completely but there is a region outside the values f_1 and/or f_2 where frequencies are attenuated, but not rejected. This is known as the filter roll-off and it is expressed in dB of attenuation per octave or decade of frequency. The bandwidth of a bandpass filter is the difference between the upper and lower cutoff frequencies.

We show on figure 2.3 an application of a band pass filter to a seismogram. The original signal is not clearly visible but after the filter application the signal and its phases seem well identified.

The most common anti-alias filter is a linear phase finite impulse response filter (FIR). The most common bandpass filters work on range 3-20Hz but some corrections

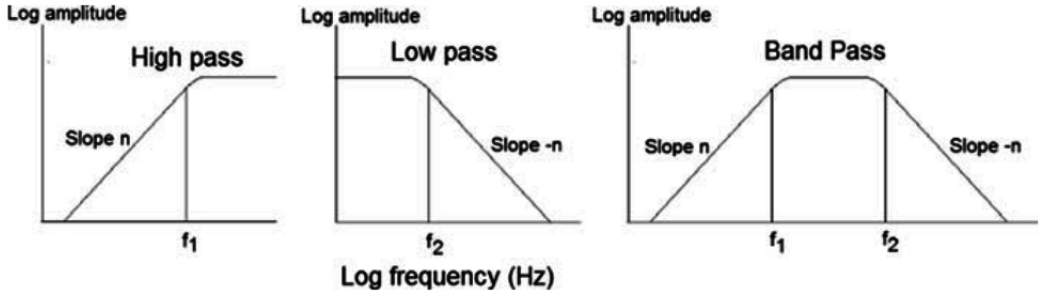


Figure 2.2. Filter types

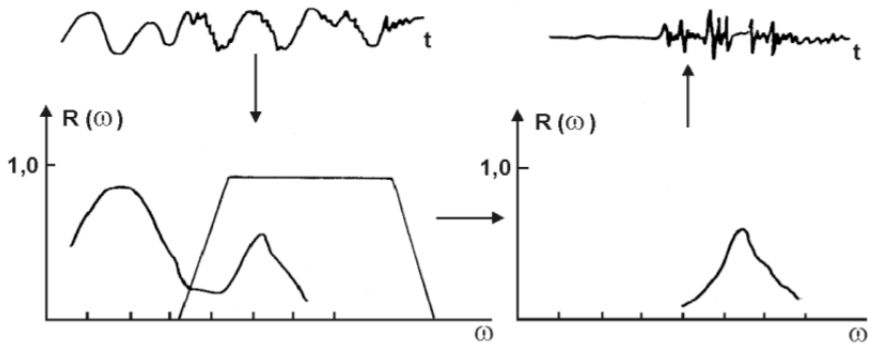


Figure 2.3. Band pass filter applied to a seismogram

depend on data. The values changes are fixed by the nature of data and by the used record stations. Of coarse a simulated event doesn't contain environmental or cultural noise.

2.3.2 Triggering and Picking

The main goal of the trigger algorithms is the automated recognition of the seismic event regardless the noise background while the picking goal is to identify the two main phases p and s. As described in [WAY+98] a great work has be done on triggering by scientists in the past years and some well tested methods are now used in seismograms analysis. The development started very fast with digital acquisition of the seismograms.

The two techniques are applied togheter on seismograms analysis in order to select an event inside a whole signal and to detect the phases on it. Indeed a trigger algorithm usually found also the *p-phase* of the event as start so that a picking algorithm can find others.

We will describe in the following text the most known triggering techniques.

The first one is called STA/LTA (Short Time Average over Long Time Average), a time domain method which evaluates the ratio of short-to-long-term energy density (squared data). This method is very common in computer applications but it requires the setting of the windows size used to evaluate the signal. These parameter may depend on sampling rate but it is a common practice to set these around values of 3 and 24 seconds respectively. After all the ratio between the two values is in range

between 8 and 10.

The implementation of STA/LTA proposed in [WAY+98] has two not overlapping window where the long term follows the short term without any delay. STA and LTA are defined as:

$$STA_i = \frac{x_i^2 - x_{i-N_{sta}}^2}{N_{sta}} + STA_{i-1} \tag{2.3}$$

$$LTA_i = \frac{x_{i-N_{sta}-1}^2 - x_{i-N_{sta}-N_{lta}-1}^2}{N_{lta}} + LTA_{i-1} \tag{2.4}$$

where $x = (x_1, \ldots, x_n)$ is the signal, N_{sta} is the length of the short time windows and N_{lta} is the length of the long time window.

The classic definition of STA/LTA has a great variability on short term windows lenght changes. On figure 2.4 we show an example of STA/LTA application to a long signal.

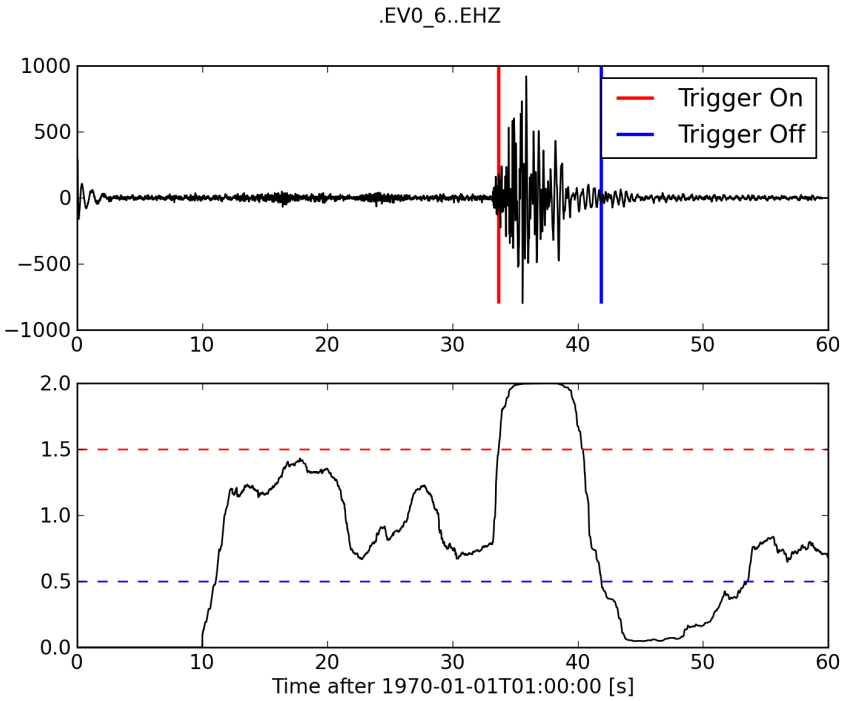


Figure 2.4. Application of the classical triggering algorithm STA/LTA

The result of the computation of the ration STA/LTA is shown on figure 2.4. The lines on figure, one red and one blue, are the triggers which define the start and the stop of the seismic event. With a start threshold fixed to a value less than 1.4, the algorithms would choose a time between 10 and 20 seconds as start, an error. The problem is the noise which have an increment between 10 and 20 seconds. The algorithm doesn't require to fix the thresholds before the run but it is clear that a wrong choice could detect an incorrect start and end of the earthquake.

An alternative to classic STA/LTA is the *recursive STA/LTA* which tries to to have a smaller range of uncertainty for threshold selection. The recursive method

is more suitable for long signal and to avoid to keep a long data vector in memory. The recursive STA/LTA is similar to the standard STA/LTA except that for each successive time step, a fraction of the average data value, rather than a specific data point value is removed. We show the recursive STA/LTA ratio of the previous signal on figure 2.5.

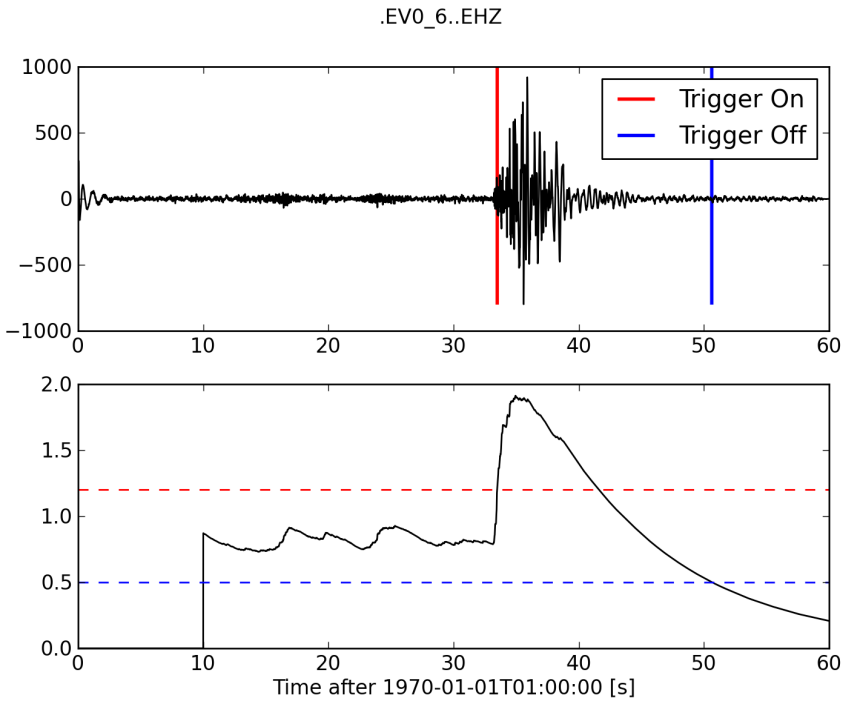


Figure 2.5. Application of the recursive triggering algorithm STA/LTA

Another well known triggering algorithm is the *Z-detector* introduced in [Ste77]. This method use a variable Z computed on the original signal as the standardized variable defined as:

$$Z(x_i) = \frac{x_i - \mu}{\sigma} \tag{2.5}$$

The algorithm is well tested on its variant which uses the STA as x. A great advantage of use Z-detector is a good behaviour in background noise's presence. Of course the thresholds levels, required to select the start and the end of the event, depends from the background noise. The application of *Z-detector* is shown on figure 2.6.

On it the signal is the same of the previous but the detection is quite different. In a comparison with STA/LTA, the Z-detector sometimes performs better because it is less sensible to noise. With the latter, the choose of the threshold is more simple than other. Indeed the start is identified with less precision and with a little delay but during experiments may be possible to alter the start threshold to find the begin of the event.

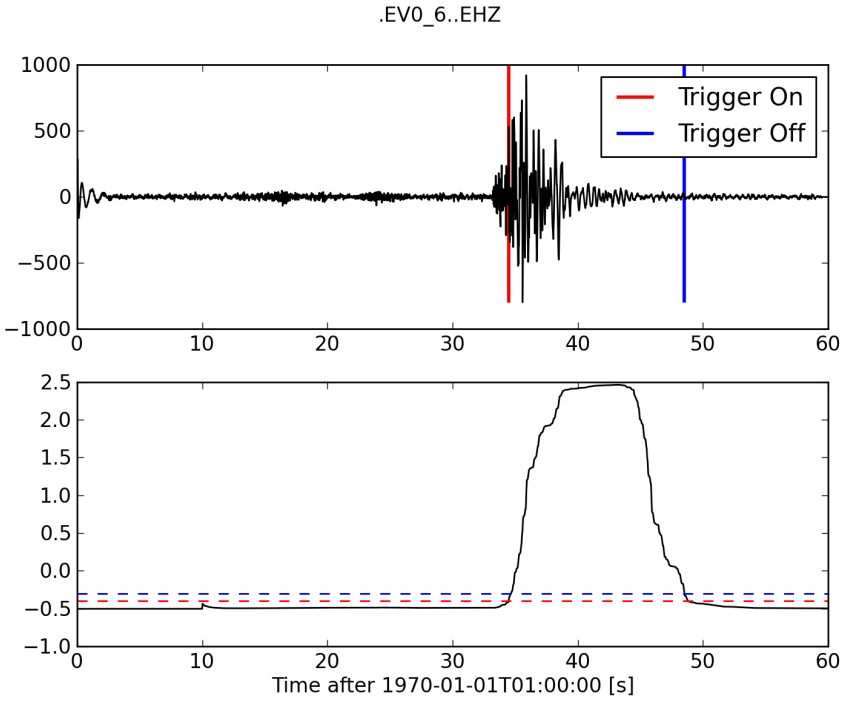


Figure 2.6. Application of the triggering algorithm Z-detector

So far we have described some triggering algorithms of time-domain energy. These algorithms use the energy of the signal in some way to detect variations of the power. A great change of the energy may be the start of an event or one of its phase.

Another class of triggering algorithms is based on frequency-domain. While time-domain algorithms find changes of amplitude on time, the latters find changes on spectral content. Also frequency-domain methods use often two windows like STA/LTA to detect change on frequency which may locate an event inside a seismogram.

Finally, we can describe the picking and their most important implemented tecniques. While triggering tecniques find an event inside a seismogram, the picking tecniques try to find the phases of the detected event. In some way the methods are overlapping because a start time indentified by a triggering algorithm may be a p-phase identified by a picking algorithm. Indeed the picking's goal is to identify at least the two main phases p and s.

The picking algorithms can be divided into three main groups [BK00]: energy analysis, polarization analysis, and autoregressive representation of the trace. However any particular method will fail when the difference between the noise and the signal is small, particularly when the signal-to-noise ratio[1] is low. The energy

[1] the signal-to-noise ratio, abbreviated as SNR, is a number which denotes how much the signal contains information over the not significant part of the signal (noise). This number may have very different values and it is often measured in dB

approach, which is commonly the most useful, it is also the more sensitive when the noise level is high and bandpass filtering does not work when the noise and the signal have almost the same frequency content.

As described on previous chapter the phases can be used to locate an event: the estimation of the source with a single station requires in order to be able to pick, at least four phases from two component of the same seismogram (two P and two S arrivals), and to be able to detect S-wave and P-wave phases among the detected arrivals. By S-wave detection and an azimuth measurement, it is make possible to estimate epicentral using travel-time information.

Automatic procedures for the detection and processing of seismic events are required to process large datasets and to analyze them in real time. In their article [VSL12], Vssallo, Satriano and Lomax did a good overview on picking algorithms and their improvements. We will describe two methods which use a single component signal although the newest tecniques use 3-component signal stored by 3D station. Of course, a picking problem that works on single component z is harder than one which works on 3-component. It is more difficult to detect both p-wave and s-wave on single vertical component while on horizontal component the s-wave are well defined.

The group of energy based methods includes one of the most used algorithms created by R.V. Allen [All78]. This algorithm works with two windows like the STA/LTA triggering algorithm. These two windows are used to compute a characteristic function based on combination of two elements: the signal and its derivative. In this way the method is enabled to detect some variations on amplitude and on frequency. When the STA/LTA ratio exceeds a threshold, the position will be investigated to check if it is a "true" trigger or something realted to the seismic noise. The trigger is also accepted if and only if some constraints like duration, amplitude, number of zero of crossing event and end of event are verified.

Another picking algorithm based on energy in described on [KPV07]. This algorithm divides the detection process into two steps: p-picking based on signal energy and s-picking based on frequency energy. It differs from STA/LTA algorithm in windows length which have both the same size. For each point is computed a function $f(t)$ as the ratio between the signal energy after and before the time t:

$$f(t) = \frac{\sum_{k=t}^{t+W-1} z^2(k)}{\sum_{k=t-W}^{t-1} z^2(k)} \tag{2.6}$$

The $f(t)$ is computed over the all signal. The maximum of the function is the candidate to be the p-phase arrival. The S-phase arrival computation is based on evaluation of both the frequency-domain and energy-domain: the frequency of the s-waves is higher than p-waves and also the energy has a sensible increment. By the combination of these properties it is possible to detect in reliable way the s-phase arrival.

Others methods are based on neural networks and was proposed in many versions. Some techniques use single-component signals while others use the full three compo-

nents. Of coarse the architectures of the proposed networks have many differences in number of layers and in computation of neuron weights.

An example of previous methods is described in [DM97] which uses a back-propagation network (BPNN) working on single component signals. The authors work an automatic picking tool which detects over 83% for p arrivals and over 75% for s arrivals. The BPNN is trained by several p arrivals with background noise.

So far, we have described several triggering and picking techniques used in seismology. Many tools for seismogram analysis implement the well known *STA/LTA* or *Z-detector* as triggering algorithms while picking algorithms implementation are less diffused.

2.4 Measures

In this section we give some notes about measures for waveforms and their application on clustering. The details are explained on the next chapter where we'll introduce a new dissimilarity measure between two seismic signal. The measure is developed in a specific scientific context with focus on evaluating the shape of the seismograms.

A seismic signal is stored like one or multiple waveforms. Which is stored like a digital sequence of intensity of released energy. Unsupervised learning techniques, like clustering, usually require a distance to compare two element of the dataset. Not all cluster methods require a dissimilarity measure but in every way some function is computed to create a configuration with many groups. The requirement depends on the selected algorithm used to analyze data. Hierarchical or partitional methods require a proximity measure to compute the distance matrix and to give the solution from that.

While cluster algorithms are generic and can be applied to an huge number of problems, a selected measure can be designed to work on a specific data type: a well identified problem can be resolved only with specific solution. We show on the next chapter that the design of a specific solution could be simple if exists a characterization of the data to analyze.

Although some measures come from signal processing and statistics, it is mandatory to do some consideration about waveforms and their structure from seismology point of view. For this reason we have dedicated the whole next chapter on similarity and dissimilarity measures.

After all, the most used measures for seismic waveforms are euclidean distance, dynamic time warping and cross-correlation. In the seismology context we can see also some work about similarity with wavelets or neural networks but scientists use almost always classical mean like the cross-correlation. We have focused our effort on shape based discrimination of the seismic waveforms because we think that the shape is the first element used by human experts in classification work. So a good measure should exploits this knowledge to do a coarse grained classification which can be analyzed for further conclusions.

2.5 Clustering

The clustering is an unsupervised learning technique used on pattern recognition for partition a large amount of data. The data are grouped by specific criterions which sometimes resemble the human activity. As described in [TK08] a clustering task is also a sequence of activities:

Feature Selection is the activity related to selection of the minimum information needed to compute the task. A dataset is very often a composition of many several data so an analyst needs to select the information useful to group them.

Proximity measure as described on the previous section, the *similarity* or *dissimilarity* is computed between data vectors and should be ensured that all the feature selected are evaluated with the same weight or in some way that doesn't hide one of them.

Clustering criterion is usually expressed by a cost function or other type of rules. Different criterions can lead to different cluster solutions. Of course also the choice of the criterion must be weighted respect to nature of the data and the obtained solution,too.

Clustering algorithms the algorithms combine the proximity measure and the criterions to compute a solution over the given dataset.

Validation of the results the solution obtained must be validated by an expert or by some functions useful to check the correctness of the computed results. Moreover the choice of the validation tests requires a bit of work because a cluster solution may be correct for a test but not for others. For this reason the results of the validation tests are interpreted to understand the nature of the data. Indeed, a first requirement of a cluster solution is an high internal homogeneity and a high external heterogeneity. These two aspects drive to a solution quite stable where groups are clearly separated from each others.

Interpretation of the results it gives meaning to the previous work. The clustering is usually computed to infer some characterization of the original data. The characterization is used to find some relations not immediately visible among data. Unluckily any hypothesis requires a domain expert and a deep knowledge of the analyzed data to infer some discrimination property.

A number of cluster applications can be found in literature and they lead in many science fields. Some successful applications exist on marketing where cluster analysis was applied to understand the customer behavior. The following are common applications of the cluster analysis:

- *data reduction*: some analysis on large amount of data requires a lot of time and very often it is too long to retrieve needed results. The aim of data reduction aim is to reduce the size of the dataset so that the analysis of the remaining data is done within a reasonable time. Through the clustering a large set of element belonging to the same group can be represented by few

prototypes of them. If the original size is N while the new size is m should be $m \ll N$. This is sometimes a preprocessing activity on supervised learning where a model is created on training set which must have dimensions with computable training time. At a high level point of view, the data reduction can be see as data compression. The term *data simplification* is also used to describe this application because all of the observations can be viewed as members of clusters and profiled by their general characteristics.

- *hypothesis generation*: this use is related to research of some unknown hypothesis about the nature of the data. The cluster solution can suggest relationships among the elements of a single group supported by some proximity measure computed between all pairs. This aspect, called *relationships identification*, is part of researches involving the study of the features and the observations of the dataset. More investigations are required to understand one or more features which group data. On hypothesis generation an aspect of clustering with great importance is the *taxonomy description* that may be obtained from the element structures extrapolated from results.

- *hypothesis testing*: is the complementary application to the previous. The testing starts with fixed hypothesis by scientists and the clustering is the mean to proof them. It's a typical task with a deep knowledge of the source's parameters.

- *prediction based on groups*: this application of clustering seems a supervised learning technique because the cluster solution is used like a model to infer the membership to a single group. Of course, such type of application is useful when some hypothesis on cluster groups are defined and verified so that the characteristics of a new element are deducted only by the assignment to a single group.

The dataset features can be of two main types: continuous or discrete. Indeed, there are also categorical and nominal types which are usually codified in some way into discrete data. The convergence to discrete or continuous data is needed to compute proximity measures over the samples of the dataset. Of course, each data transformation must be weighted in some way to avoid every excess or defect of the contribution of each feature.

So far we have given a clustering background, now we can describe the clustering in a more formal way giving several mathematical definitions. Let X be the dataset:

$$X = \{x_1, x_2, x_3, \ldots, x_N\} \tag{2.7}$$

A clustering of X, in $\Re$, of size m is the partition of X into m subsets (clusters) $\{C_1, C_2, C_3, \ldots, C_m\}$ so that the following conditions are met:

- $C_i \neq \emptyset, i = 1, \ldots, m$
- $\bigcup_{i=1}^{m} C_i = X$
- $C_i \bigcap C_j = \emptyset, i \neq j, i, j = 1, \ldots, m$

The above definitions fix, in some natural way, that an element belongs only to one cluster. This is true for almost applications but not for all. When the previous conditions are satisfied we say *hard* clustering. But some applications require to know how an element is bound to its cluster and to others. For example a lion and a mountain lion that belong to the same cluster are closer to tiger cluster than dogs so the three cluster are different but tigers and lions have some common characteristics, both are big cats, while dogs not.

A different definition of the clustering in terms of the fuzzy set is:

$$u_j : X \to [0,1], \quad j = 1, \ldots, m \tag{2.8}$$

and

$$\sum_{j=1}^{m} u_j(x_i) = 1, \quad i = 1, \ldots, N, \qquad 0 < \sum_{i=1}^{N} u_j(x_i) < N, \quad j = 1, \ldots, m \tag{2.9}$$

The equations 2.9 are the so called *membership functions* which define how a single cluster is mathematical characterized. An element $x \in X$ may belongs to more than one cluster in the same solution with different values of degree to single cluster. The degree value u_i is in interval $[0, 1]$. Stronger is the membership to a cluster then more close to 1 is the value of u_i.

Of course, the *fuzzy clustering* is used when the clusters are not well defined and each member may belongs to one or more. In this thesis we'll not use fuzzy clustering because the aim of this work is to test that the clusters are in some way a characterization of several physical phenomena. We are in a situation like a *hypothesis testing*: we think that different dynamic aspects of seismic events lead to several cluster where each of them is characterized by a fixed phenomena.

There are many categories of clustering algorithms but we choose to select only the two most diffused and known by seismology community. The main reason of this decision is that few algorithms are implemented on seismology software and some graphic tool is used to infer some characteristics of the data. In the following we'll describe the hierarchical and partitional algorithms.

A simple example of clustering on two dimension is shown on 2.7 and 2.8. In the first a dataset includes a large number of points in a single group while in the second there are four groups identified after a clustering application.

2.5.1 Hierarchical Algorithms

It is the most known cluster algorithm, it is simple to implement and to understand also for naive users. A great advantage of these algorithms is that the cluster solution is composed by a tree structure called *dendrogram*. An horizontal line over the tree cuts into two parts the dendrogram and the intersection over the line define the number of the clusters. So if you want a solution with K cluster you have to find a cut with K intersections. Indeed, the previous steps are useful to test different cluster solutions and to understand the overall structure of the data.

Hierarchical techniques are subdivided into *agglomerative* and *divisive* methods. In some way the behavior of the first is the inverse of the second: agglomerative join one subgroup into group while divisive divide group into subgroups. In this way the

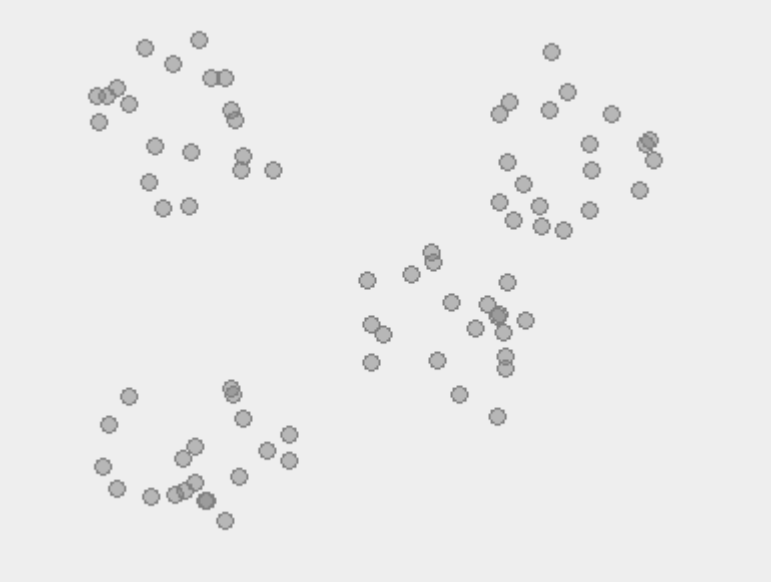


Figure 2.7. A dataset of elements before the cluster application

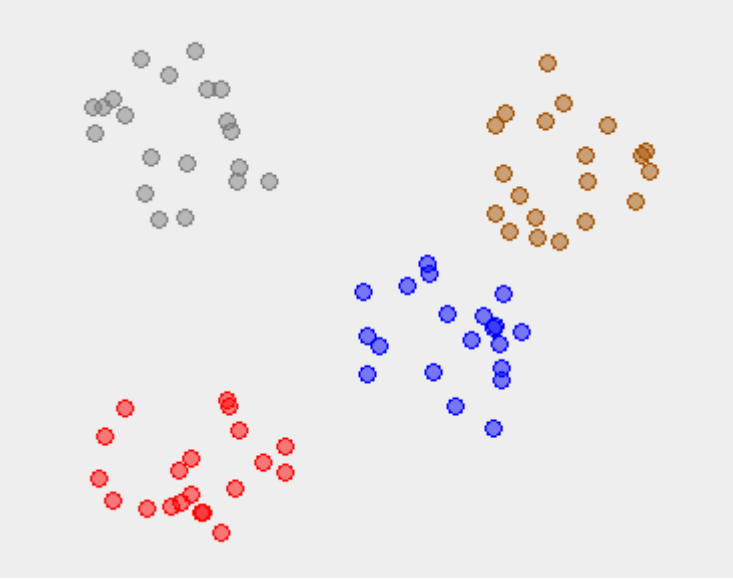


Figure 2.8. A dataset of elements after the cluster application with a partitional method

agglomerative algorithms starts from single point and end with one big cluster while divisive conversely.

A schema proposed on figure 2.9 in [ELLS11] show a simple tree that can be traversed into two directions: agglomerative and divisive. With agglomerative, two joined groups in the previous step cannot be divided in the following while in divisive two split groups cannot be joined. This aspect is in some way a disadvantage of the hierarchical techniques because every preceding error in previous steps cannot be repaired in the following.

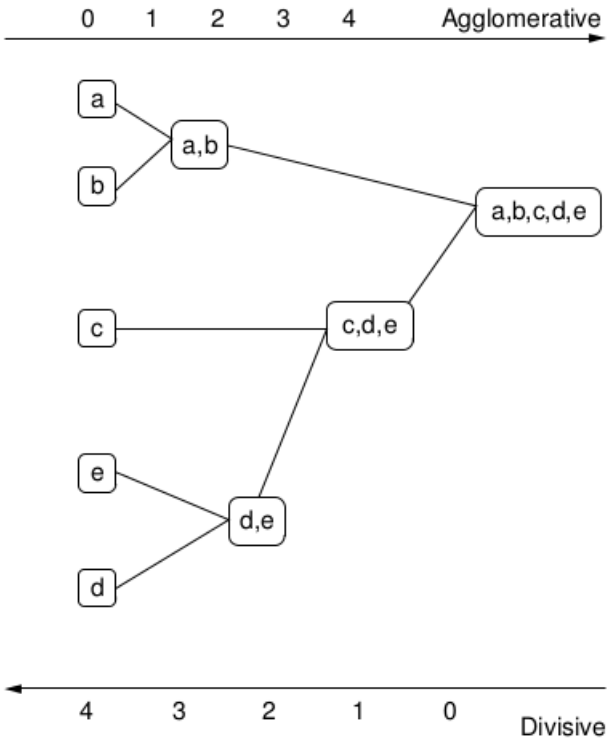


Figure 2.9. A dataset of elements after the cluster application with a partitional method

The dendrogram is the tool used in hierarchical clustering to define a cluster configuration. In figure 2.10 we show an example of this diagram computed on a sample dataset.

So, let the dataset $X = \{x_i,\ i = 1, \ldots, N\}$ a set of l-dimensional vectors, a cluster configuration is

$$\Re = \{C_j,\ j = 1, \ldots, m\} \quad with\ \ C_j \subseteq X \tag{2.10}$$

A clustering configuration $\Re_1$ is said to be *nested* (see [TK08]) in the clustering $\Re_2$ if:

- $|\Re_2| < |\Re_1|$
- $\forall C_j \in \Re_1, \quad C_i \subseteq C_d \in \Re_2$

The result of hierarchical clustering is hierarchy of the subset represented as structure tree or a dendrogram.

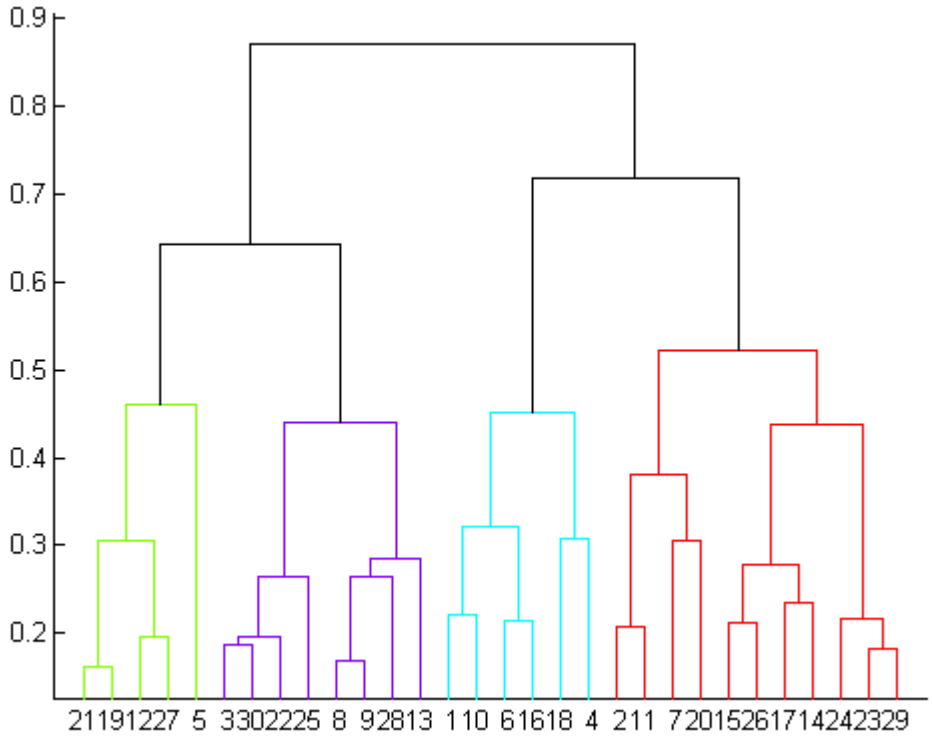


Figure 2.10. A dataset of elements after the cluster application with a hierarchical method

The most diffused techniques are agglomerative methods. These methods differ one for other on linkage criterion type. The general schema algorithm is:

- Place each item into its own group: one group for each item
- Repeat: iteratively merge the two closest groups
- Until: all the data are merged into a single cluster

The *linkage criterion* is how the distance between subsets is measured to evaluate the join or not. The discussion about distance measures between elements is exposed in the next chapter. There are several methods to measure the distance between subsets:

single linkage (also called nearest-neighbor method)
the distance between subsets is compute as the distance between the closest pair

$$D(S_1, S_2) = \min_{i \in S_1, j \in S_2} d_{i,j} \tag{2.11}$$

complete linkage (also called farthest-neighbor method)
the distance between subsets is compute as the distance between the furthest pair

$$D(S_1, S_2) = \max_{i \in S_1, j \in S_2} d_{i,j} \tag{2.12}$$

average linkage (also called group average or UPGMA, which stands for "unweighted pair group method using arithmetic averages")
the distance between subsets is compute as the average distance among all pairs

$$D(S_1, S_2) = \frac{1}{|S_1||S_2|} \sum_{i \in S_1} \sum_{j \in S_2} d_{i,j} \tag{2.13}$$

where d is a chosen metric. Figure 2.11 is shown how the group distances are computed for each method:

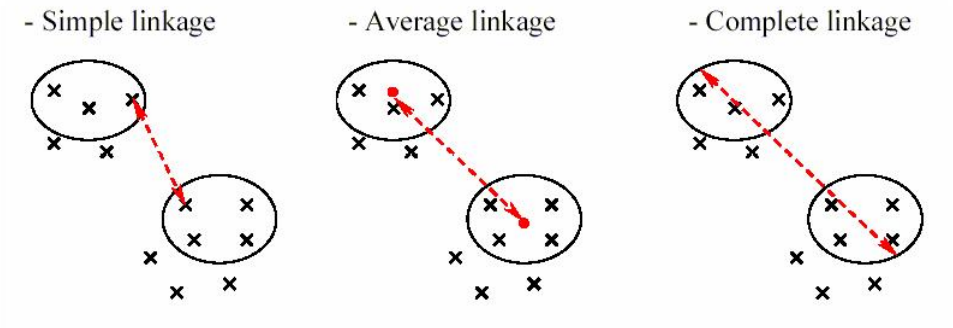


Figure 2.11. Agglomerative Hirarchical Clustering Linkage Criteria

From previous definitions we can summarize that *single linkage* can be affected by chaining which is cause of the early merges between subgroups, while *complete linkage* has the opposite problem with close subgroups which are merged late if some outlier is present. The latter *average linkage* is in practice a good compromise.

The average linkage is the algorithm the we chose to run our experiments. In detail, the group average method, compute the distance between two groups as the average of the distances between all possible pairs of data points that are made up of one data point from each group.

Of course each techniques has advantages and disadvantages. One problem of the hierarchical algorithms is that they impose a hierarchical structure that sometimes is not real compared to other methods or over well known solution.

Another problem is the choice of the correct number of clusters. As proposed in [TK08] an intuitive approach is to find inside the dendrogram the clusters with a large lifetime. These clusters are candidates to be a real configuration of the dataset. In [ELLS11] is reported a long list of indexes used to choose the correct number of clusters for both hierarchical and for partitional.

Nevertheless the hierarchical methods are diffused and the results can easily be interpreted.

2.5.2 Partitional Algorithms

The partitional clustering techniques create a flat configuration, a partitioning, of the data with a desired number of clusters K. The basic idea is to find a clustering structure that minimizes a certain error criterion which measures the "distance" of each instance to its representative value. The most well-known criterion is the Sum

of Squared Error (SSE). SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0. SSE may be globally optimized by exhaustively enumerating all partitions, which is very time-consuming, or by giving an approximate solution (not necessarily leading to a global minimum) using heuristics. The simplest and most well known algorithms are two based on the idea that a center point can represent a cluster and such center can be computed in several ways:

k-means the algorithm works by partitioning the data into a fixed K number of clusters and then iteratively reassigning elements to cluster until some criteria are met. These criteria are usually related on minimizing the distances among the elements of each cluster and maximizing the distance between clusters. For K-means the center point is a *centroid*, which is the mean or median point that almost never corresponds to real data point of a group of points.

k-medoids it is like the *k-means* but with the difference that the *centroid* is not a computed element but always a representative observation belonging to the cluster. In this case we denote the centroid with term *medoid*.

Before the detailed description of the algorithms cited above we show two simple figure examples which denote the difference between the two cluster centroid. On figure 2.12 we have a k-means application with an artificial centroid while on figure 2.13 we have k-medoids application with real observations that are centroids.

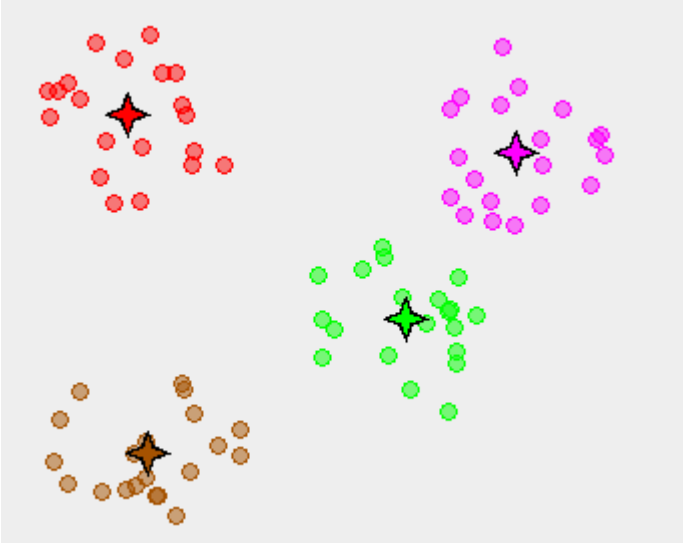


Figure 2.12. A dataset of elements after the cluster application with k-means

We can see that the centroids computed over the sample dataset are close but not the same. Of course the figures shown a simple example with little differences but in many real applications the centroids of one algorithm may be far from others.

The *k-means* algorithm belongs to hard clustering algorithmic techniques. The k-means method starts with k clusters and allows each individual to be moved from its current cluster to another cluster. Individuals are moved between clusters until

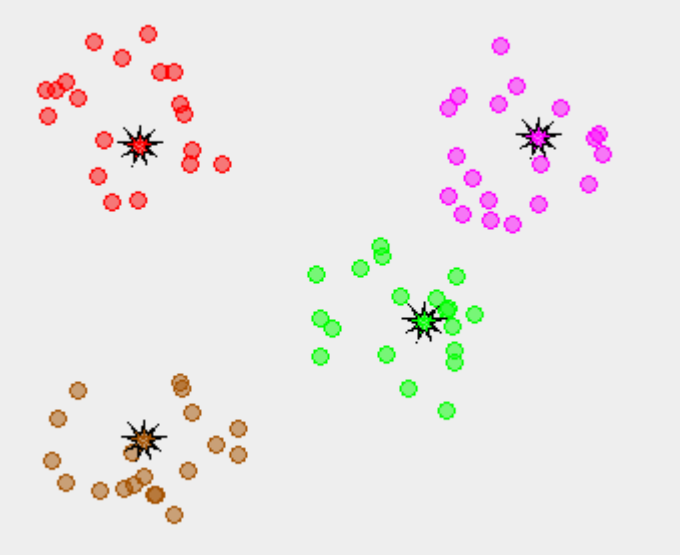


Figure 2.13. A dataset of elements after the cluster application with k-medoids

it becomes impossible to improve the measure of clustering. There is no guarantee that the global optimum will be achieved. The algorithm scheme is:

Input: S (instance set), K (number of cluster)
Initialize K cluster centers.
while termination condition is not satisfied **do**
    Assign instances to the closest cluster center
    Update cluster centers based on the assignment
end while
Output: clusters

One problem, that k-means algorithm starts with the selection of the initial partition. The algorithm is very sensitive to this selection, which may make the difference between global and local minimum. In addition, this algorithm is sensitive to noisy data and outliers because a cluster center is computed as mean over all point. The choice of the number of clusters in advance is not trivial when no prior knowledge is available.

An implementation of the well known *k-medoids* which attempts to minimize the SSE is the PAM introduced by [KR87]. This algorithm is very similar to the K-means algorithm. It differs from the latter in its selection of the different clusters. Each cluster is represented by the most centric object in the cluster, rather than by the computed mean that may not belong to the cluster. For this reason the K-medoids method is more robust than the K-means algorithm in the presence of outliers because a medoid. However, its processing is more expensive than the K-means method because the selection of the cluster centroid require a comparison between all pairs of the cluster. Another problem, that k-medoids resolves is the synthesizing of the centroid for complex object: sometimes the created centroid

as mean of all elements may be incorrect if the result is a new element with not conforming characteristics of the others.

From nature of the data can be suitable an algorithm rather than another. In the case of seismic events, a seismogram is a waveform with special features defined by seismological aspects so the artificial generation of a new waveform as cluster centroid is unsuitable. The risk is to have a new signal that miss the main characteristics of the seismic events. So the *k-medoids* is more suitable than *k-means*. In our experiments we used the *k-medoids* as select algorithm in the class of the partitioning algorithms to group data.

The most used k-medoids algorithms are PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications), and CLARANS (Clustering Large Applications based on RANdomized Search). The implementation of *k-medoids* used in this work is based on [KR87]. In the following the PAM algorithm:

Let Θ be the set of medoids for all clusters and I_Θ the set of indexes of the point in X. The quality of the clustering is defined by a cost function:

$$J(\Theta, U) = \sum_{i \in I_{X-\Theta}} \sum_{j \in I_\Theta} u_{ij} d(x_i, x_j) \tag{2.14}$$

where

$$u_{ij} = \begin{cases} 1, & \text{if } d(x_i, x_j) = min_{q \in I_\Theta} d(x_i, x_j) \\ 0, & \text{otherwise} \end{cases} \qquad i = 1, \ldots, N \tag{2.15}$$

The final set of medoids are obtained from minimization of the 2.14. The PAM algorithm focus the computation of the best solution on minimizing the cost from a step to other. Let $\Delta J_{ij} = J(\Theta_{ij}, U_{ij}) - J(\Theta, U)$ the difference of the cost obtained by the replacement of medoid $x_i, i \in I_\Theta$ with element $x_j, j \in I_{X-\Theta}$. PAM starts with a set m medoids, which are randomly selected out of X and at each step try a new element r as medoid. The number of the steps is uqual to $m(N-m)$ where m is the number of the initial medoids and N is the number of all elements in the dataset. If for some medoid q we have ΔJ_{ij} negative, then the new set of medoids will include j but exclude i. If for all pair $\Delta J_{ij} \geq 0$ the algoritm has reached a local minimum and terminates. PAM becomes inefficient for large data sets because its time complexity per iteration increases quadratically with respect to N.

2.5.3 Proximity matrix

The basic data for most applications of cluster analysis is the usual $m \times n$ multivariate data matrix, X, containing the variable values describing each object to be clustered:

$$X_{m,n} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} \tag{2.16}$$

where n is the number of the element in the dataset ad m is the length of each vector. Of course this is a column-vector representation but sometimes also row-vector is used. It depends on convention applied in the analysis process. In this case the entry x_{ij} in X gives the value of the i component of the ith element.

The described algorithms have in their computation an element in common: the *similarity (dissimilarity) matrix*. Its is a matrix $N \times N$, where N is the number of elements in the dataset X. Each element (i, j) of the matrix contains the similarity $s(i, j)$ (or dissimilarity $d(i, j)$) between vectors x_i and x_i in X. The matrix is also referred as *proximity matrix* as generalization of both cases.

The proximity matrix P in general is a symmetric matrix but if P is a similarity matrix, its diagonal elements are equal to the maximum value of similarity s otherwise if P is a dissimilarity matrix, its diagonal elements are equal to the minimum value of distance d (usually equal to zero). After all the proximity matrix values depend on the selected proximity function.

But, the question is how the proximity matrix is related to described cluster algorithms? The proximity matrix is a common element on the cluster algorithms because these methods use intensively the proximity function to execute a single step of the computation. So, a precomputed proximity matrix can speed up the execution time avoiding repeated and redundant computation of the measure between all pairs.

All the details related to proximity measure in cluster analysis applied to seismology are shown into the next chapter.

2.6 Validation

The cluster validation is on field of research with a lot of techniques and results. Many scientists have spent for a long time their resources on the definition of strong tool which aim is to asses a cluster configuration.

After all, we must remember as described in [HBV02] that clustering is an unsupervised technique and clustering algorithms behave differently depending on the features of the data and the initial assumptions for defining groups. As unsupervised process, there are no predefined classes and no examples that can show that the clusters found by the clustering algorithms are valid.

On clustering validation many aspects can be evaluated:

- when a solution is unknown it is important determining the clustering tendency of a set of data, distinguishing whether non-random structure actually exists in the data (for example hierarchical clustering is very useful on this task because the dendrogram visualize how data are joined to form a cluster)
- when some characteristics of the true solution are available then comparing the results of a cluster analysis to externally known results
- comparing the results of two different sets of cluster analyses to determine which is better
- determining the ”correct” number of clusters.

The validation technique try to solve the hard problem of solution validation without any prior knowledge. The cluster validation criteria can be of three types:

internal
: evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix).

external
evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set.

relative
evaluate the clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values.

In [BL97] are described two well known relative criteria for cluster assessment:

compacteness the members of each cluster should be as close to each other as possible

separation the clusters themselves should be widely spaced

In order to evaluate the performance of a dissimilarity, we have adopted three different indices. Two of them are related to the partitioning inducted by a clustering algorithm which make use of the dissimilarity, while the other one does not consider any partitioning information.

When using a dissimilarity measure in conjunction with a clustering algorithm, it is possible to evaluate its performance by means of *clustering internal and external indices*: the former gives a reliable indication of how well a partitioning solution captures the inherent separation of the data into clusters [SS01], the latter measures how well a clustering solution agrees with the *gold solution* for a given data set.

A gold solutions for a dataset is a partition based on external knowledge of the data in classes, that can be also inferred by the use of internal knowledge via data analysis tools such as clustering algorithms. When the gold solution is not known, the internal criteria must give a reliable indication of how well a partitioning solution, and indirectly the used dissimilarity, captures the inherent separation of the data into clusters.

Let X a set of generic items $X = \{x_1, \ldots, x_N\}$, and $\mathcal{P} = \{p_1, \cdots, p_t\}$ a partitioning of X.

In our experiment we have adopted the `Homogeneity (H)` and `Separation (S)` as internal indices [SS01] of a partitioning $\mathcal{P}$ produced by a clustering algorithm with a dissimilarity δ, whose formulas are here reported:

$$H = \frac{1}{|X|} \sum_{i=1}^{t} \sum_{x \in p_i} 1 - \delta(x, \mu_i) \tag{2.17}$$

$$S = \frac{1}{\sum_{i \neq j} |p_i||p_j|} \sum_{i \neq j} |p_i||p_j| \delta(\mu_i, \mu_j) \tag{2.18}$$

where μ_i represent the centroid of a cluster p_i.

Note that both of the indices have to be considered: if $\forall x, y \; 0 \leq \delta(x, y) \leq 1$, they assume value in $[0, 1]$ and, the closer H and S are to 1, the better the partitioning of the data, and consequently the used dissimilarity.

When the gold solution is known, the external indices can be computed. Giving the partitioning $\mathcal{C} = \{c_1, \cdots, c_r\}$ corresponding to the gold solution for the dataset,

an external index measures the level of agreement between $\mathcal{C}$ and $\mathcal{P}$. For our experiment we have used the `Adjusted Rand index` as described in [Ran71] and [HA85]

$$R_A = \frac{\sum_{i,j}\binom{T_{ij}}{2} - \frac{[\sum_i \binom{T_{i\cdot}}{2}\sum_j \binom{T_{\cdot j}}{2}]}{\binom{N}{2}}}{\frac{1}{2}[\sum_i \binom{T_{i\cdot}}{2} + \sum_j \binom{T_{\cdot j}}{2}] - \frac{[\sum_i \binom{T_{i\cdot}}{2}\sum_j \binom{T_{\cdot j}}{2}]}{\binom{N}{2}}} \tag{2.19}$$

where $T_{i\cdot} = |c_i|$ and $T_{\cdot j} = |p_j|$. Also in this case, the closer R_A is to 1, the better the partitioning of the data, and consequently the used dissimilarity.

Besides the assessment of a dissimilarity function by making use of clustering validation indices, for this purpose it is also possible to use other a priori information.

In the following, we will define a new index, called `Dissimilarity Optimality index` which make use of the sort of data items. This index was defined to validate a special dataset used in this thesis. This dataset, described in detail in the following chapters, is a sequence of simulated seismic events where one differs a bit from the previous. So we have a sequence of events that in a little range has very similar events.

Let us assume now that X is a partially ordered set of generic items, whose sorting permutation $P = (i_1, i_2, \ldots, i_N)$ is known. In this case, the goodness of a generic dissimilarity δ on X can be established by comparing the sorting it induces on X with the sorting permutation P. In particular, what we expect from a good dissimilarity δ is that for each item x_i, its closest item with respect to δ is x_{i+k} with a small $|k| \geq 1$. The Dissimilarity Optimality index is so defined:

$$do = \sum_{i=1}^{n} \frac{|i - j - 1|}{N - 2} \text{ with } j = \underset{1 \leq k \leq N, k \neq i}{\operatorname{argmin}} \delta(x_i, x_k) \tag{2.20}$$

$do \approx 0$ is what we expect in case of good dissimilarity measure.

Chapter 3

Waveform measures

3.1 Proximity, similarity and dissimilarity

On cluster analysis, the common terms used for measures are: proximity, similarity and dissimilarity. Proximity is a general term used for both similarity and dissimilarity without an effective distinction. The similarity is the expression on how two items have common characteristics while the dissimilarity how two items differ among them. Many algorithms require a proximity matrix to group data and such matrix is constructed with a similarity or a dissimilarity measure.

A similarity coefficient is the value of the relationship between two elements computed by a similarity function $s(x, y)$. That coefficient assume usually values between 0 and 1: 1 for equal relation and 0 for uncorrelated. Otherwise a dissimilarity function $\delta(x, y)$ is the value of the difference between two elements. It is simple to convert a similarity $s(x, y)$ into a dissimilarity $\delta(x, y)$ by taking $d(x, y) = 1 - s(x, y)$.

A dissimilarity measure δ is called a *distance measure* if it fulfils the triangular inequality property. The properties which usually are satisfied by a dissimilarity measure are:

- nonnegativity: $\delta(x, y) \geq 0$
- reflexivity: $\delta(x, y) = 0 \iff x = y$
- commutativity: $\delta(x, y) = \delta(y, x)$

and if it is a distance

- triangle inequality: $\delta(x, y) \leq \delta(x, z) + \delta(y, z)$

As defined in [ELLS11] an $n \times n$ matrix Δ of dissimilarities, with element $\delta_{i,j}$, where $\delta_{i,i} = 0$ for all i, is a *metric*, if the triangular inequality property is satisfied for all triples (i, j, m).

Given a dataset $X = x_1, x_2, \ldots, x_n$ a distance matrix M_{dist} is defined as:

$$M_{dist}(D) = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & 0 & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & \cdots & 0 \end{pmatrix} \tag{3.1}$$

where $d_{i,j} = d(x_i, x_j)$ with $d(x, y)$ a distance measure. Otherwise a similarity matrix M_{sim} is defined as

$$M_{sim}(D) = \begin{pmatrix} 1 & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & 1 & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & 1 \end{pmatrix} \tag{3.2}$$

where $s_{i,j} = s(x_i, x_j)$ with $s(x, y)$ a similarity measure.

Of course, when a similarity or distance is symmetric then the matrix is symmetric and only half of the element needs to be computed.

Author in [Cor07] reports an overview of dissimilarity measure for data mining. The author proposes a classification into three main types: shape dissimilarity, dissimilarity by feature extraction and structural dissimilarity. Euclidean distance is the most simple shape-based dissimilarity. Also cross-correlation based measures are shape based. Dynamic Time Warping is another example of shape-based dissimilarity with some interesting characteristics related to distortion in time axis: DTW allows non-linear alignments between time series. On time series the frequency domain measures belong to group of dissimilarities by feature extraction. The two most diffused techniques are based on well known Discrete Time Fourier and Discrete Wavelet Transform. To the third group of the structural dissimilarities belong some techniques based on a representation of time series as recorded geometric trajectory.

In the next sections we'll describe some common used similarity and distance measures and we'll define a new dissimilarity measure devoted to seismic waveforms.

3.2 Euclidean Distance

The simplest similarity measures for time series is the Euclidean distance measure. Let two sequences with the same length n [1], the measure is defined as:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \tag{3.3}$$

Referring to formula 3.3 , if two series are equal, their distance is zero, which means two series are completely similar. The euclidean distance is the most used distance in pattern recognition. Of course as general purpose distance is good on large number of applications but in some cases it is not good or it is not the best. On seismogram comparison the distance is too sensible on variations of the signal on the second half or to background noise.

One variation of euclidean distance applied to seismic events is the application to the spectrum of the signal rather on the time series. On [OAEGODCD06] there is an application of the euclidean distance for dissimilarity-based classification of the seismic signals. The proposed method is oriented to computation of the spectrum, normalization, computing dissimilarities between spectra with pointwise euclidean distance, area difference between non-overlapping parts. However some problem's solutions are related to spectrum analysis because the spectrum of a seismic event

[1] if the sequences have different length it is simple to pad one of the two signals to have signals with the same length

depends on overall signal where a first half has more important details than the second. Because the spectrum is frequency-based and not time-based every change on frequency may cause a loss of information by adding weight where is not required.

On figure 3.1 we show four different seismograms and their spectra. We can see that the signals have different waveforms but in two case the differences on the spectra are less meaningful than the seimosgrams. The first and the third signal have a very different waveform related to p-wave and s-wave but the resulting spectra are quite similar. So a difference on spectra value may join these two signals while they have a different form with a different type of source. The same observation is valid for the second and the fourth signal which are different but have a similar spectra. In some way the spectra have a loss of information related to the shape of the seismic event. A problem of the spectra is that some frequencies are equal on different waveforms which have a similar aspect.

The measure has some advantage but also some disadvantages. The advantages:

- it is simple and well known by everyone
- it is fast because the computational time is dependent on the vector length

The disadvantages:

- great changes on little translations: a bit translated signal takes great difference on computed value
- soft dilation or compression may cause very different results

Despite to this problem the euclidean distance is still used in the last years. A good example is the work proposed in [PMDOA+10]. The measure uses a representation of three-way data as a dissimilarity on multidimensional objects.

3.3 Minkowski distance

The Euclidean distance is particular cases of the Minkowski distance defined by

$$d(x,y) = \left(\sum_{j=1}^{d} |x_j - y_j|^r \right)^{\frac{1}{r}}, \quad r \geq 1 \tag{3.4}$$

where the parameter r is the order of the distance. Of course, with $r = 2$ we have the Euclidean distance.

3.4 Mahalanobis Distance

The Mahalanobis distance is based on correlations between variables by which different patterns can be identified and analyzed. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant. In some way it is more suitable on seismograms than euclidean distance. It is defined as:

$$d(x,y) = \sqrt{(x-y)^t S^{-1} (x-y)} \tag{3.5}$$

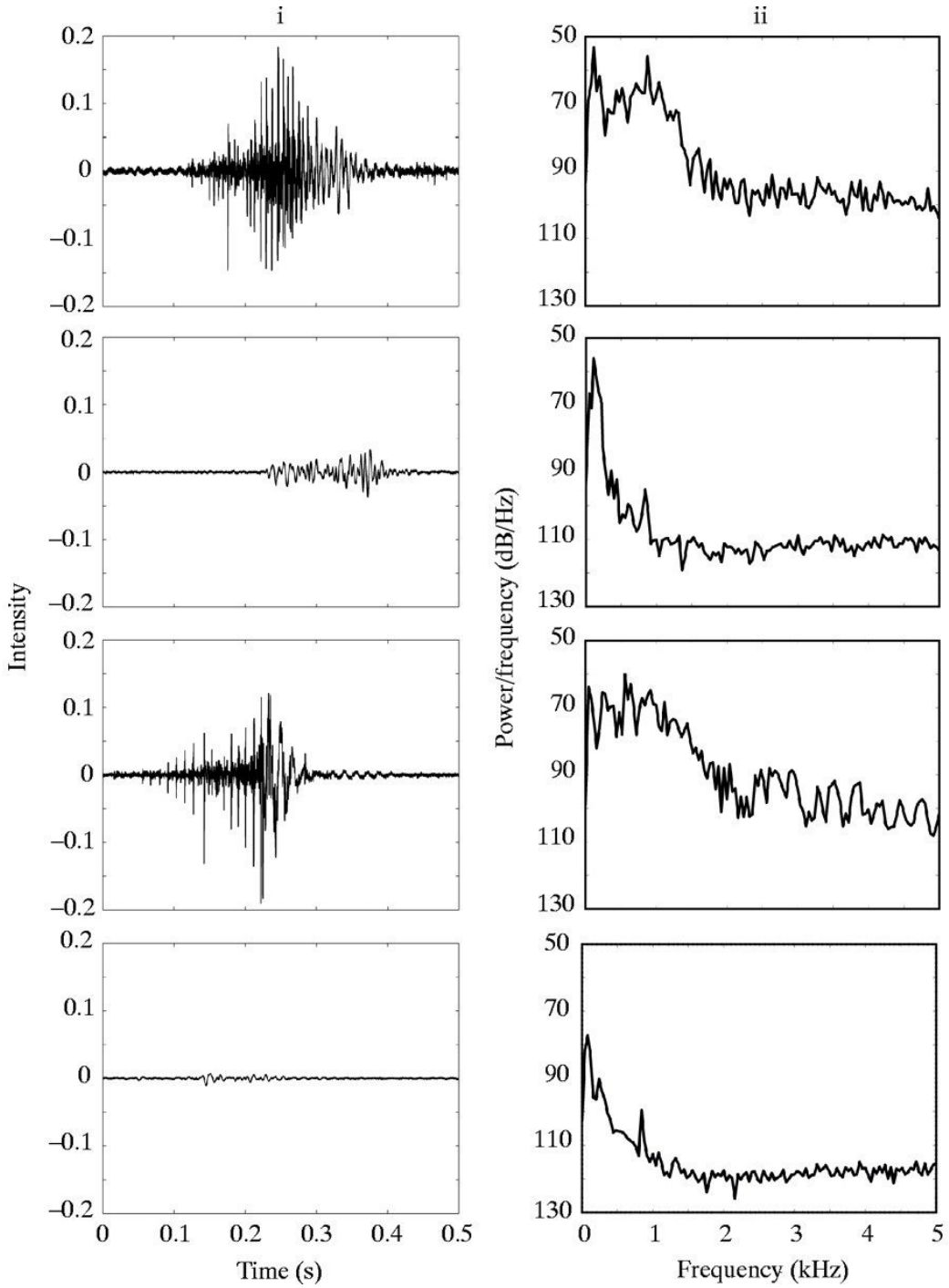
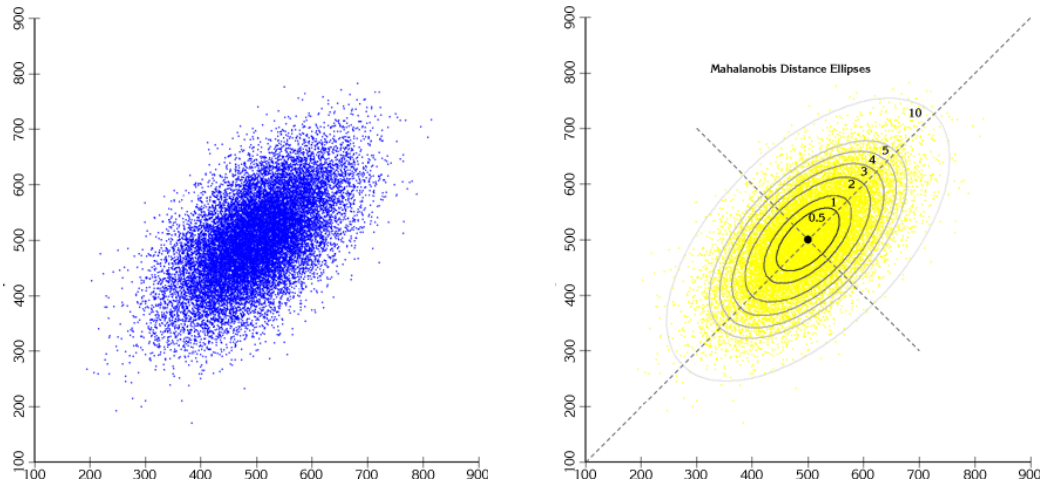


Figure 3.1. Signal examples and their spectra



(a) A sample distribution of points **(b)** The Mahalanobis distance values

Figure 3.2. A Mahalanobis example

where S is the $d \times d$ (d number of attributes) covariance matrix:

$$S = \frac{1}{n}\mathbf{X}^T\mathbf{X} \tag{3.6}$$

and $\mathbf{X}$ is an $n \times d$ (n number of elements in dataset) matrix.

$$\mathbf{X} = (x_{ij} - \bar{x}_j)_{n\times d} = \begin{pmatrix} \mathbf{x_1} - \bar{x}_1\mathbf{e}_d \\ \mathbf{x_2} - \bar{x}_2\mathbf{e}_d \\ \vdots \\ \mathbf{x_d} - \bar{x}_d\mathbf{e}_d \end{pmatrix} \tag{3.7}$$

where $\mathbf{e}_d$ is the d-dimensional identity vector.

Mahalanobis distance was prompted by the problem of identifying the similarities of skulls based on measurements in 1927. The Mahalanobis distance is a metric which is better adapted than the usual Euclidian distance to non spherically symmetric distributions. Figure 3.2 (b) shows the ellipsis which have the same value of distance around the centre and respect to distribution on (a).

The Mahalanobis distance accounts for the variance of each variable and the covariance between variables. Geometrically, it transforms the data into standardized uncorrelated data and computes the ordinary Euclidean distance for the transformed data. The Mahalanobis distance is like a univariate z-score because it measures the distance into the scale of the data.

In [AAHS06] has been introduced a discrimination using the Mahalanobis distance for artificial seismic signals generated by mine blasts. The application of the distance doesn't use the raw data but rather it is based on some discriminants: total of seven separate discriminants are computed, based on the spectrograms of recorded events. This work is a typical example of an application of the distance on extracted data. The feature extraction step is executed to extract the real information from data or when the computation over the raw data is not feasible.

3.5 Dynamic Time Warping

Sometimes a very simple distance measure such as the Euclidean distance is sufficient: it is often the case that two sequences have approximately the same overall component shapes, but these shapes do not line up along the time.

Dynamic time warping (DTW) is an algorithm for measuring similarity between two sequences which may vary in time or speed. Dynamic time warping is an extensively used technique in speech recognition which allows acceleration-deceleration of signals along the time dimension and many applications exist on seismology.

Finally we can define the DTW as described in [LCW10]. Let $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ be two series with the length of n and m, respectively, and an $n \times m$ matrix M called **local cost matrix**. We define M_{ij} as the distance (Euclidean) $d(x_i, y_i)$ between x_i and y_i. The relationship between X and Y can be defined as by the **warping path**:

$$W = (w_1, w_2, \ldots, w_k), \quad \text{with} \quad max(m, n) \leq k \leq m + n - 1 \tag{3.8}$$

where $w_k = (i, j)$.

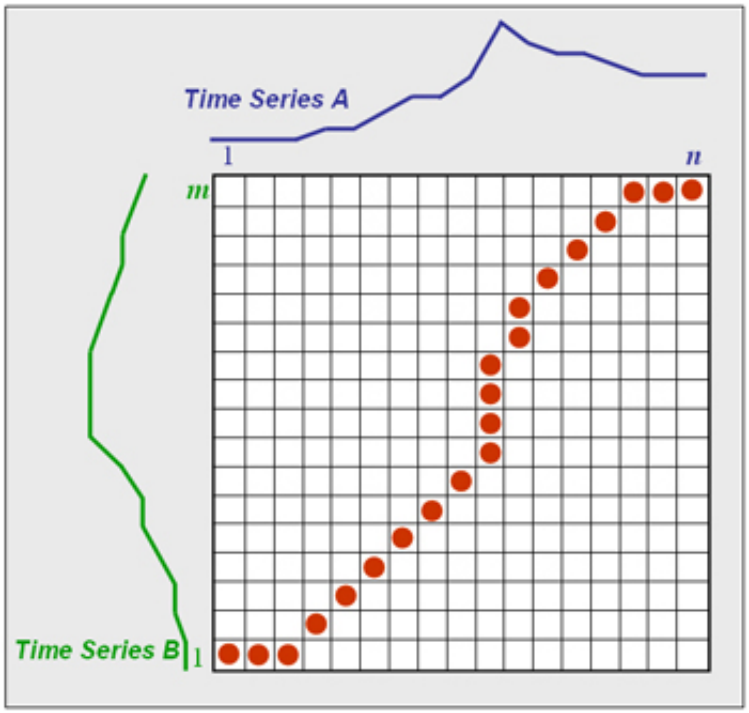


Figure 3.3. Dynamic Time Warping application between two signals

In the figure 3.3 we shown the application of dynamic time warping between two simple time series. Along the axis there are the two signals while inside the matrix there is the optimal path, in red color, which define the value of the DTW like the sum of distances between the single points.

The warping path must satisfy some criteria. The starting and ending points of the warping path must be the first and the last points of aligned sequences. The path indexes must meet a monotonicity condition that preserves the time-ordering

of points. A limit is applied to the the warping path from long jumps (shifts in time) while aligning sequence.

The cost function associated with a warping path computed over the local cost matrix M, a matrix of *local distance measure* between each pairs of the two compared sequences, is:

$$c_W = \sum_{k=1}^{K} M(w_k) \quad \text{with } w_k = (i, j) \tag{3.9}$$

The warping path which has a minimal cost associated with alignment called the optimal warping path. The dynamic time warping is defined as:

$$DTW(X, Y) = \min_{W}\{c_W\} \tag{3.10}$$

The dynamic time warping described above is the "naive" version but in many case a customized measure is used to be suitable on certain dataset. In [Sen08] the authors have proposed three customization based on step function, wighting and global path constraints. Every one of these techniques try to resolve problem related to excessive dilation or compression of some piece of the waveform. The application of the measure to specific signals, like seismic signals, requires the analysis of the results and a correction of the constraints to preserve a comparison over the shape of the signal. Furthermore the DTW requires a massive computation which is unsuitable on large dataset or long time series.

3.6 Cross Correlation

In this section we'll describe the classical *cross correlation dissimilarity* used very often in signal processing, it is a measure of similarity of two waveforms as a function of a time-lag applied to one of them.

We recall that the *cross correlation* between two vectors $\bar{x}_1$ and $\bar{x}_2$, both of length n, is so defined

$$R_{\bar{x}_1,\bar{x}_2}(k) = \begin{cases} \sum_{i=0}^{n-k-1} (\bar{x}_1(i+k) - \mu_{\bar{x}_1}) \times (\bar{x}_2(i) - \mu_{\bar{x}_2}) & \text{if } k \geq 0 \\ R_{\bar{x}_1,\bar{x}_2}(-k) & \text{otherwise} \end{cases} \tag{3.11}$$

for $k = 1 - n, .., n - 1$, and where μ_{x_1} and μ_{x_2} indicate the means of $\bar{x}_1$ and $\bar{x}_2$ respectively. Consequently, the cross correlation dissimilarity between $\bar{x}_1$ and $\bar{x}_2$ is

$$\delta_R(\bar{x}_1, \bar{x}_2) = 1 - \frac{1}{\sigma_{\bar{x}_1}\sigma_{\bar{x}_2}} \max_{k=1,..,2n-1} R_{\bar{x}_1,\bar{x}_2}(k - n). \tag{3.12}$$

Where $\sigma_{\bar{x}_1}$ and $\sigma_{\bar{x}_2}$ are the standard deviations of $\bar{x}_1$ and $\bar{x}_2$ respectively. Such dissimilarity is largely used to catch difference in shape between seismic signals, but in this context it has also shown some drawbacks. Moreover, for a signal of length n its computational time is $O(n^2)$.

The cross-correlation is one of the most used proximity measure in seismology. It is widely used for clustering and classification of seismic events. Indeed the cross-correlation is also a well known tool for alignment. The cross-correlation is used in

seismology as proximity measure. The value computed between two elements is in interval [-1,1]. The value 1 is obtained when the two vectors are totally correlated while -1 when they are opposite. The value 0 when they aren't correlated.

On figure 3.4 we show two equal signal where the second is a translation of the first and tail was cut to have the same length. Of course a computation of the cross-correlation will have a value near 1 because the signal are equal with some difference on the tail.

The cross-correlation between two signal of size N is a vector of length $2 * N - 1$. When the two vector don't have the same length, the shorter is zero-padded to the length of the longer vector. The computation between two signal is usually limited by a lag so that each value is obtained on interval [-lag:lag]. It is useless and expensive to evaluate the coefficient over a very large range when the computed value is over a little intersection of the two signals.

On experiments section of this thesis, the lag parameter of two signals S_0 and S_1 is fixed to value equal to a $max(length(S_0), length(S_1))/2$. In figure 3.5 we show the lag used to compute the cross-correlation. Finally, in figure 3.6 the cross-correlation of the two signals S_0 and S_1 of the figure 3.4. The length of the correlation vector is $2 * lag + length(signal)$.

As proximity measure we take the maximum of the cross-correlation vector. The maximum value computed between the signals in figure 3.4 is 0.99806166. As we can see, the value is close to one because the signals are the same.

In seismology there are a lot of references on cross-correlation. One of them, [BFMS07] use the application of cross-correlation analysis to define groups of dependent events (multiplets) characterized by similar location, fault mechanism and propagation pattern. On [BF01] the authors did a good analysis on source parameters and fault plane determinations by use of cross-correlation. They use the cross-correlation distance in a classification phase before to develop a focal mechanisms solution.

3.7 Cumulative shape

This work propose a dissimilarity measure which is an ensemble of the measures on the single components. The properties of the three dimension detection station are used to boost the comparison between element pairs.

The definition of the new dissimilarity was inspired by a simple observation

> *a seismic signal is characterized by two types of waves: body waves and surface waves.*

The body wave, the waves between the P and S arrivals, are less sensitive to the travel path and have no phase overlapping. Moreover, these seismic phases have often the better signal to noise ratio, so we can use them to discriminate one wave from the others. A seismic dataset is often a set of aligned (or not[2]) signals which contain the two types of body waves: P wave and S wave. Both waves have a

[2]many technics are used to cut and to align the signals: a common phase is the pre-processing of the signal with denoising, P phase identification and cut.

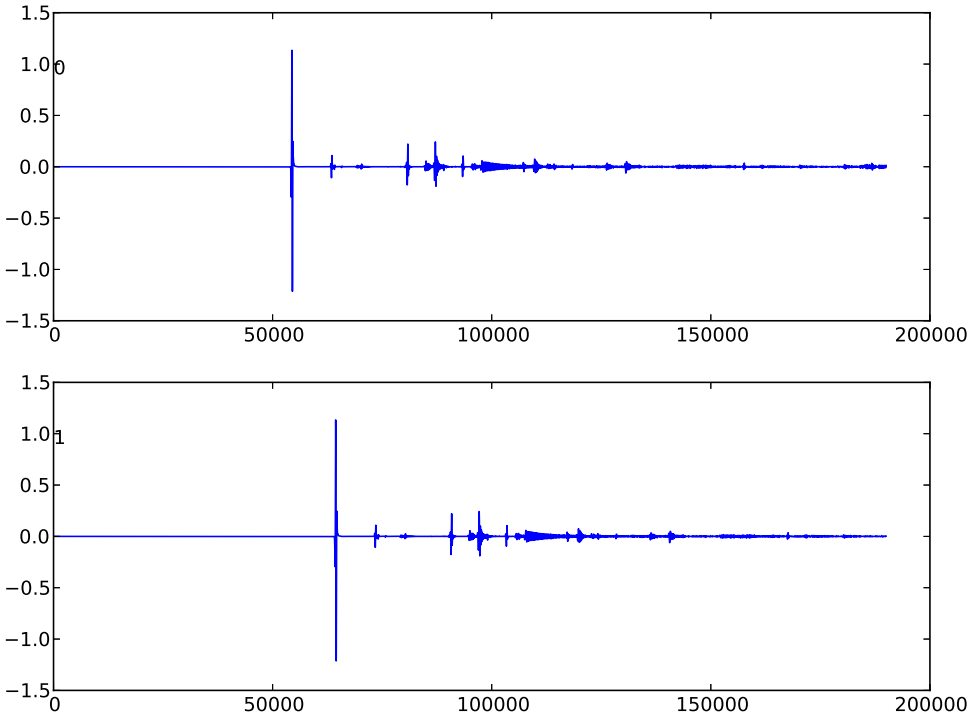


Figure 3.4. Signal translation for cross-correlation test S0 and S1

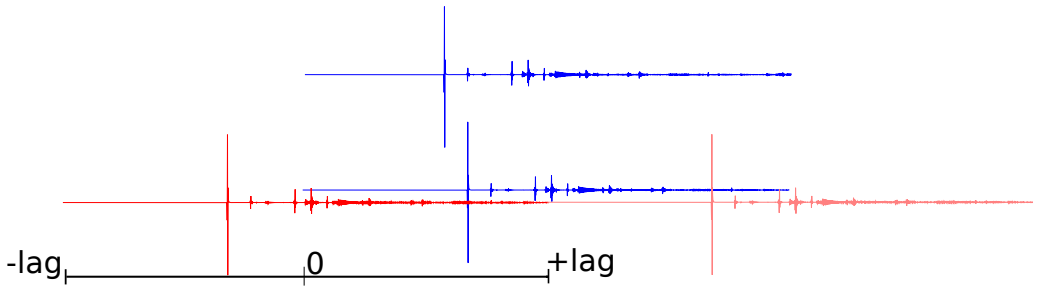


Figure 3.5. Signal translation for cross-correlation test S0 and S1

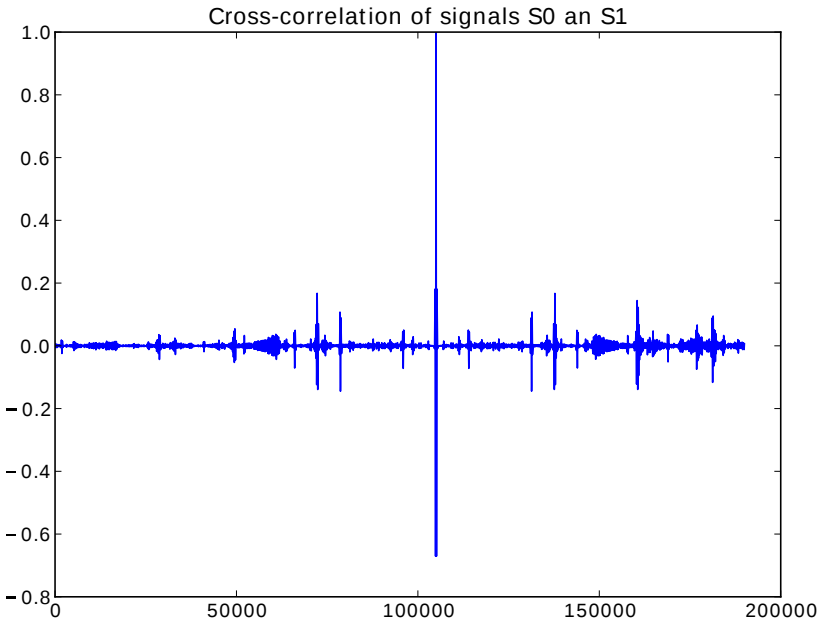


Figure 3.6. Cross-correlation of the signals S0 and S1 with lags=length/2

magnitude peak with high energy. Consideration about the nature of the data leads to state the main properties of a good dissimilarity measure for seismic signals :

- it should give high weight to the difference among the initial part of the signals,
- it should be low sensitive to background and impulsive noise,
- it should be capable of detecting where two wave shapes are similar regardless of magnitude

The first property can be graphically verified on figure 3.7. The second half of the signal, after the arrival of the s-waves, has a lot of waves component due to body waves and surface waves. The original signal is altered by the structure of the propagation path between the source and the detection station. Only the first half of the signal, between p-waves arrival and s-waves arrival, is more similar to original because the propagation mechanisms related to body waves are less sensible than surface waves to the earth structure.

The second property must be satisfied to make a proximity measure more stable on the presence of background noise. The nature of the background noise on seismogram can be natural or artificial. The filtering should clean the signal in the right way without loss of information. A measure less sensible to noise can be applied with a light filtering which preserve the information about the real seismic event.

The third property is required to identify events which are generated by the same source but with different intensity. A proximity measure like euclidean measure has a great value on amplitude difference due to a point-to-point difference. A

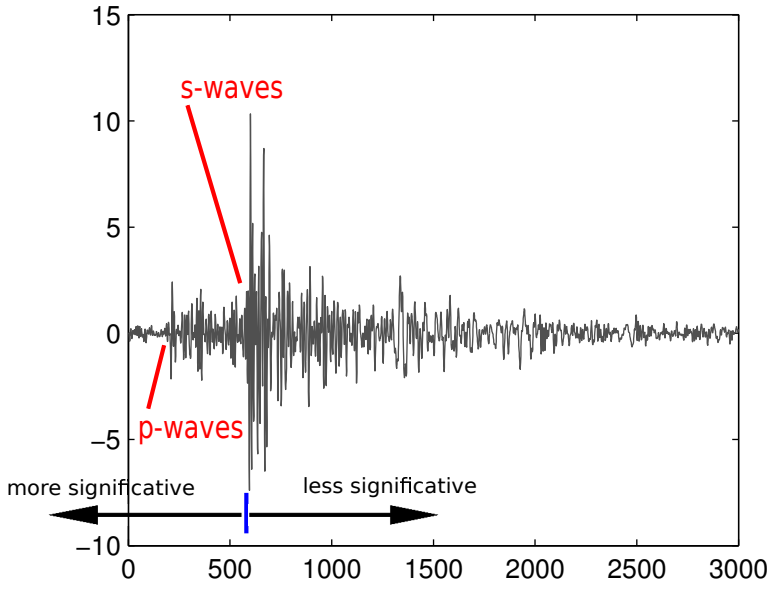


Figure 3.7. Properties related to seismic signals

good proximity measure should ensure that similar shape are discovered also with amplitude difference.

So far, we have defined some properties that a good proximity measure must have to evaluate seismic signals. But how these properties can be satisfied?

The first two properties can be satisfied by a dissimilarity representation based on the cumulative energy of the signals rather than on their original waveforms. Of course, the peaks of the P wave and S wave are well visible on cumulative energy graph whereas the tail of the signal has a tiny impact. All the properties are finally satisfied by a dissimilarity that take into account the evaluation of the difference between cumulative energies.

Given two vectors x_i and x_j both of the same length n, and let s_i and s_j be their cumulative sums

$$s_i(k) = \frac{\sum_{r=1}^{k} x_i^2(r)}{\sum_{r=1}^{n} x_i^2(r)} \qquad i = 1, 2 \qquad K \leq N \tag{3.13}$$

The previous equation define a normalized non-decreasing curve with values between 0 and 1. The curve is not decreasing because each cumulative point value at k is equal to value at cumulative $k-1$ plus energy at k. The normalization of the maximum value at 1 makes the measure invariant to amplitude of the seismogram but not to the shape. So we can calculate their absolute difference as

$$sd_{ij}(k) = |s_i(k) - s_j(k)| \tag{3.14}$$

Finally, the new proposed dissimilarity, called *cumulative shape dissimilarity* δ_s is defined as

$$\delta_s(x_i, x_j) = \sum_k |sd_{ij}(k+1) - sd_{ij}(k)| \tag{3.15}$$

The new measure satisfies the defined properties of an dissimilarity meaure with some changes:

- nonnegativity: $\delta_s(x, y) \geq 0$ because equation 3.15 is defined as sum of non negative differences so the result is non negative
- reflexivity: $\delta_s(x, y) = 0 \iff |x| = |y|$ is true because the numerator of the 3.15 is equal to zero only if $|sd(k+1) - sd(k)| = 0$ for each k and then if $sd(k+1) = sd(k)$, but 3.14 $\implies |s_1(k) - s_2(k)| = |s_1(k+1) - s_2(k+1)|$. So the latter two differences must be equal to a value P for each $0 \leq k \leq n$. But both $s_1(n)$ and $s_2(n)$ are equals to 1 so the difference $sd(n)$ is 0 and then must be $P = 0$ for each $0 \leq k \leq n$. $P = 0 \implies s_1(k) = s_2(k)$ for each $0 \leq k \leq n$.
- commutativity: $\delta_s(x, y) = \delta_s(y, x)$ because from 3.14 follows $sd(k) = |s_1(k) - s_2(k)| = |s_2(k) - s_1(k)|$

The constraint on reflexivity is less restrictive because it's clear that the transition from waveform to energy curve is with loss of information: two opposite waveform in some point have the same energy curve. It's a great problem in general but when applied to seismograms we have seismograms as very long sequences with a low probability that two waveform are opposite in sign. Indeed is more likely to have two different shapes rather then two same shapes with point differences in sign.

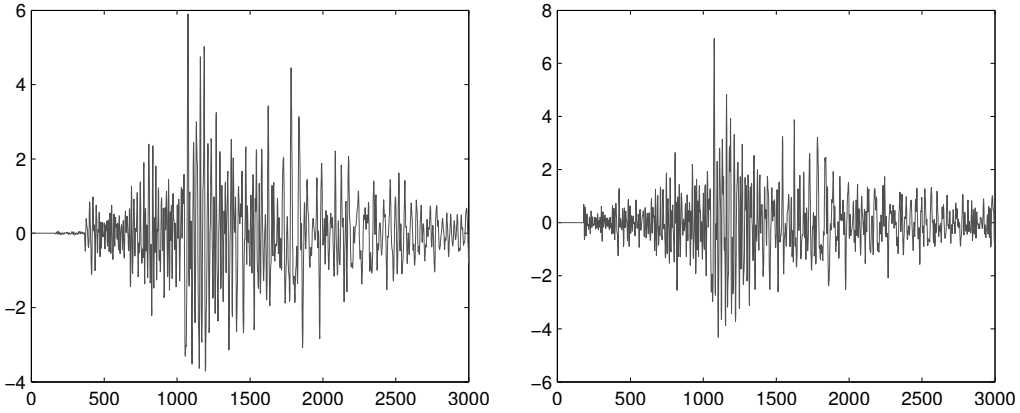
The definition of δ_s in 3.15 represents the sum of the derivative of the difference between the cumulative sums of x_1 and x_2. In some way the cumulative shape checks that the two cumulative energy curves rise in the same way at the same time. The value of the measure is higher when the energy curves have the same shape.

In figure 3.8 we report 4 examples of signal, in figure 3.9 their cumulative sums and in 3.10 the pairwise dissimilarities.

We can see that in the cumulative sum it is possible to identify the p-waves and s-waves arrivals in the graph where the curve has two concavities. The cumulative of energy is less sensible in time to amplitude values of the tail: the curve rises quickly at first, but less when the value of the initial energy is added to that remaining. The background noise is less evident because its value is constantly added to the energy curve. Two curve with similar shape have a similar cumulative energy curve so is more simple to detect near events.

Finally, in figure 3.13 we show the value of $|sd(i+1) - sd(i)|$ used to compute $\delta_s(x_1, x_2)$. Such example shows how similar shapes have lower dissimilarity values. It is important to note that the new measure δ_s have a remarkable computational time of $O(n)$.

The application of the cumulative shape requires a good cut of the signal at p-waves arrival. Although a fast alignment can be applied between two signal on the first part of them is is preferable to have each signal that starts with p-waves. Although the cumulative shape is more sensible than cross-correlation to the triggering of the signal, a good cut is in every way a necessary condition to have a good result. Also the value of cross-correlation is affected by the cut because the



(a) event 1 of Palermo dataset

(b) event 6 of Palermo dataset

(c) event 32 of Palermo dataset

(d) event 79 of Palermo dataset

Figure 3.8. Sample events of the Palermo dataset earthquakes

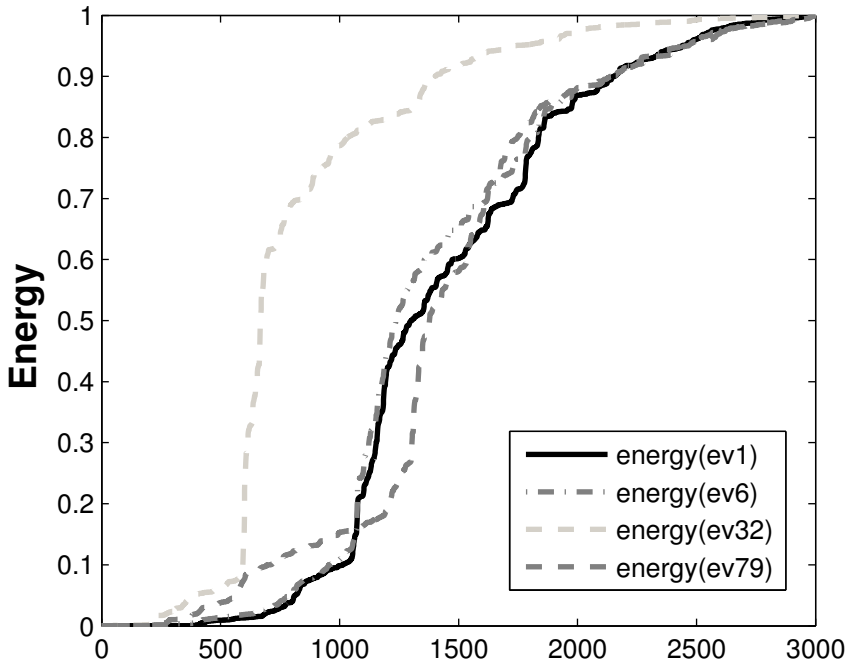


Figure 3.9. Cumulative energy of the events

value is computed over all the signal so that a not required portion or a miss can affect the computed value with different weights.

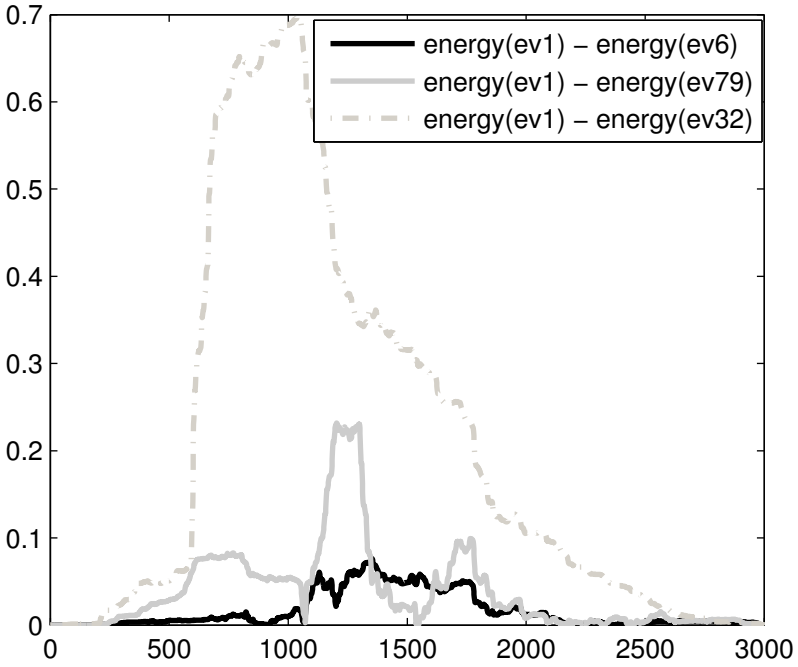


Figure 3.10. Difference between cumulative energies

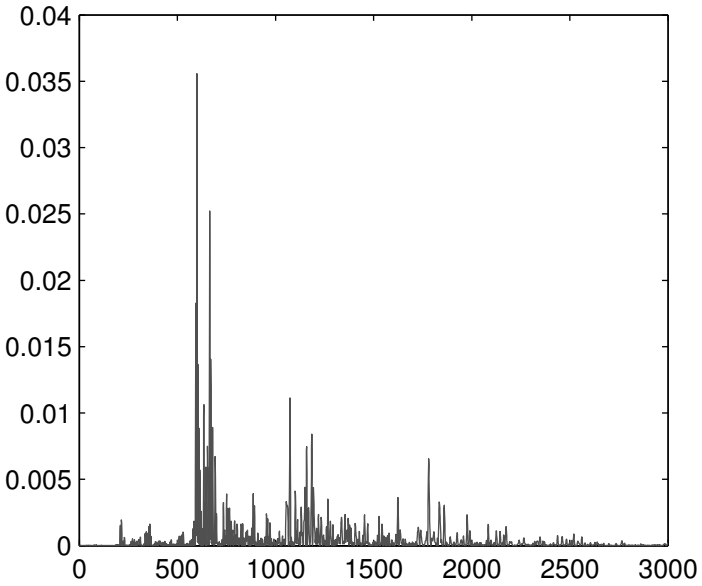


Figure 3.11. Derivative at sample point i of the difference between cumulative energies ($|sd(i+1) - sd(i)|$) (a) event 1 - event 32

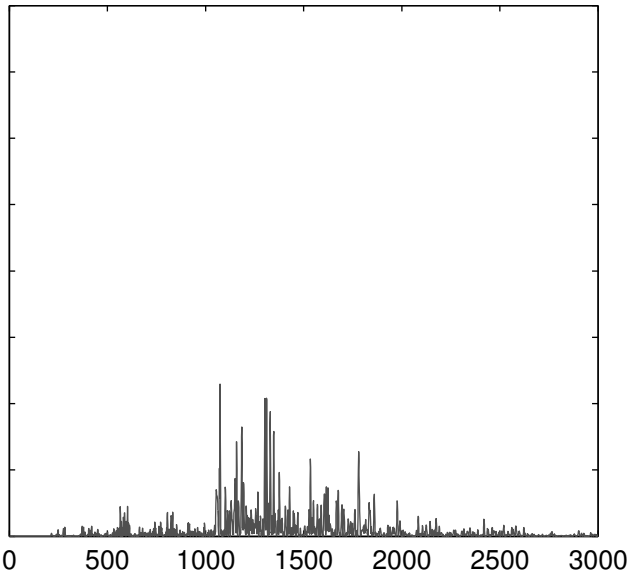


Figure 3.12. Derivative at sample point i of the difference between cumulative energies ($|sd(i+1) - sd(i)|$) (b) event 1 - event 79

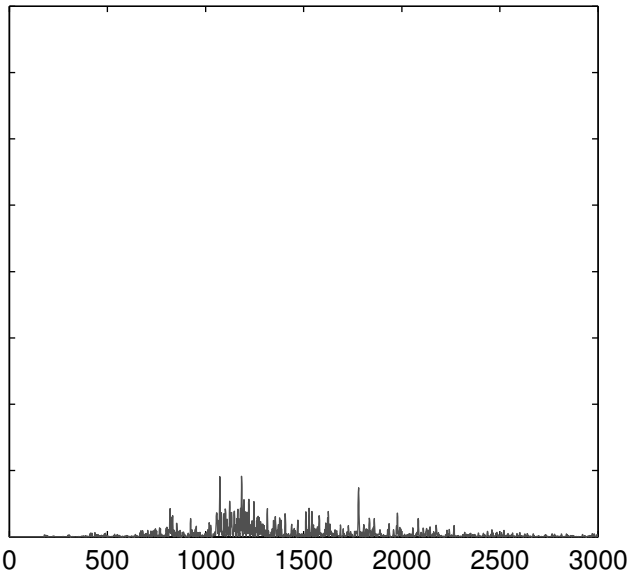


Figure 3.13. Derivative at sample point i of the difference between cumulative energies ($|sd(i+1) - sd(i)|$) (c) event 1 - event 6

Chapter 4

Evaluating the measures

4.1 Preparing the experiments

A number of experiments have to be done to evaluate the performance of a proximity measure. A dissimilarity measure for clustering and its application on specific seismic signals has been developed in this contribution. The tests are executed on several contexts with different types of signals and environments.

The evaluated dataset are mainly of two type: natural and artificial. The natural data are recording of real events detected by some station in a time interval ranged between one or two months. The artificial data are generated by simulation that can be executed in real or in virtually environment. In the following we'll describe three type of dataset:

artificial on virtual environment the events are fully generated by simulation tools. The simulation requires a lot of work on design and configuration of the parameters of the model. The scientists design several models to test algorithms and infer properties on real environment

artificial on real environment the events are artificially generated and they are recorded by the same detection station used to store real earthquakes.

real the dataset element are real seismic events. The elements are recorded by a detection station in a digital format like SAC (Seismic Analysis Code). All the elements are generated in a large delimited zone and localized by seismology experts

Three different dataset have been used on this work, one for each type. They are used to test the proposed distance versus the most used instance used by seismologists: maximum of the cross-correlation. In literature the cross-correlation is the most used tool by seismologists. One of the main reason of this large use is the simple access to software tools and the interpretation of the results. The cross-correlation is used as distance but it is also used to investigate about relationship between seismic events record. Also it is widely used on classification and location of the earthquakes source. Some complex methods and measures based on wavelets or neural systems exist but they are diffused on computer science community articles rather than in seismology.

The references about cross correlation on geological and sismology journals are many, so we cite some of the most recent and more relevant:

- classification: [BFMS07] use the application of cross-correlation analysis to define groups of dependent events (multiplets) characterized by similar location, fault mechanism and propagation pattern; [GN06] use cross-correlation techniques to quantify the degree of similarity, they developed a method to sort events into families containing comparable waveforms; [SC01] use cross-correlation to exploit the high degree of similarity among waveforms; [UTS+08] conducted a cross-correlation analysis of waveforms to detect earthquake families
- characterization of the waves: [KP12] identified 146 event families that occurred within this suite of selected events using a cross correlation technique; [MTI09] use cross-correlation to characterize the phases; [ZYD+12] use cross-correlations between many station pairs at the same inter-station distance to characterize the phases; [TWS10] use cross-correlation as similarity measure to detect multiplets and characterize seismic events before and after a volcanic eruption, [JCM12] use cross-correlation to analyse the volcanic tremor
- clustering: [MMP11] use a cross-correlation technique and hierarchical clustering algorithm in order to determine the origin of waveform similarity and the distribution of earthquakes producing similar waveforms; [MFB+11] use cross-correlation of seismic waveforms and clustering to improve the earthquake locations and determine the best-constrained focal mechanisms ; [BWT12] use cross correlation-based clustering techniques applied to waveforms from one representative data channel; [PBB+12] use waveform cross-correlations to cluster events and to improve the selected phase picks.

Sometimes the cross-correlation is used in simple way on the flat signal data while sometimes it is applied on spectra or on wavelets. In every case it is used as tool to discover the relationship based on shape similarity among the signals. The reason of the latter observation is in the cross-correlation definition: the value is higher when both the signal rise or fall at the same point and for a long time.

Of course, a full list is available on specialized search engine like *ScienceDirect* to assess that cross-correlation is today the most used tool for pattern recognition on seismic events.

The classification of the seismic events is used by sismologists to infer some labeling on new events respect to a known dataset. When a group of seismic events is collected and the sismologists did a classification of them, the next step is to assign new element to a known group. Otherwise the clustering is used by sismologists to infer an unknown grouping of the collected dataset. It is simple to understand that the classification permits to check the obtained results in some way while the clustering is characterized by no knowledge about the nature of the data. These two problems are complementary but have usually something in common in their algorithms: a proximity measure.

Many classification and clustering algorithms use a proximity measure to detect a label of a new element or to group a unknown dataset. The previous chapters have depicted some clustering algorithms which require a distance measure.

4.1.1 Simulated dataset

This dataset was generated by the using the E3D tool created by the Lawrence Livermore National Laboratory of the University of California. The simulation model is based on models defined by [Mad76], [Vir86], [Lev88] and [LH93].

E3D is able to simulate seismic wave propagation in a 3D heterogeneous earth. Seismic waves are initiated by earthquake, explosive, and/or other sources. These waves propagate through a 3D geologic model, and are simulated as synthetic seismograms or other graphical output.

The software simulates wave propagation by solving the elastodynamic formulation of the full wave equation on a staggered grid. The solution scheme is 4th-order accurate in space, 2nd-order accurate in time.

The computation of a simulated dataset requires a long computational time and lot of resources. Theses simulations are usually executed on high performance clusters (HPC) because they use a massive parallelism. For these reasons we have created many models with a balance between feasibility and precision. Feasibility in term of computational time is needed to generate a single seismogram and precision to have a signal as suitable as real. These two parameters are inversely proportional to previous ones because greater details imply longer times of computation.

All the models created have in common the earth structure. Each simulation needs the definition of the earth structure where the waves propagate because the propagation path leverages the shape of the waveform at detection station. The velocity model is composed of five block. Each block is an horizontal layer over the distance between the detection station and projection of the source in the earth surface. The hypothesis of an horizontal layer is used to simplify the model without losing some real characteristics of the Earth's crust. The block are defined by six parameters:

start depth starting z position of block element

end depth ending z position of block element

gradient vertical gradient (units per km)

p P-wave velocity in km/sec

s S-wave velocity km/sec

r density g/cm^3

The previous parameters are used to describe the physics of the propagation path in a more realistic way. The values are fixed to simulate a real mean of several rock types. The used parameters of the crust are defined on table 4.1.

On our simulation we have chosen a 2D model on a grid of size $30Km$ of length and $60km$ of depth. The source event are located at $7.5km$ of distance from origin while the detection station is at $22.5km$. The full environment is shown on figure 4.1.

The dataset is composed of two main set of events: explosive and fault. The first type is a compressional or explosive P source while the second is defined in terms of

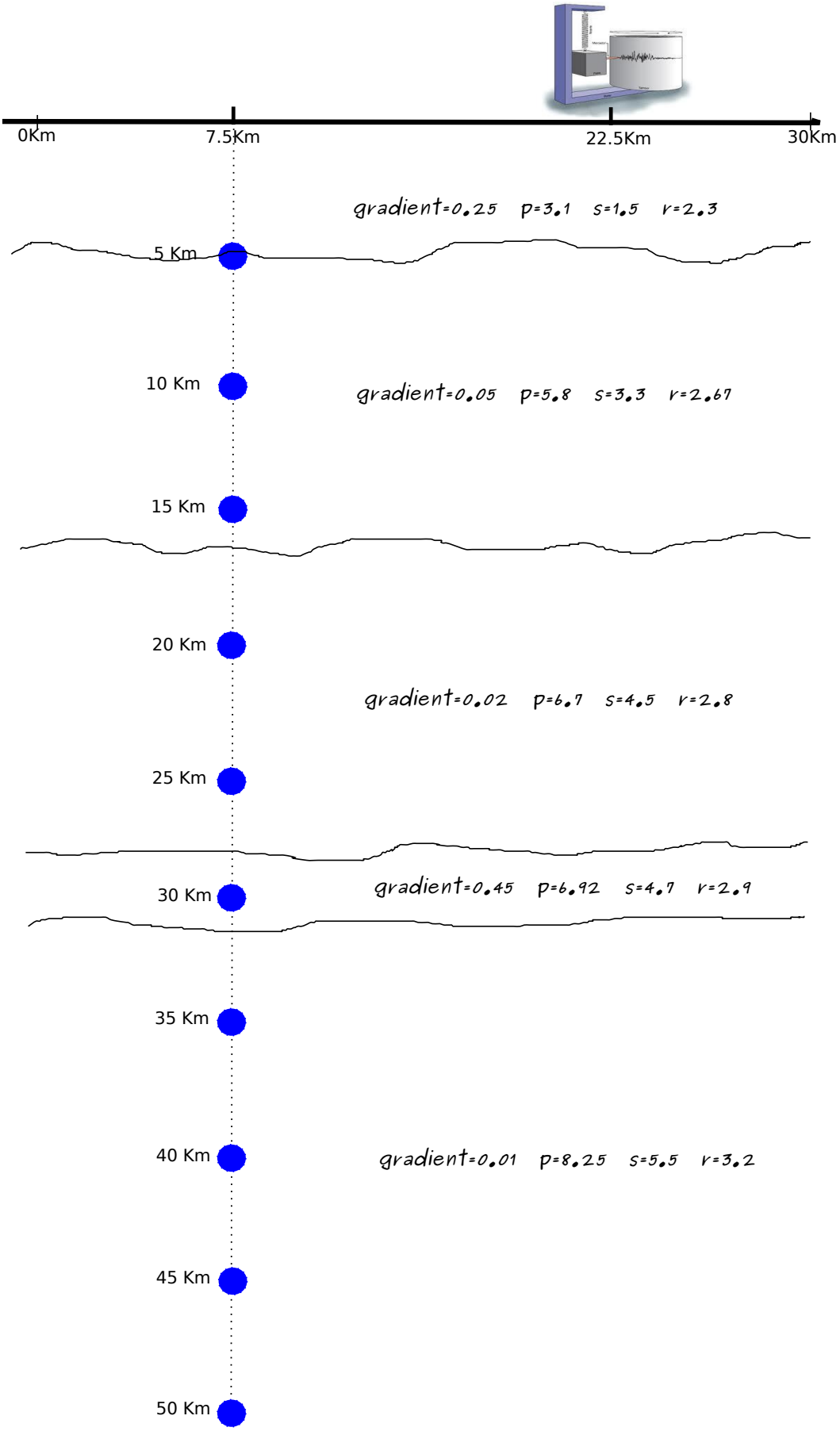


Figure 4.1. Velocity model used on simulation with software E3D

start depth	end depth	gradient	p-wave velocity	s-wave velocity	density
0	5	0.25	3.1	1.5	2.3
5	17	0.05	5.8	3.3	2.67
17	28	0.02	6.7	4.5	2.8
28	31	0.45	6.92	4.7	2.9
31	60	0.01	8.25	5.5	3.2

Table 4.1. Layer of the simulated model

a finite-length fault with uniform moment. We have defined 10 explosive dataset with 20 sample for each. We have a dataset for each depth from $5km$ to $50km$ with a step of $5km$. For each depth we have simulated 20 explosion with some coordinate variations in a a range about $4km$. The source coordinate variations are added by two normal distribution with mean 0 e variance 1. For each depth we have a cloud of seismic source around the coordinates of a center.

The coordinate differences are a problem for proximity measure on comparison between events because the detection station is at short distance of $15km$ from the event at surface level and a variation on propagation path length brings a large variation on the timing of the signal and its phases. Due to this reason a good triggering/picking must be done before to do any measure application between signals.

The second type of dataset, fault based, is defined on the same velocity model but with a different source. The fault source is defined by three main parameters: strike, dip and rake. As described in [AR02] we report th definition of the previous parameters:

strike

Fault strike is the direction of a line created by the intersection of a fault plane and a horizontal surface, 0° to 360°, relative to North. Strike is always defined such that a fault dips to the right side of the trace when moving along the trace in the strike direction. The hanging-wall block of a fault is therefore always to the right, and the footwall block on the left. This is important because rake (which gives the slip direction) is defined as the movement of the hanging wall relative to the footwall block.

dip

Fault dip is the angle between the fault and a horizontal plane, 0° to 90°.

rake

Rake is the direction a hanging wall block moves during rupture, as measured on the plane of the fault. It is measured relative to fault strike, ±180°. For an observer standing on a fault and looking in the strike direction, a rake of 0° means the hanging wall, or the right side of a vertical fault, moved away from the observer in the strike direction (left lateral motion). A rake of ±180° means the hanging wall moved toward the observer (right lateral motion). For any rake> 0, the hanging wall moved up, indicating thrust or reverse motion

on the fault; for any rake $< 0°$ the hanging wall moved down, indicating normal motion on the fault.

The stripe, dip and rake are shown on figure 4.2.

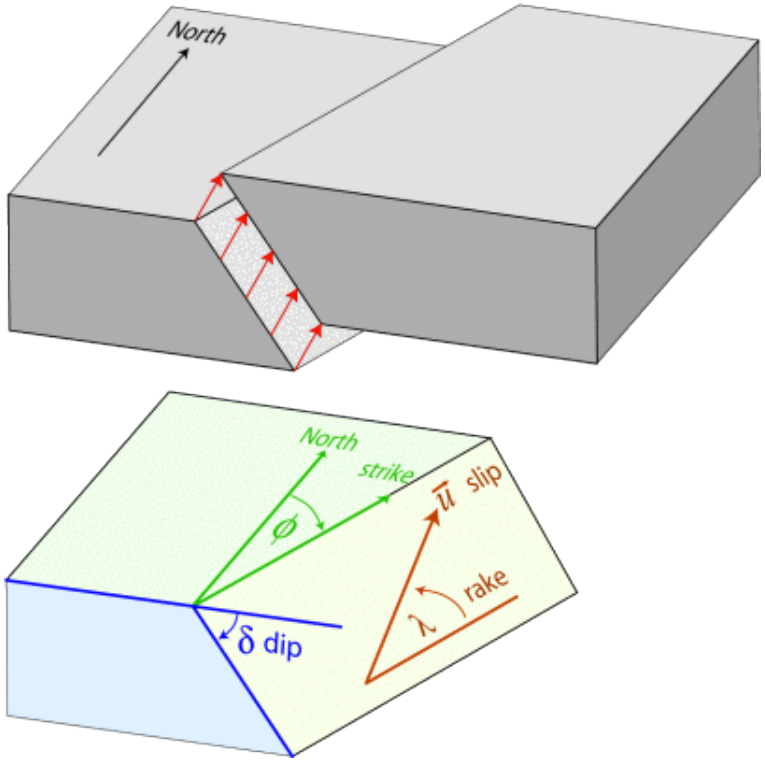


Figure 4.2. Strike, dip and rake of fault

A full 3D model must include all the parameters described above but in a 2D model we can use only the *dip* and *rake*. In details we have created a model for two types of focal mechanisms:

strike slip have walls that move sideways, not up or down. That is, the slip occurs along the strike, not up or down the dip. In these faults, the fault plane is usually vertical, so there is no hanging wall or footwall. The forces creating these faults are lateral or horizontal, carrying the sides past each other. It's shown on figure 4.3.

reverse form when the hanging wall moves up. The forces creating reverse faults are compressional, pushing the sides together. It's shown on figure 4.4.

The generated dataset are the following:

- the appendix A.1 shows the dataset generated with a source of explosion type at the depth of 5 km. The event are generated in a cloud large about 2-3 km around the basic depth.

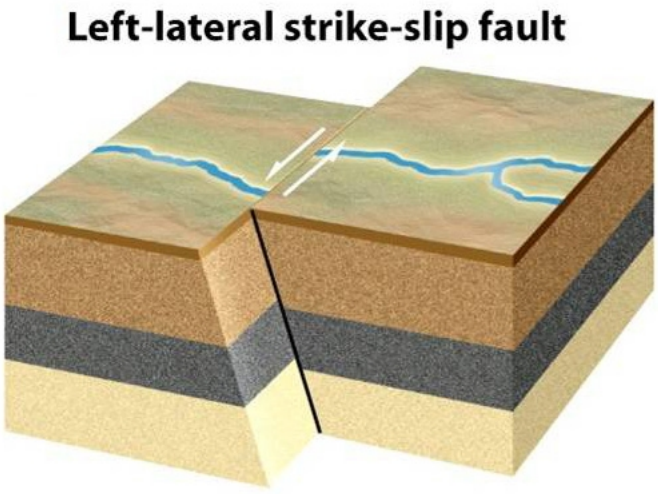


Figure 4.3. Strike slip fault

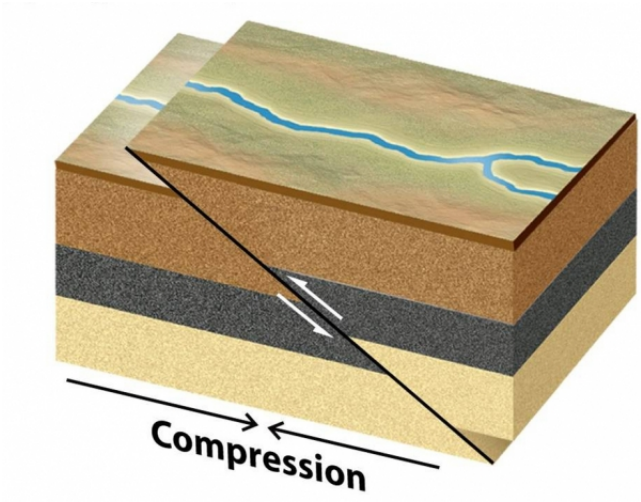


Figure 4.4. Reverse fault

- the appendix A.2 shows the dataset generated with a source of explosion type at the depth of 10 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.3 shows the dataset generated with a source of explosion type at the depth of 15 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.4 shows the dataset generated with a source of explosion type at the depth of 20 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.5 shows the dataset generated with a source of explosion type at the depth of 25 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.6 shows the dataset generated with a source of explosion type at the depth of 30 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.7 shows the dataset generated with a source of explosion type at the depth of 35 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.8 shows the dataset generated with a source of explosion type at the depth of 40 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.9 shows the dataset generated with a source of explosion type at the depth of 45 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.10 shows the dataset generated with a source of explosion type at the depth of 50 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.11 shows the dataset generated with a source of reverse fault type at the depth of 25 km. The event are generated in a cloud large about 2-3 km around the basic depth.
- the appendix A.12 shows the dataset generated with a source of slip fault type at the depth of 25 km. The event are generated in a cloud large about 2-3 km around the basic depth.

4.1.2 Bursts dataset

This dataset is classified artificial type on real environment. Although the dataset was generated for seismological research it is suitable to test some properties of the distance measures applied to seismic signals. The experiments is very useful because, it produces a series of well known events.

Between April 7 and May 8 2010, was carried out a multidisciplinary geophysical investigation in the framework of the MEDOC project. In the first part of the experiment 4 wide angle seismic profiles, crossing the entire Tyrrhenian basin in East-West direction were acquired together with a fifth profile between southern Sardinia and Sicily. The seismic energy was produced by airgun bursts operating on the Sarmiento de Gamboa vessel, located at constant distance between them, placed at different distances from the OBS/H, and recorded with high signal to noise ratio. The OBS/H is an Ocean Bottom Seismometer with Hydrophone developed by the OBS Lab of the Istituto Nazionale di Geofisica e Vulcanologia at Gibilmanna. In particular, the airgun bursts occurs at regular interval times of 45s and the seismic sensor of the OBS/H records for each burst a signal s_i at time t_i that express the variation of the pressure level over time. Figure 4.5 shows the arrangement of the experiment.

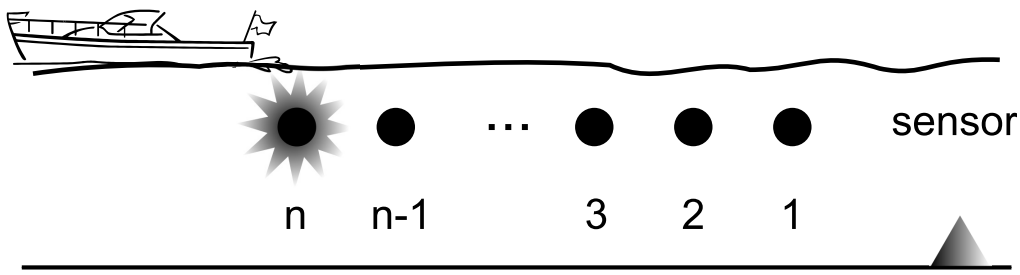


Figure 4.5. Plan of the bursts experiment

The acquired data define what is here named as *bursts dataset*, they can be considered a controlled dataset builded in order to have a well characterized set of signals to be used as a benchmark for problems involving seismic signals. The main assumption, is that close temporal explosions occurs at similar distances from the OBS/H. It is finally composed by the Up-Down component of 919 signals of maximum length 12000 sample points. Indeed the signals are recorded as a single long sequence where it is possible to identify each one of the simulated event.

The assumption about the proximity between signals is very important to check the behavior of a distance measure. Of course, we can think that close events in time, and space, are near in terms of distance. Two following events are generated by the same source, detected by the same station and have very near propagation among them. The difference of the propagation path is influenced by time that occurs among many consecutive bursts and by the motion of the boat. A difference on the propagation path affect the shape of the detected signal.

4.1.3 Palermo dataset

The Palermo dataset is a real dataset composed by seismic events recorded near the coast of the Sicily, Italy. The sea area in front of north sicilian coast is very important from seismic point of view. This area is affected from a lot of events due to faults and volcanoes. So many recording tools are located in the coast to detect any earthquake in a large area. The figure 4.6 show the area interested for monitoring with recorded events.

On the 6th September 2002, at 01:21 UTC, a strong earthquake (M_W 5.9) occurred in the northern Sicilian offshore. The seismic event was recorded by the Istituto Nazionale di Geofisica e Vulcanologia (*INGV*) network and located at about 50 km in NNE direction, from the Palermo city. In the following months, more than a thousand of aftershocks were located in the same epicentral area [GLT+09].

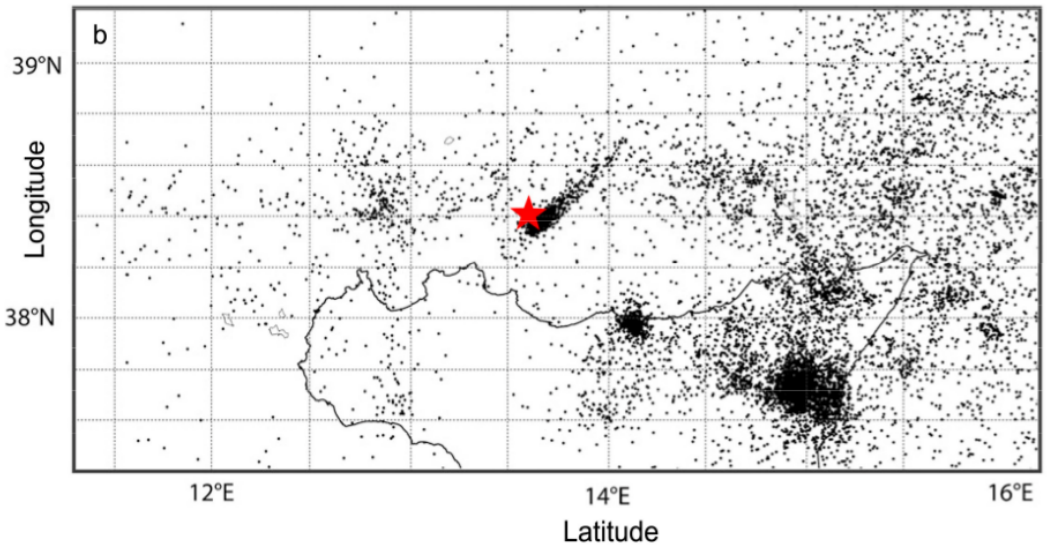


Figure 4.6. Location of the events in Palermo dataset

In December 2009, to better monitoring the seismicity of the Palermo 2002 epicentral area, the Gibilmanna OBSLab of INGV installed an Ocean Bottom Seismometers with Hydrophone (*OBS/H*) near the epicentral area of the mainshock, at a depth of about 1500 m. The 3 Component velocity signals (Up-Down, Nord-Sud, East-West) was digitized with a 21 bit data logger with a sampling frequency of 200 Hz.

The OBS/H recorded several teleseismic and regional earthquakes and about 250 local micro-events not located by the land network. The magnitude of the local events ranges between -0.5 and 2.5 M_L, and the delay between the S wave and P wave arrival times ($T_S - T_P$) ranges between 0.2 s and 5 s. A visual analysis of the seismograms revealed some similarity. To better characterize the recorded micro-seismicity we located 159 micro-events, with Signal to Noise ratio greater than a selected threshold, with a 3C single station location technique based on the polarization analysis of the signals [DLD+10]. 95 of this microevents, signed on figure 4.6 by black points, have been selected for our study. The resulting dataset, is denoted as *Palermo earthquake dataset*, and is finally composed by only the Up-Down component of 95 signals of length 3000 sample points.

The dataset is validated by an expert through a cluster analysis with the hierarchical method and maximum of cross-correlation as distance. The final configuration was verified respect to hypocenters location and waveforms shape. After a first analysis the events grouped inside the same cluster have near source location and similar shape. This is a start point for a deep research about relationship between shape and source mechanisms.

4.2 Results

4.2.1 Results on simulated dataset

The result on simulated dataset are shown on table 4.2 and 4.3. The evaluation indexes computed on the results are: Adjusted Rand Index, Mirkin Index, Homogeneity and Separation.

The table 4.2 shows results computed by hierarchical clustering with cumulative shape and cross-correlation distance. The solution computed by cumulative shape is often correct and equal to the original configuration. The worst results are obtained in comparisons which contain the datasets at 5 km and 10 km of depth. The reason of this behavior is determined by the short distance between the hypocenter and the detection station. Each hypocenter coordinate is perturbed by a normal distribution to move the source along the axis for a distance about 2-3 km. This movement of the source is very meaningful for distance of about 16 km but is less significative for higher distance at a greater depth. The values obtained with homogeneity and separation are quite high for both distances but for cross-correlation distance are less meaningful because this distance is unable to compute a correct cluster solution so that every internal index is not relevant.

The table 4.3 shows results computed by partitional clustering with cumulative shape and cross-correlation distance. The solution computed by cumulative shape is often correct and equal to the original configuration. The worst results are obtained in comparisons which contain the dataset at 5 km of depth. The reason of this behavior is described above. However the partitional clustering has better results on depth 10 km. The cross-correlation distance has better results with partitional clustering but they are still poor. The values obtained with homogeneity and separation are quite high for both distances but for cross-correlation distance they are less meaningful because this distance is unable to compute a correct cluster solution so that every internal index is not relevant.

The results on table are also reported on the following figures. The figures 4.7 and 4.8 show Adjusted Rand Index for hierarchical and partitional methods respectively. The figures 4.9 and 4.10 show homogeneity for hierarchical and partitional methods respectively. The figures 4.11 and 4.12 show separation for hierarchical and partitional methods respectively. All the figures have values computed for both cumulative shape and cross-correlation distance.

We can summarize results with the mean of the computed values:

- 0.525 and 0.005 are the mean of the adjusted rand index for cumulative shape and cross-correlation distance respectively with hierarchical method
- 0.548 and 0.513 are the mean of the homogeneity for cumulative shape and cross-correlation distance respectively computed with hierarchical method
- 0.863 and 0.690 are the mean of the separation for cumulative shape and cross-correlation distance respectively computed with hierarchical method
- 0.580 and 0.133 are the mean of the adjusted rand index for cumulative shape and cross-correlation distance respectively with partitional method

- 0.538 and 0.527 are the mean of the homogeneity for cumulative shape and cross-correlation distance respectively computed with partitional method
- 0.777 and 0.553 are the mean of the separation for cumulative shape and cross-correlation distance respectively computed with partitional method

To evaluate the measure and how it works at change of the physic distance between events we have plotted a diagram that reports on axis the physic distance and computed distance respectively. The diagram is generated for both cumulative shape and cross-correlation distance. In detail, a first diagram is printed without the dataset at 5 km of depth while a second with that dataset included. We excluded the dataset at 5 km because all the distance measures have many problem related to the short distance between the event and the detection station as described above. On figure 4.13, 4.14, 4.15 and 4.16 we can see the results. In the last two figures we can see how the dataset at 5 km of depth has many problems on comparisons with other. The proximity measure values with the 5 km dataset are too closer each to other in one way and too far in other. Finally, the values computed with cumulative shape tend to grow with the physics distance while in the cross-correlation based figure this aspect is less evident. This grow is the reason of the good behavior of the new distance.

Hypocenters			Adjusted Rand I.		Mirkin Index		Homogeneity		Separation	
distance	I.depth	II.depth	c.s.	xcorr	c.s.	xcorr	c.s.	xcorr	c.s.	xcorr
5	5	10	0.003	0.003	0.51	0.51	0.545	0.486	0.975	0.84
5	15	10	0.0	0.0	0.51	0.51	0.546	0.51	0.818	0.688
5	15	20	0.011	0.011	0.505	0.505	0.579	0.557	0.76	0.686
5	25	20	0.708	-0.003	0.146	0.513	0.54	0.504	0.755	0.576
5	25	30	0.024	0.088	0.497	0.456	0.612	0.598	0.791	0.696
5	30	35	1.0	-0.003	0.0	0.513	0.56	0.536	0.863	0.676
5	40	35	1.0	-0.003	0.0	0.513	0.574	0.568	0.831	0.697
5	40	45	0.0	-0.024	0.51	0.513	0.71	0.504	0.849	0.678
5	50	45	0.272	-0.003	0.364	0.513	0.627	0.662	0.79	0.81
10	5	15	0.003	0.003	0.51	0.51	0.532	0.454	0.979	0.841
10	20	10	0.0	-0.003	0.51	0.513	0.539	0.429	0.806	0.593
10	20	30	0.8	0.003	0.1	0.51	0.537	0.508	0.798	0.577
10	25	15	1.0	0.0	0.0	0.51	0.537	0.511	0.805	0.661
10	25	35	1.0	-0.003	0.0	0.513	0.498	0.518	0.824	0.671
10	40	30	1.0	0.024	0.0	0.497	0.55	0.621	0.798	0.746
10	45	35	0.8	-0.005	0.1	0.513	0.613	0.634	0.824	0.724
10	50	40	1.0	0.028	0.0	0.486	0.586	0.593	0.867	0.738
15	5	20	0.003	0.011	0.51	0.505	0.526	0.418	0.981	0.803
15	15	30	1.0	0.011	0.0	0.505	0.505	0.517	0.812	0.665
15	20	35	1.0	-0.003	0.0	0.513	0.483	0.549	0.788	0.694
15	25	10	0.0	-0.003	0.51	0.513	0.535	0.388	0.804	0.554
15	25	40	1.0	0.009	0.0	0.505	0.481	0.606	0.777	0.738
15	30	45	1.0	-0.003	0.0	0.513	0.569	0.552	0.89	0.688
15	50	35	1.0	-0.026	0.0	0.513	0.593	0.618	0.874	0.707
20	5	25	0.003	0.011	0.51	0.505	0.525	0.408	0.981	0.805
20	15	35	1.0	-0.003	0.0	0.513	0.478	0.586	0.78	0.69
20	20	40	1.0	-0.004	0.0	0.51	0.481	0.52	0.851	0.597
20	25	45	1.0	-0.003	0.0	0.513	0.541	0.538	0.88	0.683
20	30	10	0.0	-0.003	0.51	0.513	0.555	0.397	0.821	0.56
20	50	30	1.0	0.024	0.0	0.497	0.514	0.599	0.889	0.693
25	5	30	0.003	0.011	0.51	0.505	0.528	0.41	0.981	0.807
25	15	40	1.0	0.011	0.0	0.505	0.475	0.539	0.852	0.686
25	20	45	1.0	-0.003	0.0	0.513	0.537	0.546	0.895	0.659
25	35	10	0.0	-0.003	0.51	0.513	0.598	0.429	0.838	0.563
25	50	25	1.0	0.011	0.0	0.505	0.478	0.586	0.855	0.716
30	5	35	0.003	0.011	0.51	0.505	0.53	0.425	0.979	0.802
30	15	45	1.0	0.011	0.0	0.505	0.541	0.558	0.894	0.637
30	40	10	0.0	-0.003	0.51	0.513	0.62	0.414	0.873	0.572
30	50	20	1.0	0.024	0.0	0.497	0.49	0.568	0.871	0.635
35	5	40	0.003	0.011	0.51	0.505	0.533	0.415	0.98	0.807
35	45	10	0.0	-0.003	0.51	0.513	0.622	0.419	0.857	0.558
35	50	15	1.0	0.0	0.0	0.51	0.507	0.583	0.872	0.66
40	5	45	0.003	0.011	0.51	0.505	0.533	0.419	0.979	0.802
40	50	10	0.0	-0.003	0.51	0.513	0.646	0.435	0.875	0.571
45	5	50	0.003	0.003	0.51	0.51	0.537	0.452	0.98	0.822

Table 4.2. Result on simulated dataset with hierarchical clustering

Hypocenters			Adjusted Rand I.		Mirkin Index		Homogeneity		Separation	
distance	I.depth	II.depth	c.s.	xcorr	c.s.	xcorr	c.s.	xcorr	c.s.	xcorr
5	5	10	0.098	0.068	0.456	0.472	0.486	0.476	0.793	0.596
5	15	10	0.622	0.334	0.189	0.335	0.495	0.577	0.65	0.461
5	15	20	0.708	0.217	0.146	0.391	0.562	0.588	0.63	0.552
5	25	20	0.395	-0.026	0.302	0.513	0.566	0.501	0.73	0.513
5	25	30	0.396	0.125	0.302	0.437	0.584	0.624	0.695	0.669
5	30	35	0.8	0.009	0.1	0.497	0.593	0.534	0.831	0.552
5	40	35	0.708	-0.018	0.146	0.51	0.61	0.552	0.796	0.6
5	40	45	0.708	0.009	0.146	0.497	0.657	0.541	0.787	0.641
5	50	45	0.272	0.126	0.364	0.437	0.645	0.618	0.77	0.721
10	5	15	0.04	0.04	0.486	0.486	0.469	0.422	0.794	0.617
10	20	10	0.332	0.009	0.335	0.497	0.487	0.438	0.653	0.435
10	20	30	0.708	0.168	0.146	0.416	0.564	0.511	0.77	0.51
10	25	15	0.622	0.007	0.189	0.497	0.569	0.517	0.777	0.534
10	25	35	0.8	0.056	0.1	0.472	0.534	0.521	0.79	0.53
10	40	30	0.8	0.029	0.1	0.486	0.578	0.638	0.771	0.64
10	45	35	0.622	0.029	0.189	0.486	0.64	0.633	0.796	0.652
10	50	40	0.897	0.168	0.051	0.416	0.615	0.655	0.839	0.679
15	5	20	0.04	0.04	0.486	0.486	0.461	0.401	0.793	0.609
15	15	30	0.8	0.125	0.1	0.437	0.537	0.532	0.781	0.531
15	20	35	0.8	0.217	0.1	0.391	0.517	0.558	0.756	0.553
15	25	10	0.332	0.056	0.335	0.472	0.472	0.41	0.663	0.38
15	25	40	0.8	0.055	0.1	0.472	0.514	0.613	0.745	0.636
15	30	45	0.8	-0.008	0.1	0.505	0.608	0.558	0.852	0.558
15	50	35	0.897	0.125	0.051	0.437	0.625	0.651	0.844	0.675
20	5	25	0.04	0.04	0.486	0.486	0.46	0.391	0.793	0.606
20	15	35	0.8	0.396	0.1	0.302	0.508	0.625	0.75	0.556
20	20	40	0.8	0.087	0.1	0.456	0.522	0.52	0.812	0.548
20	25	45	0.8	0.055	0.1	0.472	0.582	0.544	0.841	0.546
20	30	10	0.396	0.126	0.302	0.437	0.487	0.395	0.682	0.413
20	50	30	0.897	0.031	0.051	0.486	0.553	0.611	0.852	0.614
25	5	30	0.095	0.04	0.456	0.486	0.441	0.393	0.772	0.607
25	15	40	0.8	-0.007	0.1	0.505	0.515	0.549	0.814	0.558
25	20	45	0.8	0.217	0.1	0.391	0.581	0.569	0.852	0.534
25	35	10	0.541	0.467	0.229	0.267	0.527	0.518	0.714	0.341
25	50	25	0.897	0.088	0.051	0.456	0.517	0.595	0.818	0.602
30	5	35	0.04	0.04	0.486	0.486	0.466	0.409	0.795	0.611
30	15	45	0.8	0.396	0.1	0.302	0.581	0.609	0.856	0.52
30	40	10	0.8	0.221	0.1	0.391	0.516	0.469	0.766	0.353
30	50	20	0.897	0.217	0.051	0.391	0.532	0.578	0.832	0.577
35	5	40	0.132	0.04	0.437	0.486	0.438	0.398	0.765	0.609
35	45	10	0.708	0.467	0.146	0.267	0.541	0.506	0.745	0.327
35	50	15	0.897	0.465	0.051	0.267	0.546	0.626	0.835	0.556
40	5	45	0.04	0.04	0.486	0.486	0.468	0.403	0.797	0.608
40	50	10	0.897	0.467	0.051	0.267	0.552	0.521	0.78	0.35
45	5	50	0.04	0.04	0.486	0.486	0.474	0.407	0.799	0.647

Table 4.3. Result on simulated dataset with partitional clustering

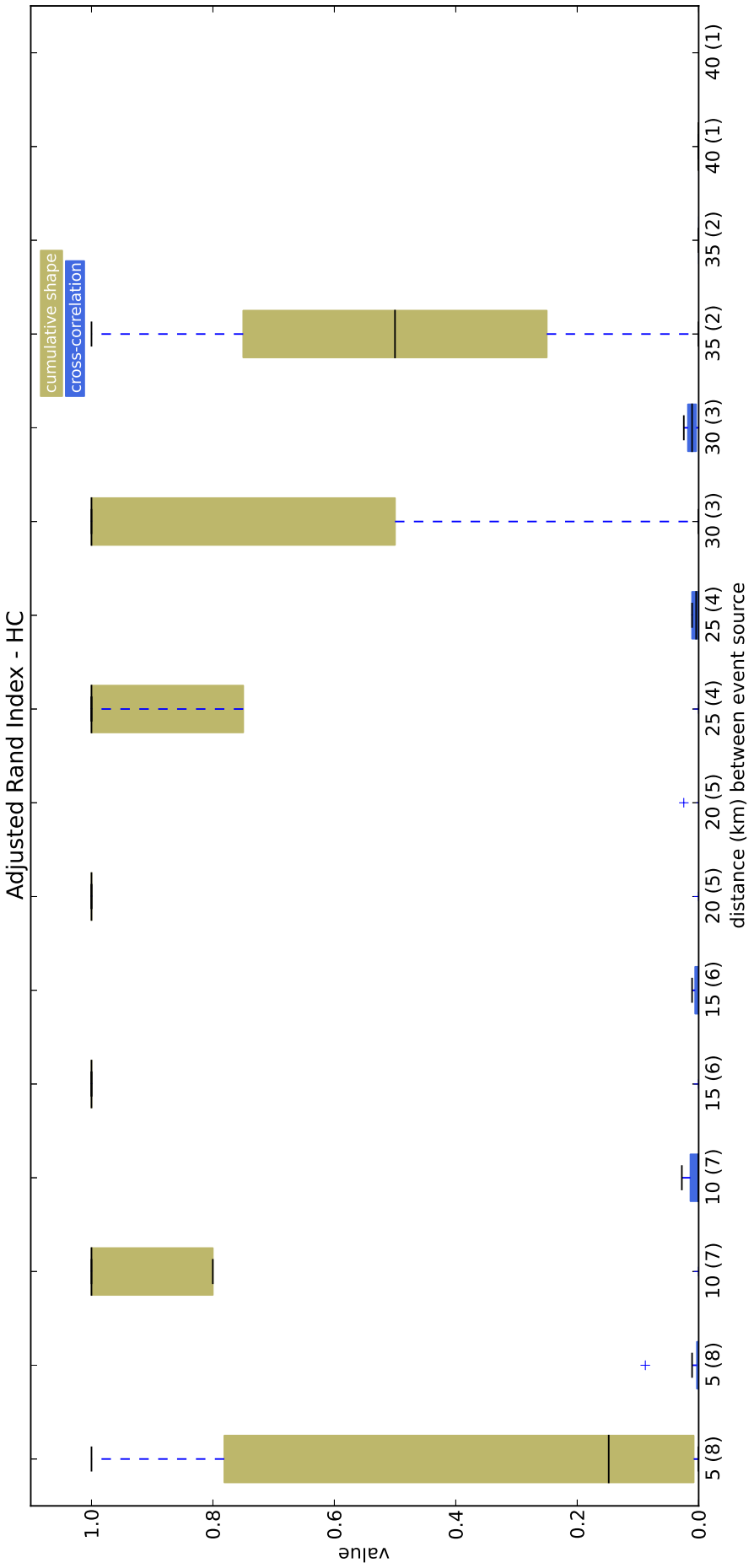


Figure 4.7. Adjusted Rand Index on simulated dataset with hierarchical clustering (depth 5km not included)

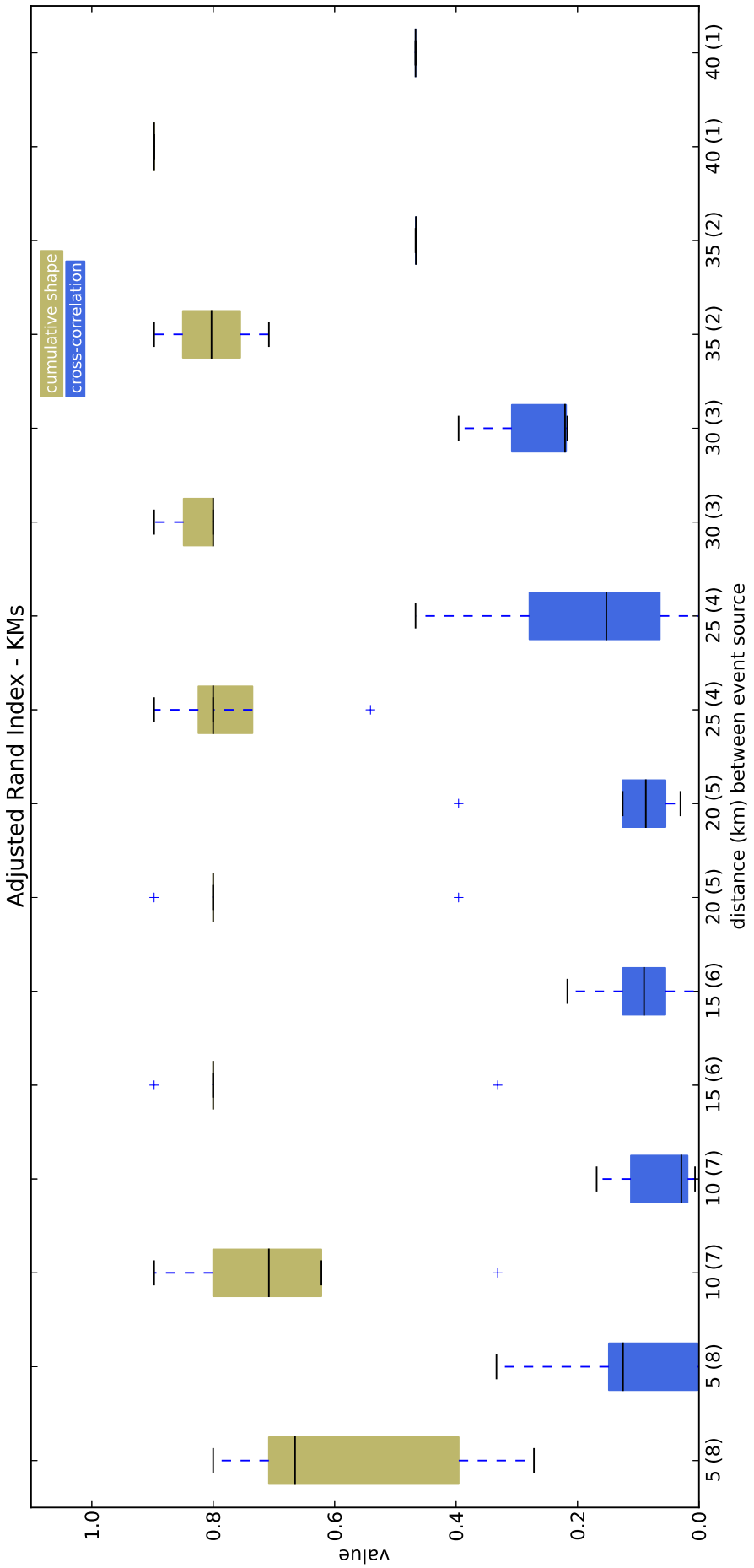


Figure 4.8. Adjusted Rand Index on simulated dataset with partitional clustering k-medoids (depth 5km not included)

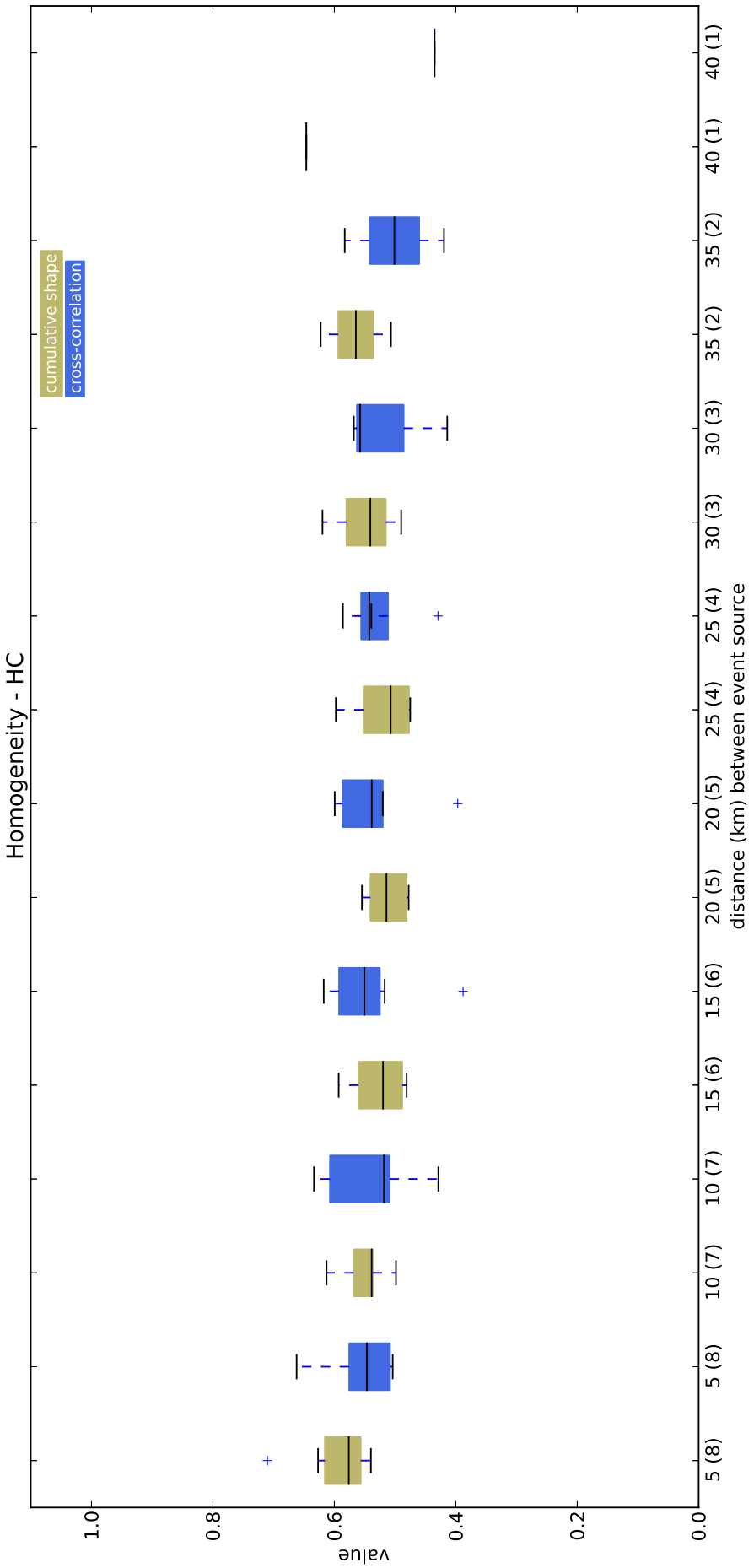


Figure 4.9. Homogeneity index on simulated dataset with hierarchical clustering (depth 5km not included)

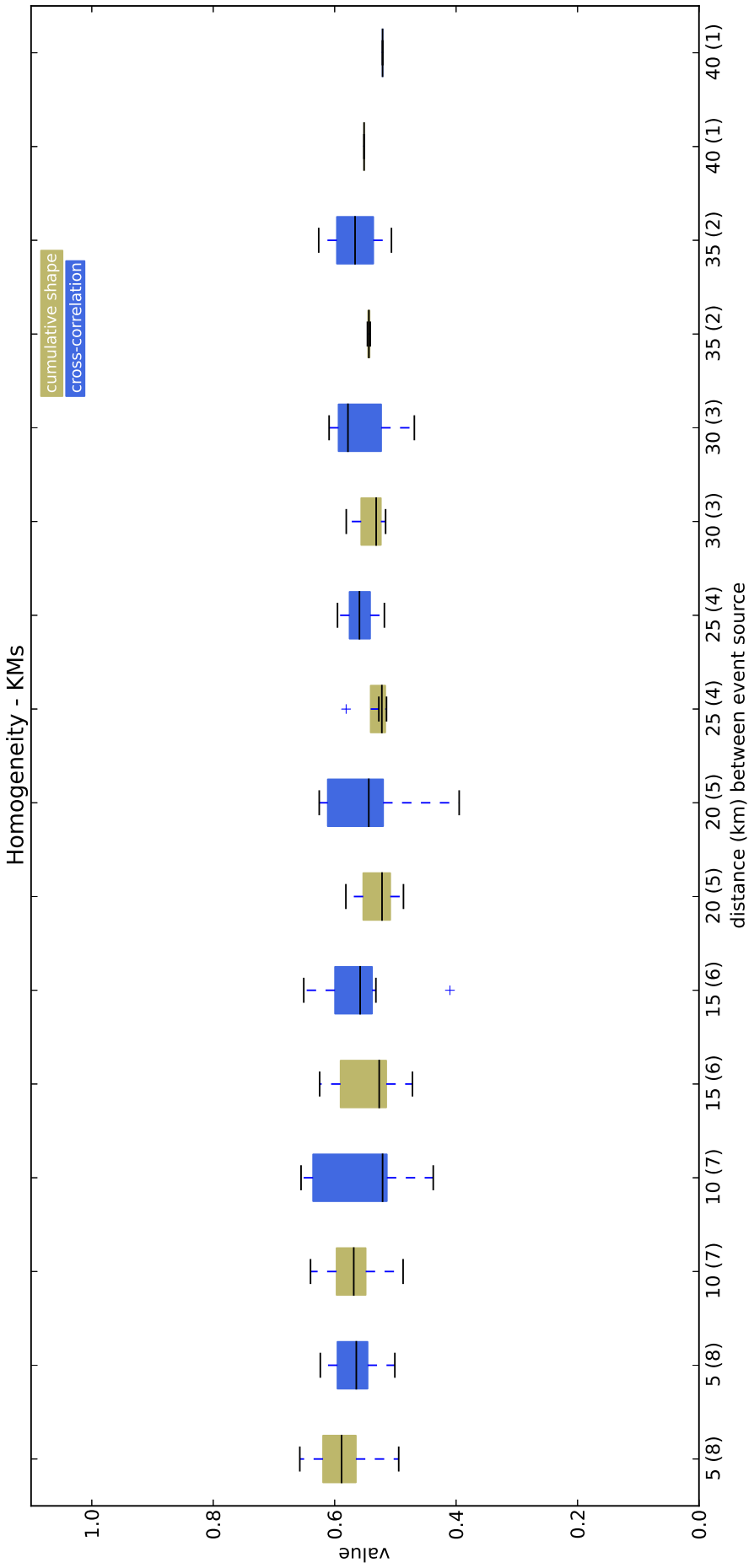


Figure 4.10. Homogeneity index on simulated dataset with k-medoids clustering (depth 5km not included)

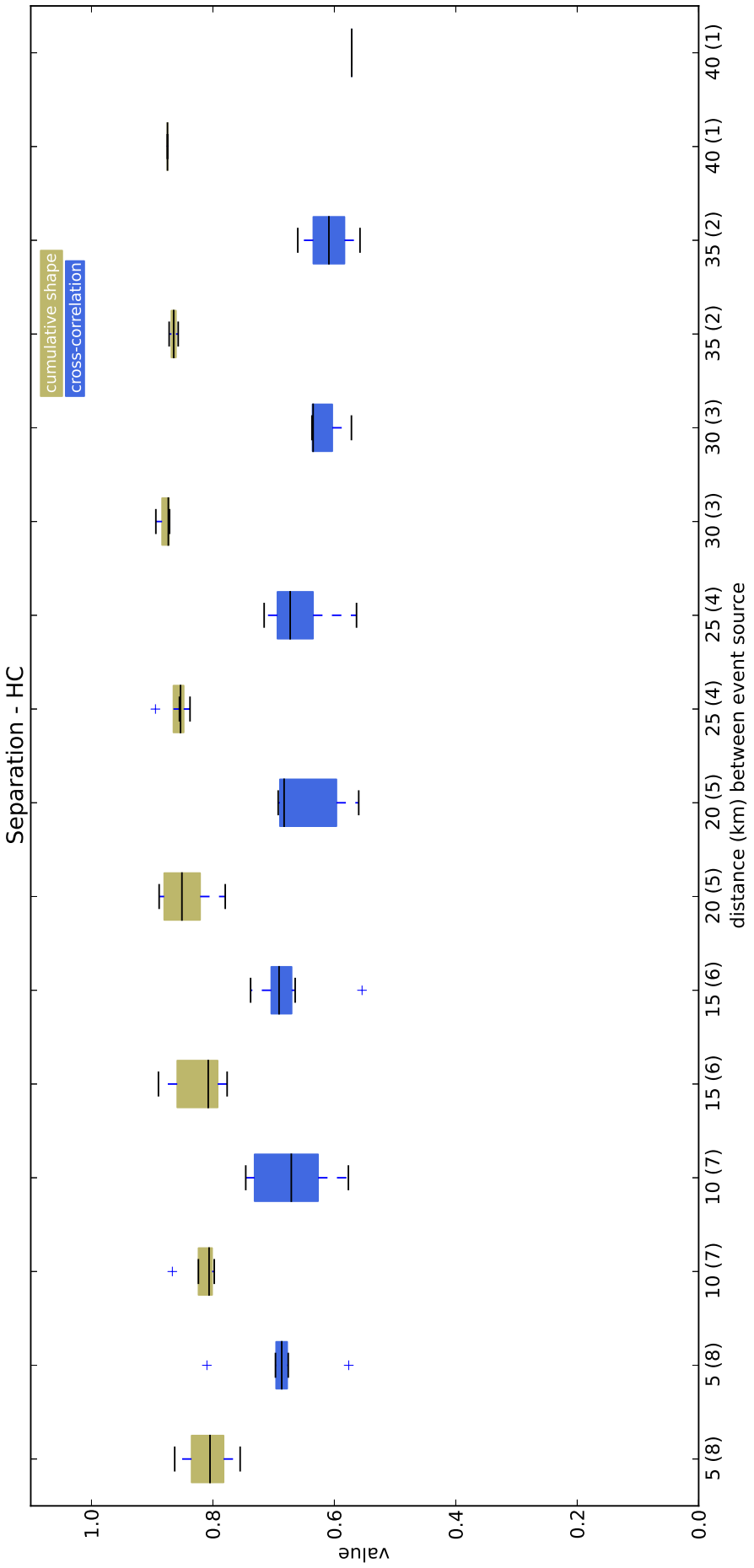


Figure 4.11. Separation index on simulated dataset with hierarchical clustering (depth 5km not included)

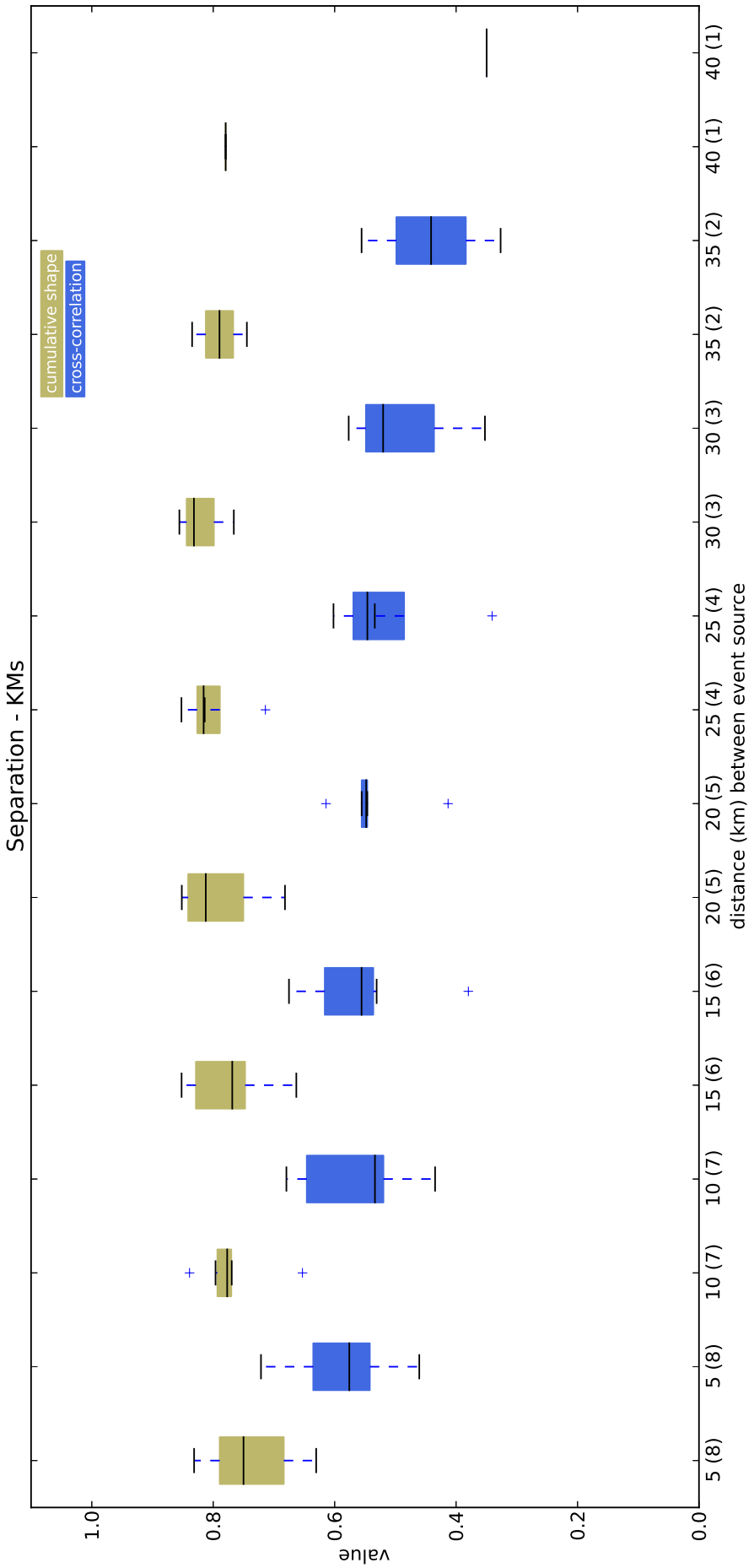


Figure 4.12. Separation index on simulated dataset with k-medoids clustering (depth 5km not included)

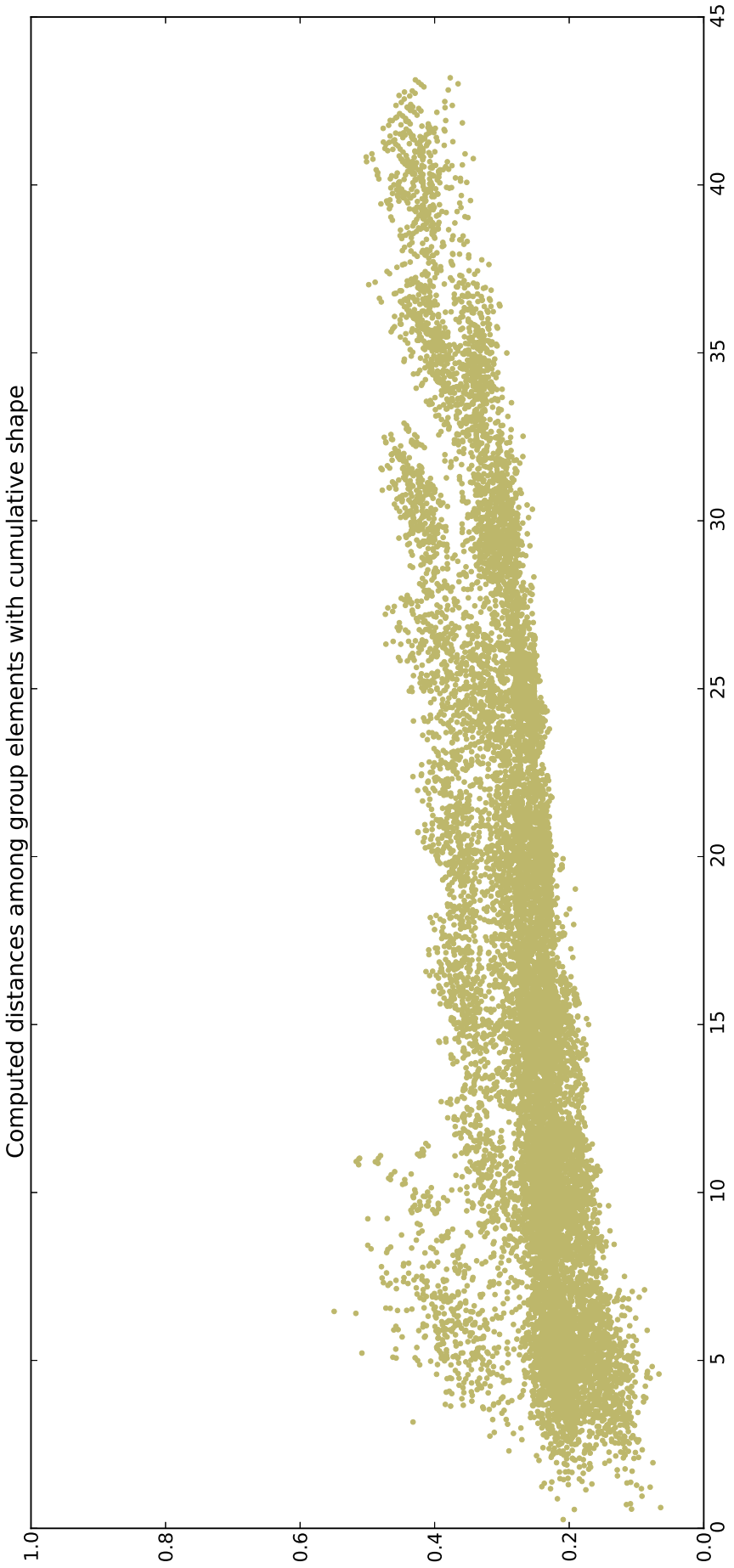


Figure 4.13. Cumulative shape distribution respect to the physical distance (depth 5km not included)

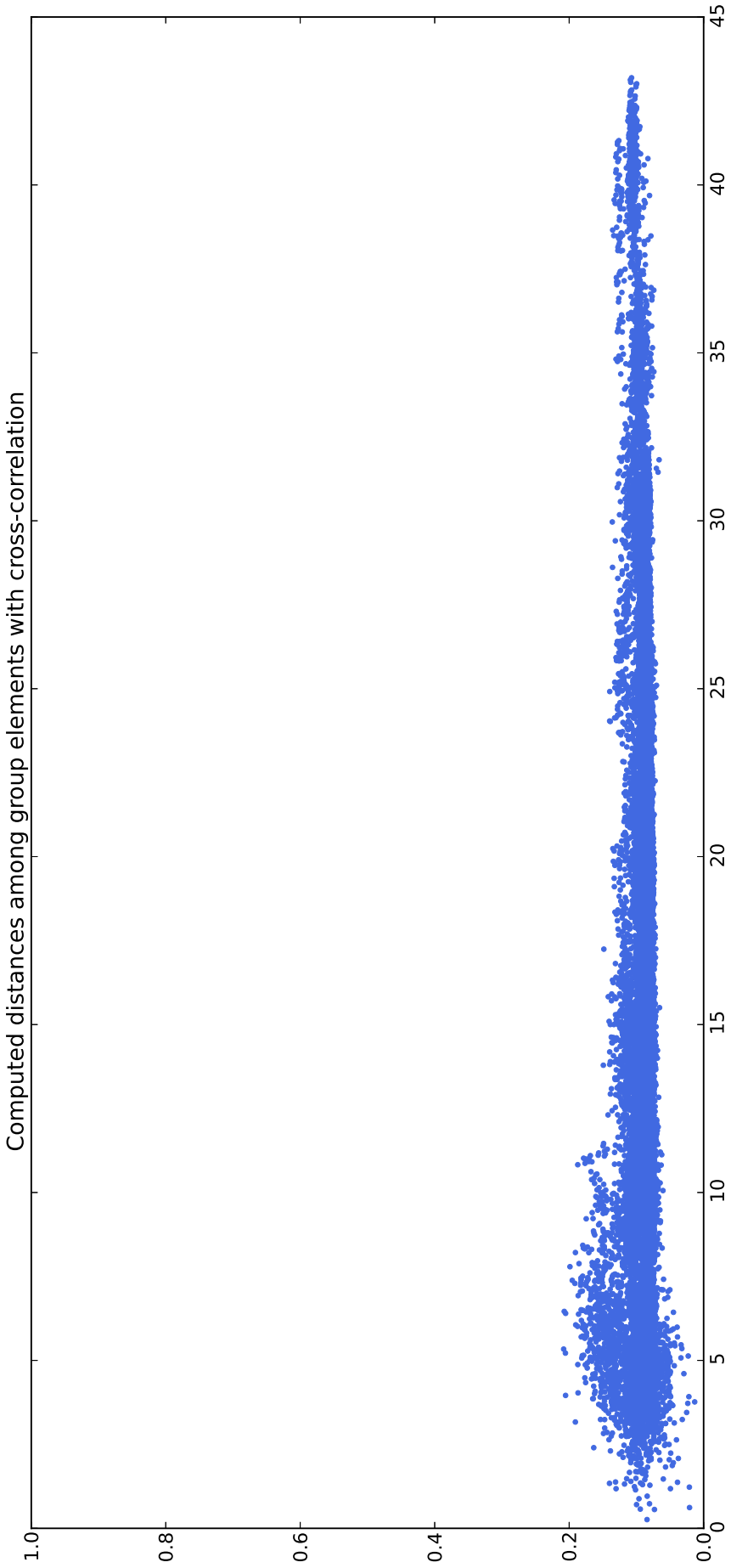


Figure 4.14. Cross-correlation distance distribution respect to the physical distance (depth 5km not included)

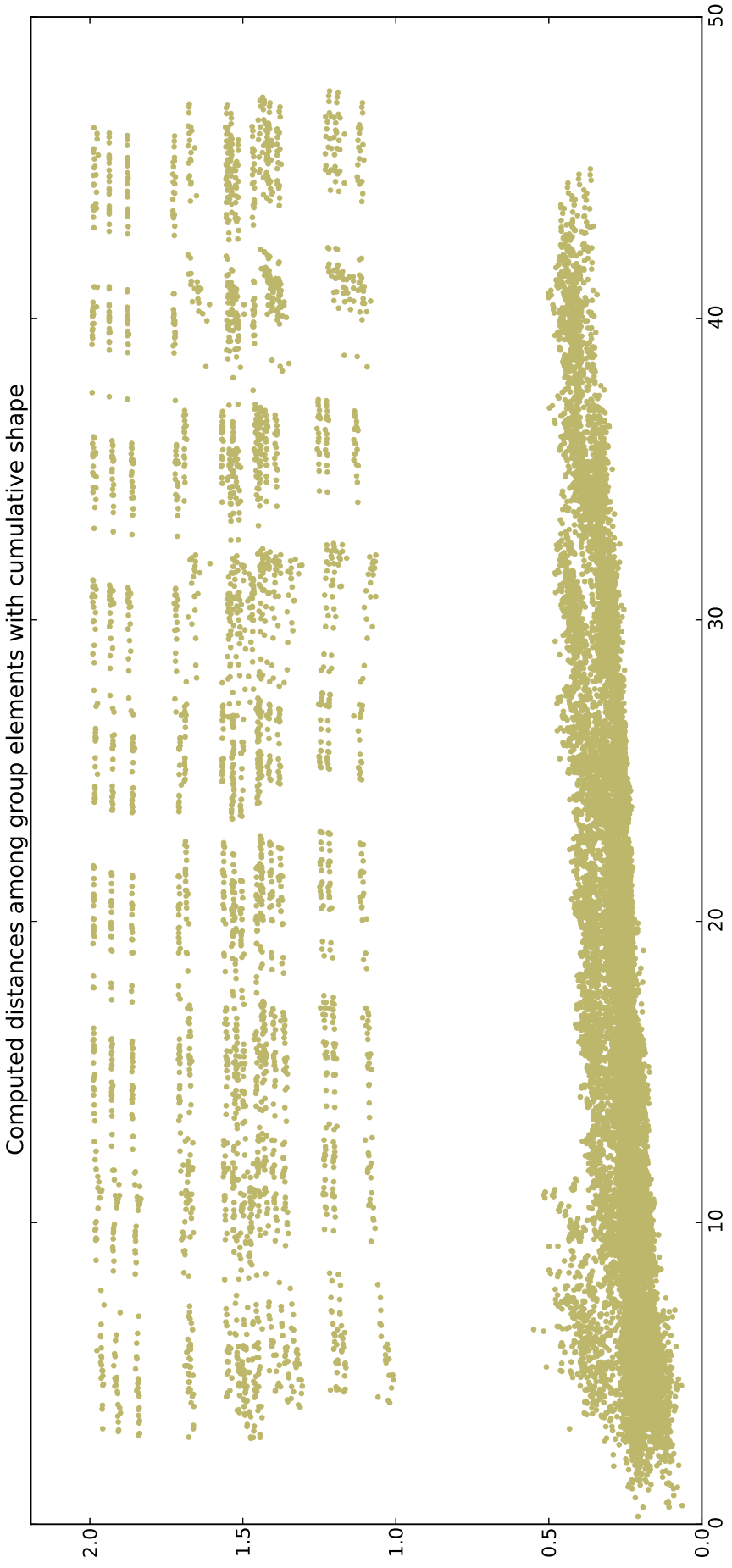


Figure 4.15. Cumulative shape distribution respect to the physical distance

4.2.2 Results on bursts dataset

In order to test the relative merit of each distance over the bursts dataset we cutted the signals to a size useful to catch the meaningful part of the simulated burst. In particular we considered the first 1000 points of each signal because this part has an higher signal to noise ratio. The performance of dissimilarities on this dataset has been measured by using a custom index: the Dissimilarity Optimality index. This is due to the fact that the experiment involve that signals recorded at closer instant times, should reveal similar shapes.

The values of the distance optimality index for the cross correlation dissimilarity and the cumulative shape dissimilarity are 0.0033 and 0.0071 respectively.

Both values are very close to 0 and their difference is very small.

We have also studied how the distance optimality index changes in terms of a temporal window w. In particular, for each signals x_i recorded at instant time t_i, we have computed the rate of how many times its closest signal x_j with $j = \underset{1 \leq k \leq N, k \neq i}{\operatorname{argmin}} \delta(x_i, x_k)$ falls into a temporal window w, i.e $|t_j - t_i| \leq w$. We indicate this rate as *coverage proximity*. Figure 4.17 shows its computation for w ranging from 1 until 17.

The results show that cumulative shapes have a coverage proximity of 80% vs 88% of cross correlation (8% difference) for $w = 1$. Anyway, this difference decreases very fast to 1% for $w > 1$. We can conclude that the performances of the two measures over the bursts dataset are almost equal.

4.2.3 Results on Palermo earthquake dataset

This dataset is composed by 95 signals of length 3000 sample points. The performance dissimilarities on this dataset has been measured by using the Homogeneity, Separation and Adjusted Rand indices. This is due to the fact that the we dispose of a gold solution established by experts taking into consideration both their knowledge about the phenomena and the result of a hierarchical clustering algorithm using cross correlation dissimilarity. In particular, the spatial distribution of the hypocenters of the acquired data, suggests at least four well separated hypocenters clouds, close to the Palermo 2002 cluster [GLT+09]. This 4 clusters, had been split into 9 clusters with a variable number of events, by using the average link clustering algorithm in conjunction with the cross-correlation dissimilarity. The same clustering algorithm has been used to compute all the indices since it has been adopted by the expert to establish the gold solution. The first result is that the partitioning computed by the average link clustering in conjunction with the cumulative shape dissimilarity is perfectly equal to the gold solution (adjusted rand index equal to 1). Moreover, in order to better characterize this partitioning, we have computed its homogeneity and separation.

We report in figure 4.18 and 4.19 the homogeneity and separation indices of the two dissimilarities for different partitionings of K clusters ranging between 2 and 20.

The results show that the cumulative shape outperforms the cross-correlation in term of homogeneity and performs almost equally on separation.

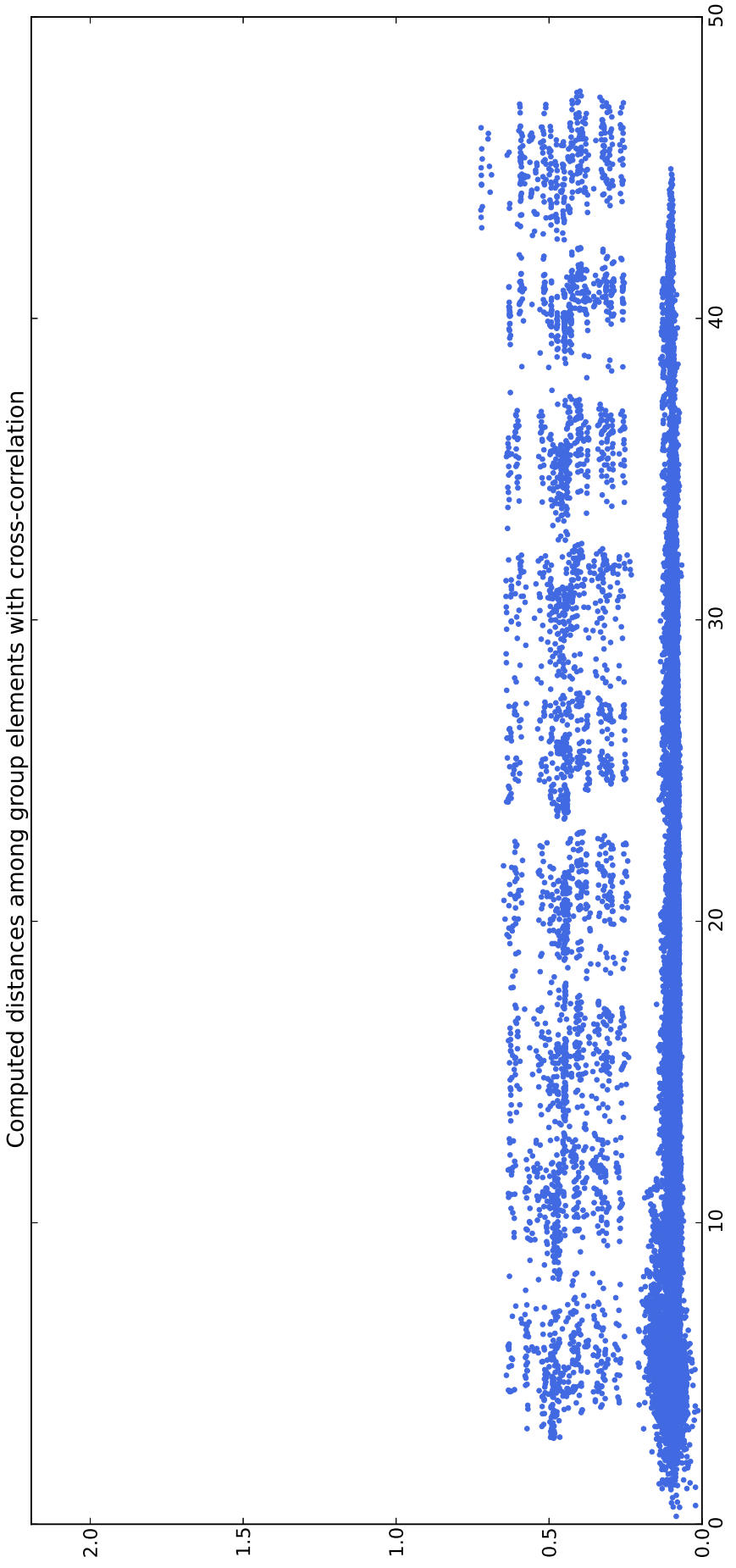


Figure 4.16. Cross-correlation distance distribution respect to the physical distance

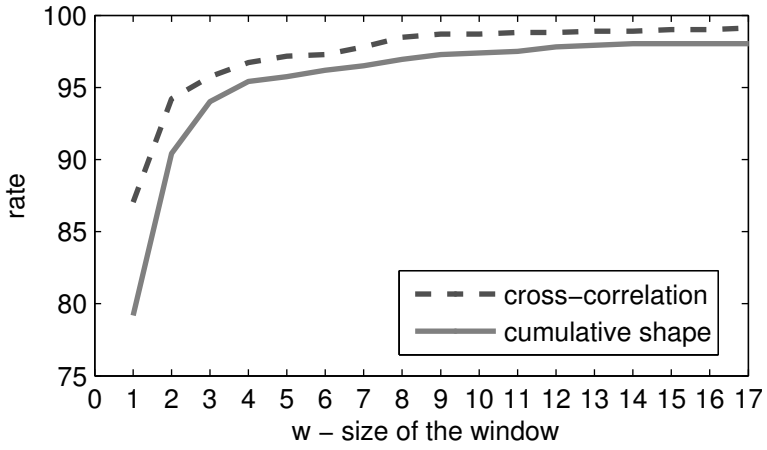


Figure 4.17. Diagram of coverage proximity for w between 1 and 17 in the bursts dataset

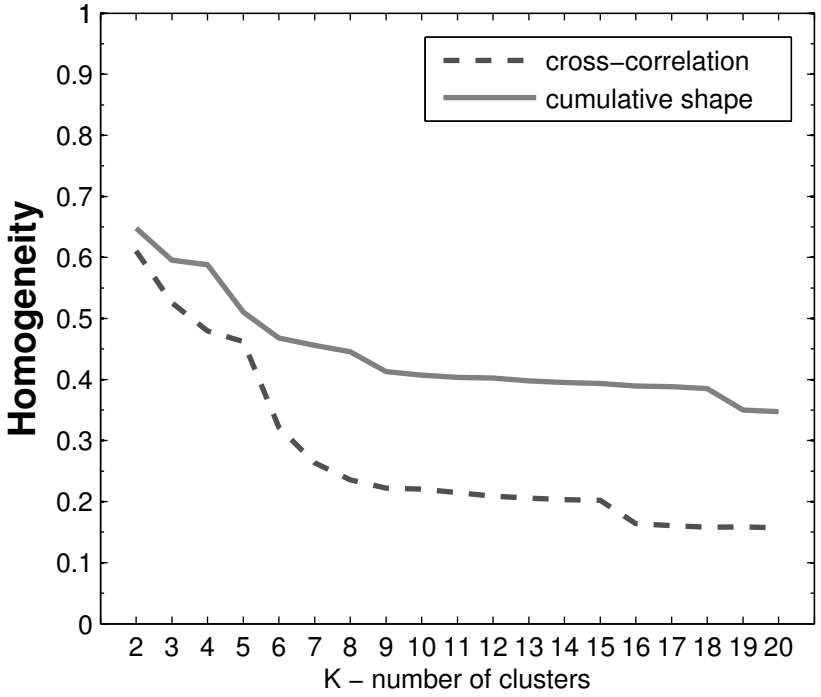


Figure 4.18. Internal Homogeneity index for Palermo earthquake dataset

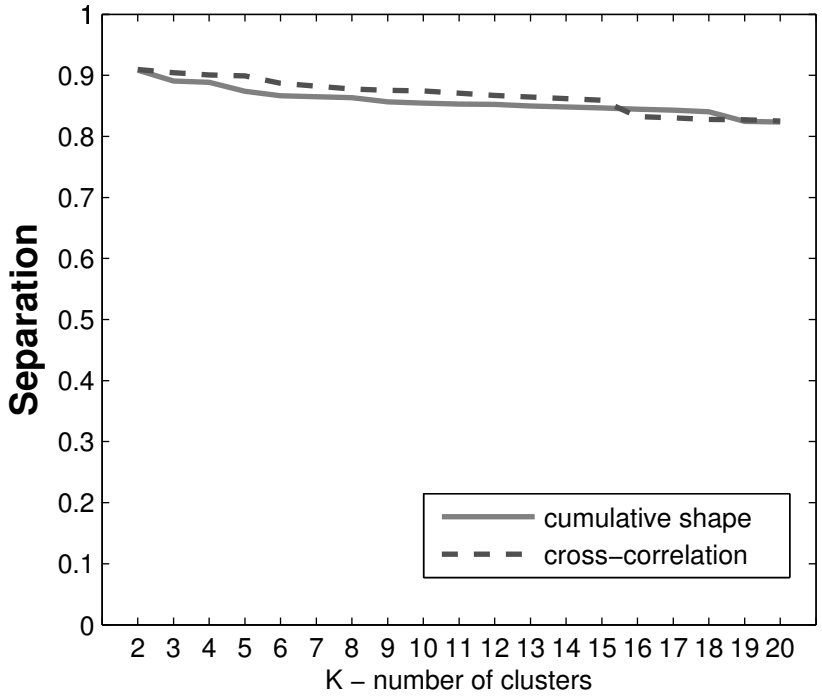


Figure 4.19. Internal Separation index for Palermo earthquake dataset

Chapter 5

Conclusion

In this thesis, a new dissimilarity measure between seismic signals called cumulative shape dissimilarity has been proposed. A number of tests have been done on three different dataset of earthquake events. These events are natural and artificial. The natural are recorded by a national network of detection stations in Italy while the artificial are simulated in a real and in fully virtual environment.

The first has a gold solution proposed by an expert providing 9 cluster with a variable number of elements. The second is characterized by synthetic signal without gold solution in spite of the third that, due to its fully simulated nature, it has a known solution of 10 cluster defined by the simulation model.

Such datasets have been used to compare the cumulative shape dissimilarity with the cross correlation dissimilarity, that is actually largely adopted to differentiate waveforms in the context of seismic signals.

In order to evaluate the goodness of the proposed measure, due to the heterogeneity of the two dataset, several indices have been considered (Dissimilarity Optimality, Homogeneity, Separation and Adjusted Rand). Furthermore for the third dataset the ratio measures/distance is drawn to show how the cumulative shape is more compliant to physics model respect to cross-correlation based distance.

The test returns that the proposed measure have Dissimilarity Optimality and a Separation indices almost equal to the cross correlation ones, and a superior Homogeneity for all clusters values ranging from 2 to 20 (in average 1%). While the differences on performance are small on first and second dataset, on the third the cumulative shape exceeds with a large margin the cross-correlation distance. On simulated environment the cumulative shape is able in a great number of cases to exploit the correct cluster solution while the other not.

Anyway, the relevant difference has to be noted on the computational time, in particular cumulative shape measure is faster than cross-correlation ($O(n)$ vs $O(n^2)$). Future developments will be devoted to an extension of the cumulative shape on all the three-component signals, a new version taking into account weights for the signal samples, and to the study of the better conjunction between the new proposed dissimilarity and several kind of clustering algorithms.

Part I

Appendix

Appendix A

Simulated Dataset

In the following pages we show the seismograms of the simulated dataset. These signals are generated in a virtual environment defined by the model defined in table 4.1.

The dataset is composed by 12 group of 20 elements. The full dataset is composed by 240 simulated signals. Each group was generated at several depths with two types of source: explosive and fault.

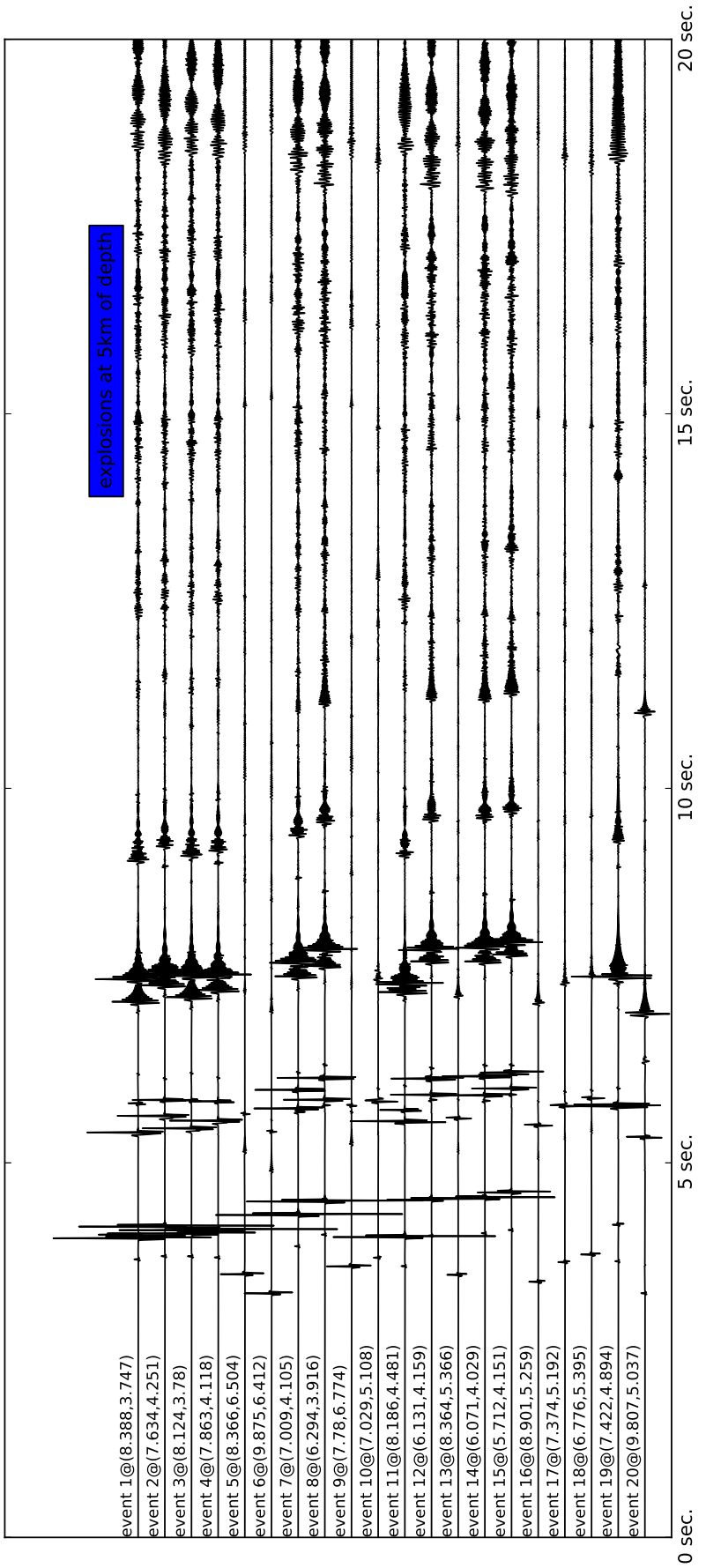


Figure A.1. Simulated dataset with explosive source at 5km of depth

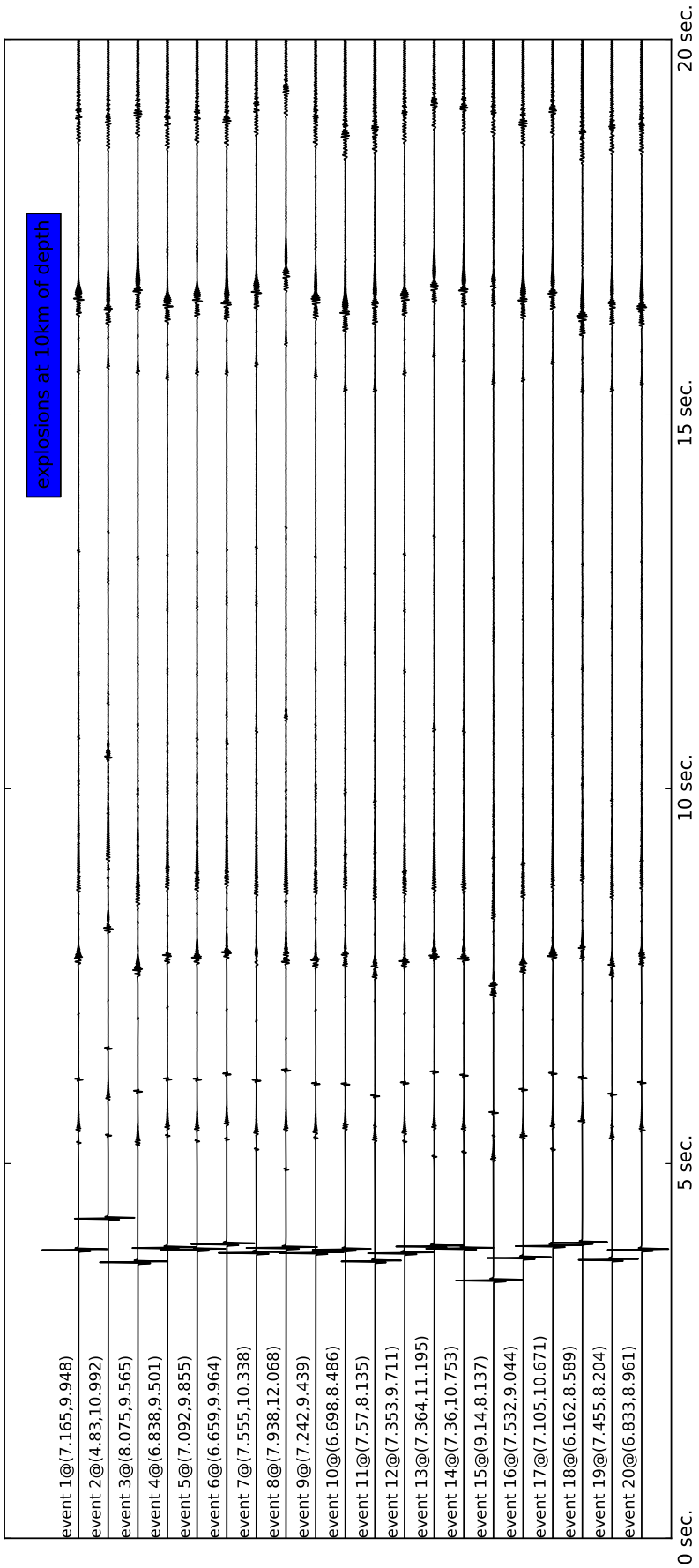


Figure A.2. Simulated dataset with explosive source at 10km of depth

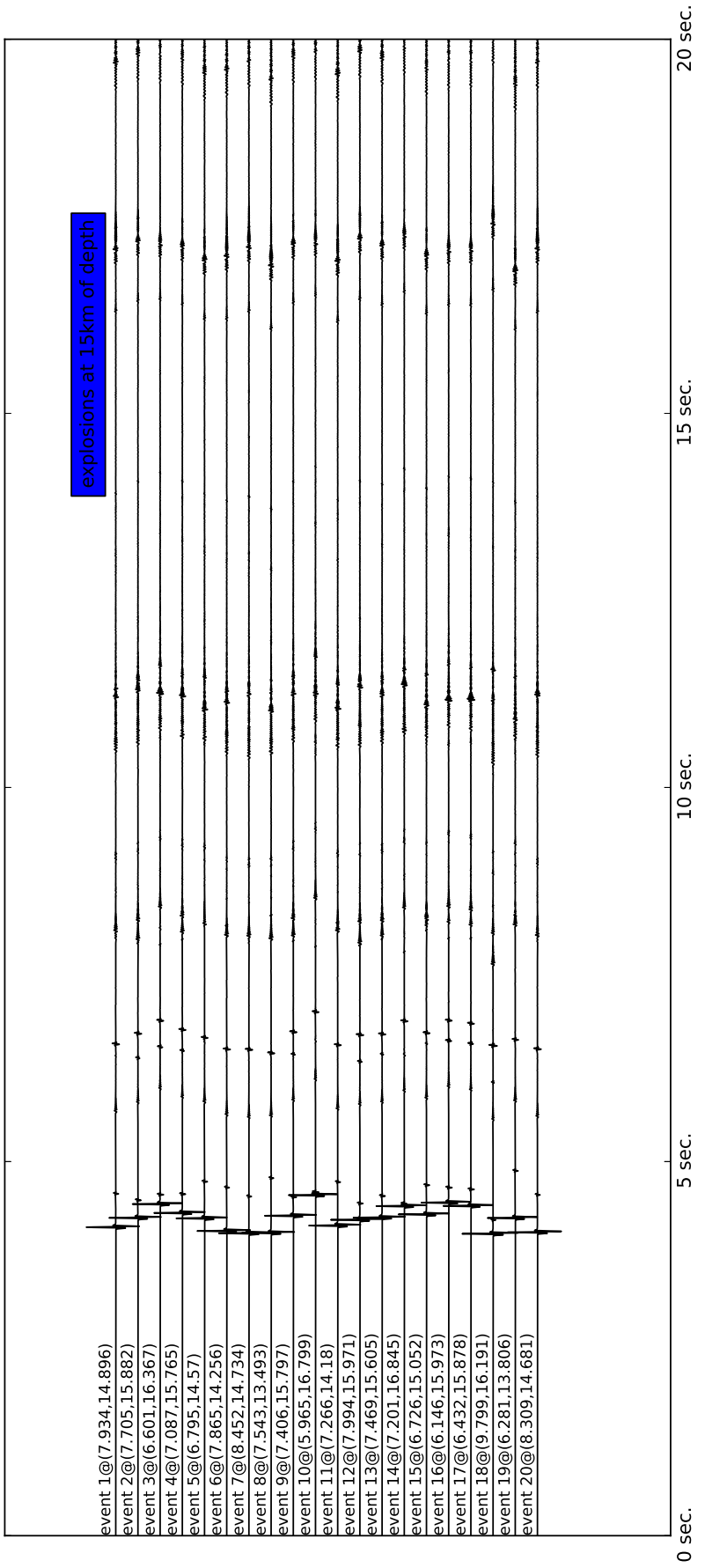


Figure A.3. Simulated dataset with explosive source at 15km of depth

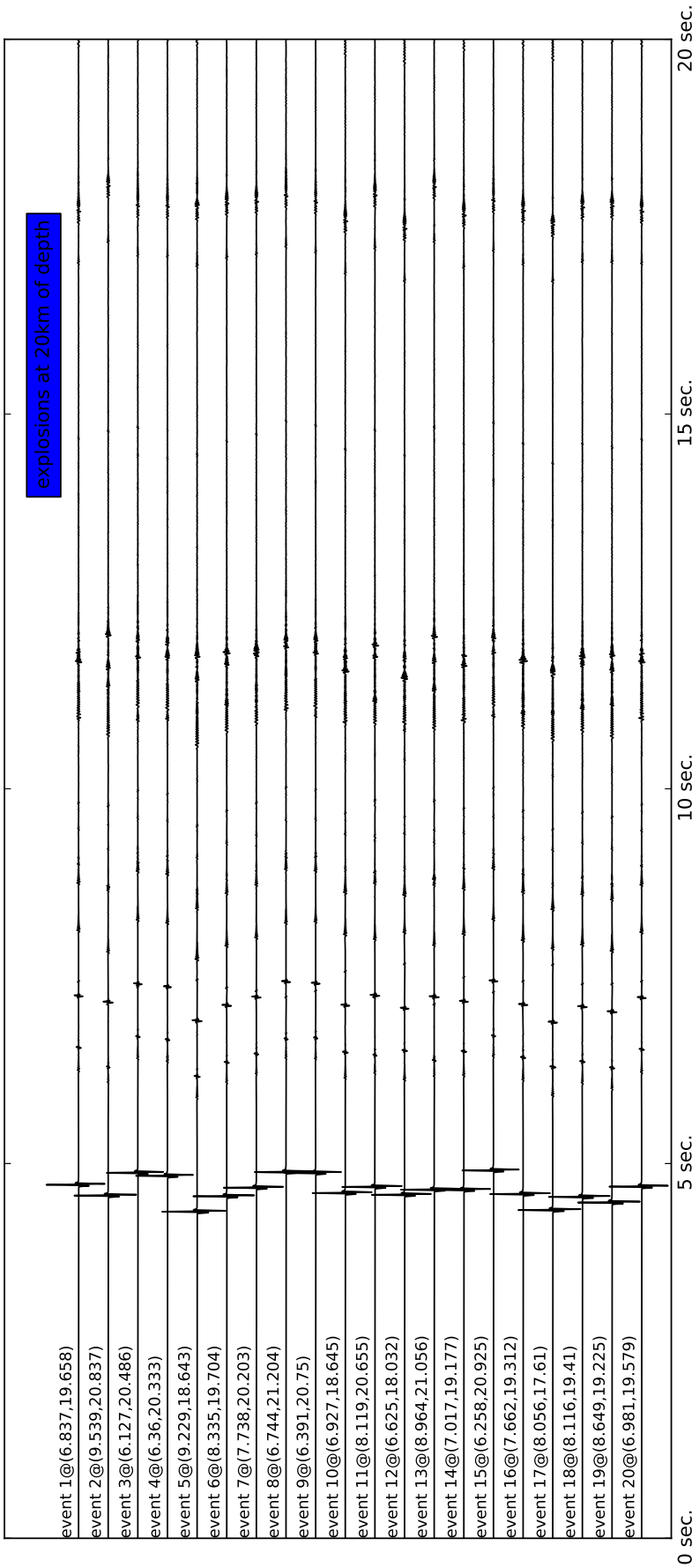


Figure A.4. Simulated dataset with explosive source at 20km of depth

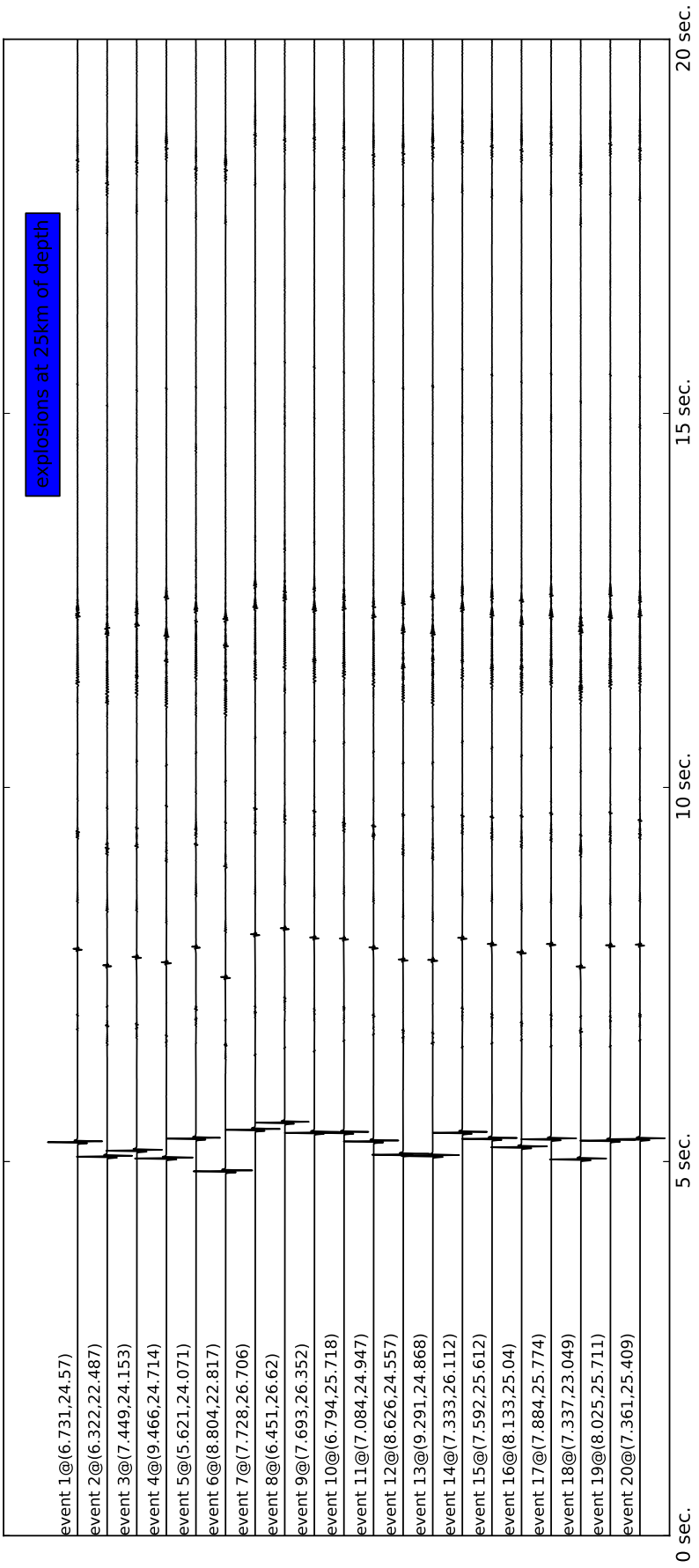


Figure A.5. Simulated dataset with explosive source at 25km of depth

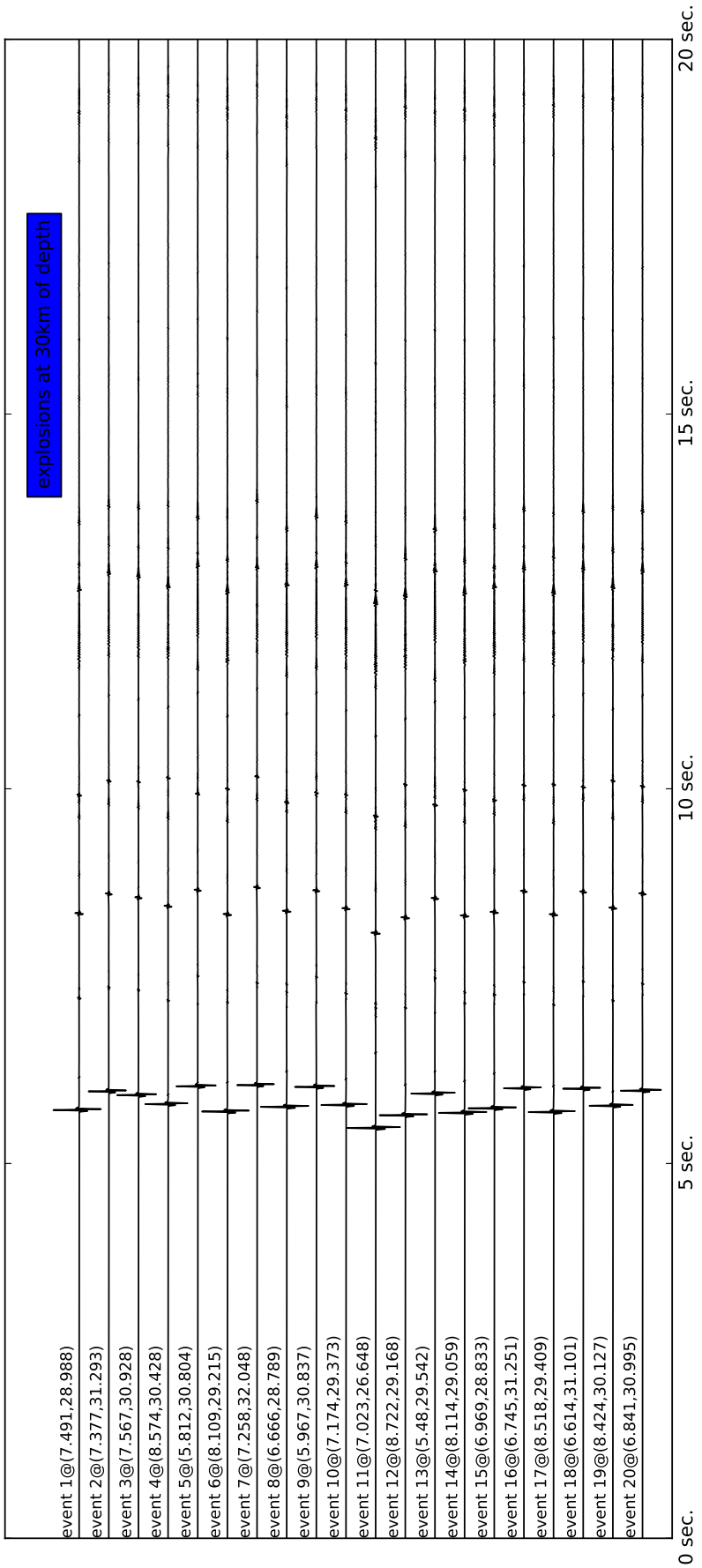


Figure A.6. Simulated dataset with explosive source at 30km of depth

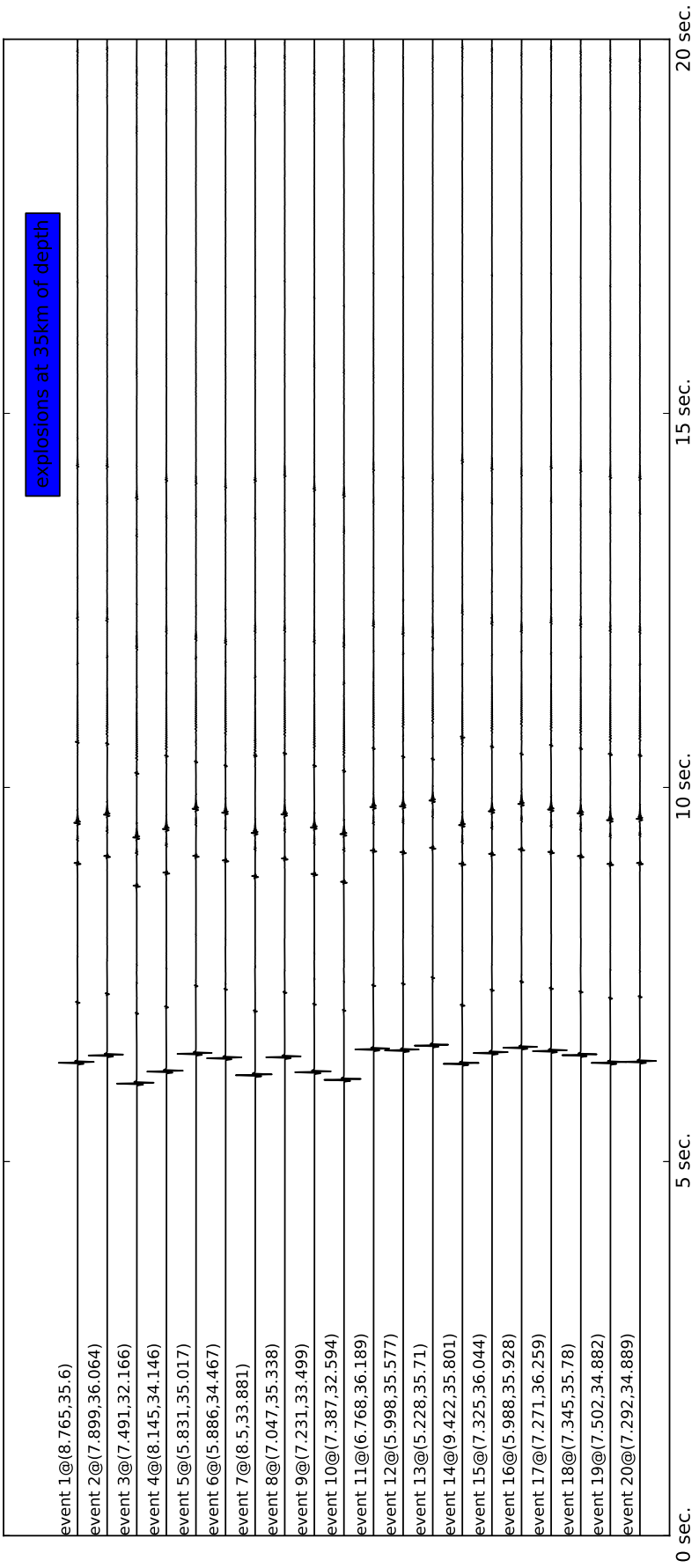


Figure A.7. Simulated dataset with explosive source at 35km of depth

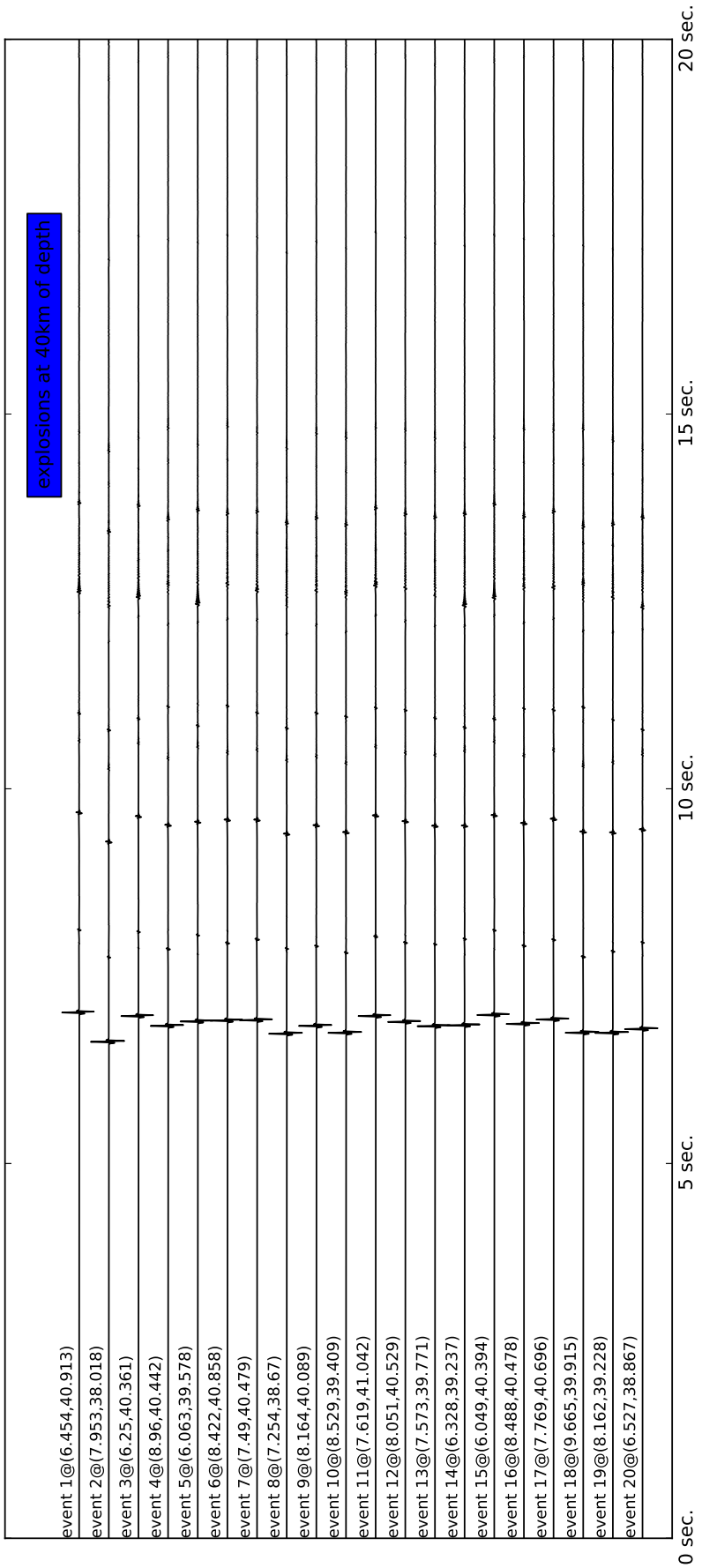


Figure A.8. Simulated dataset with explosive source at 40km of depth

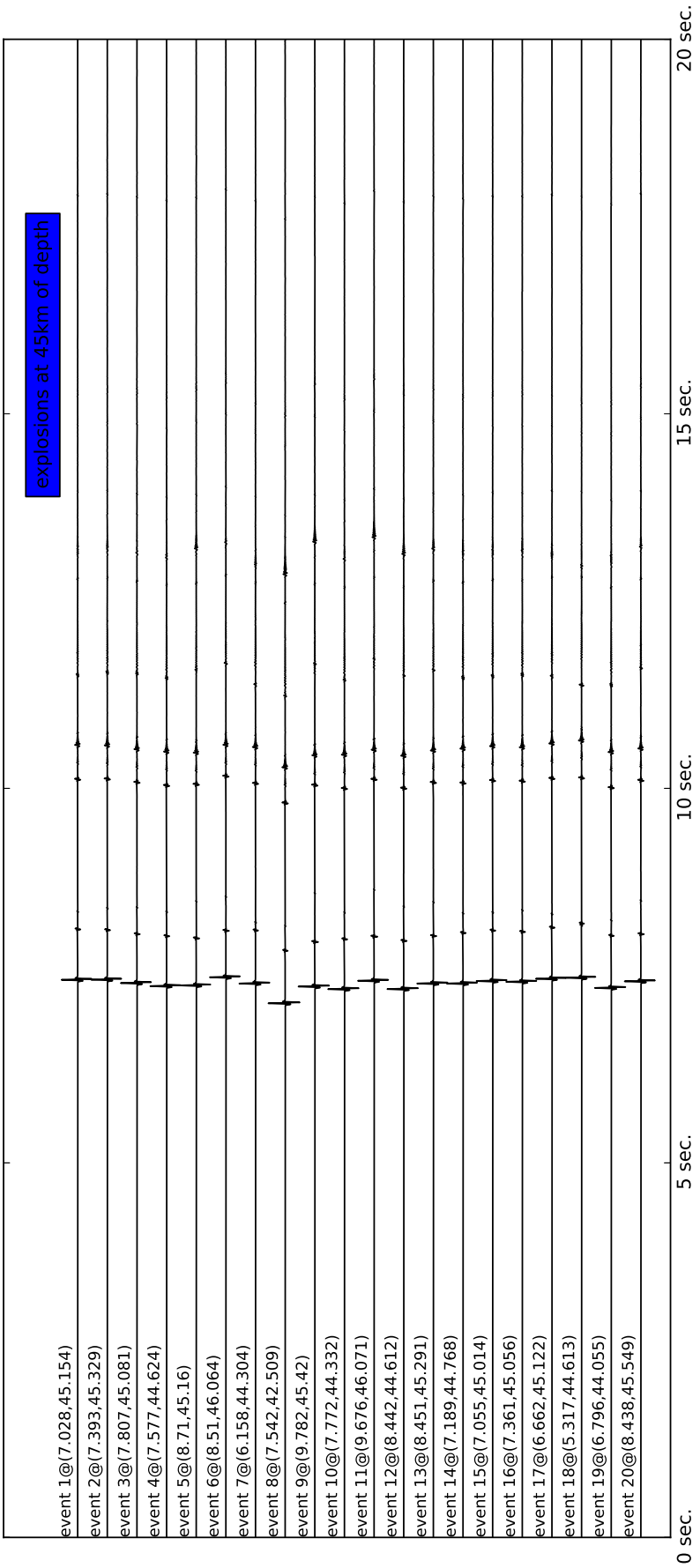


Figure A.9. Simulated dataset with explosive source at 45km of depth

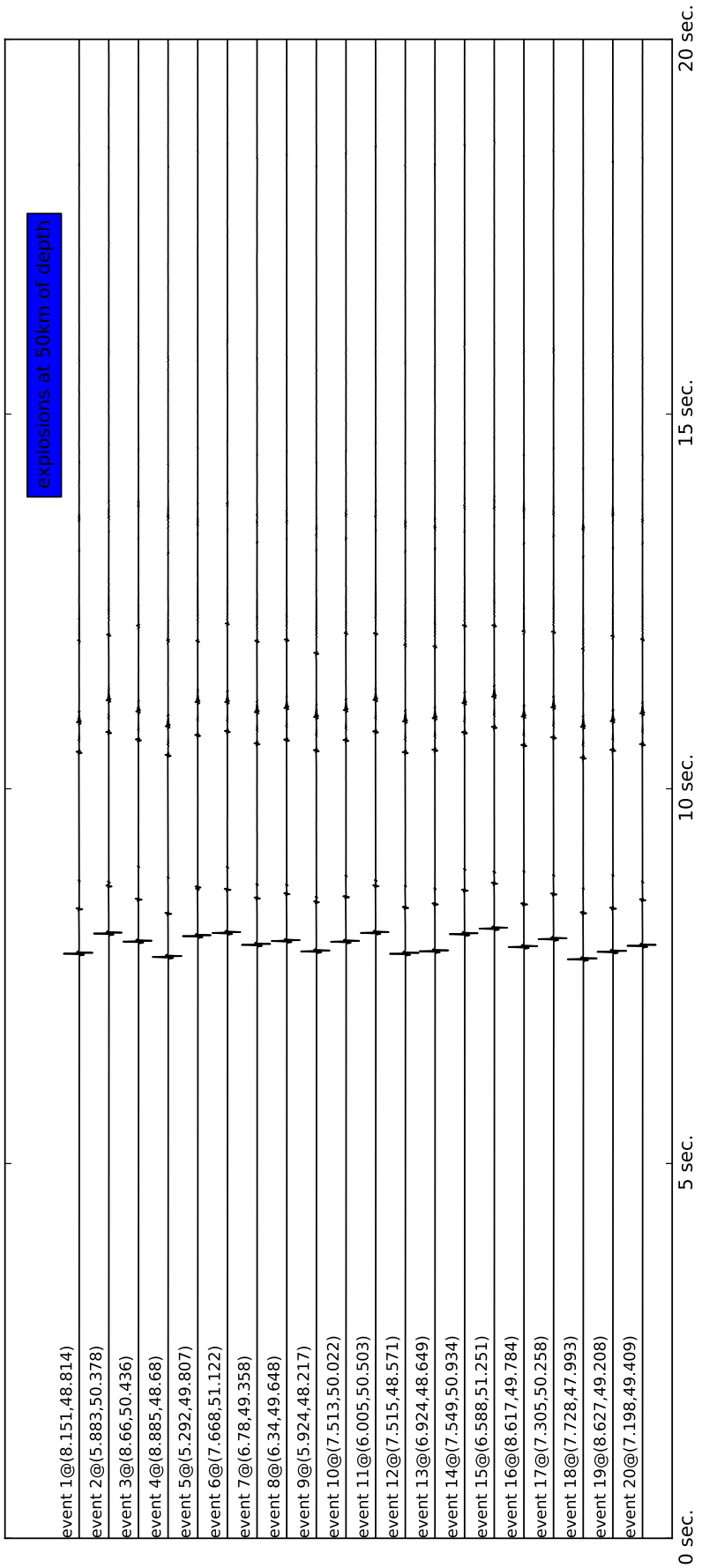


Figure A.10. Simulated dataset with explosive source at 50km of depth

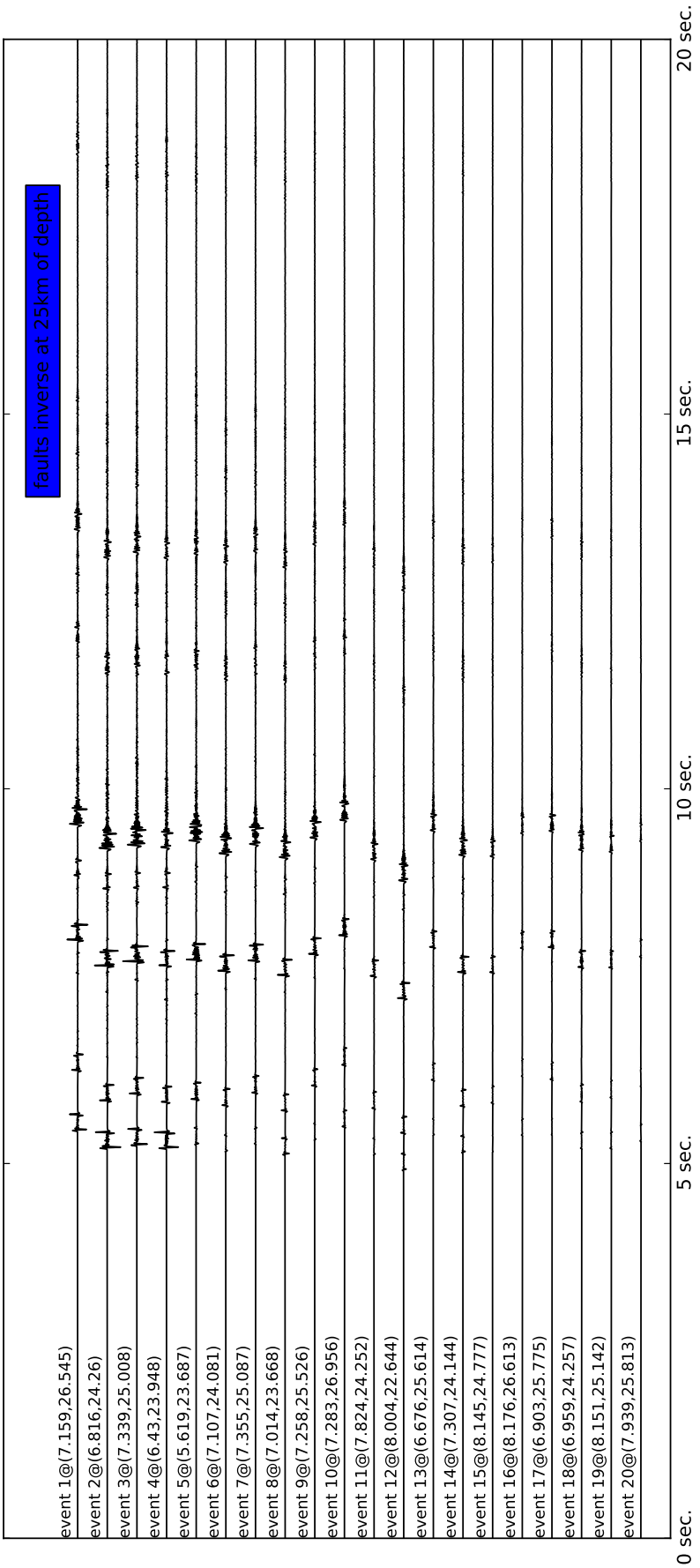


Figure A.11. Simulated dataset with inverse fault source at 25km of depth

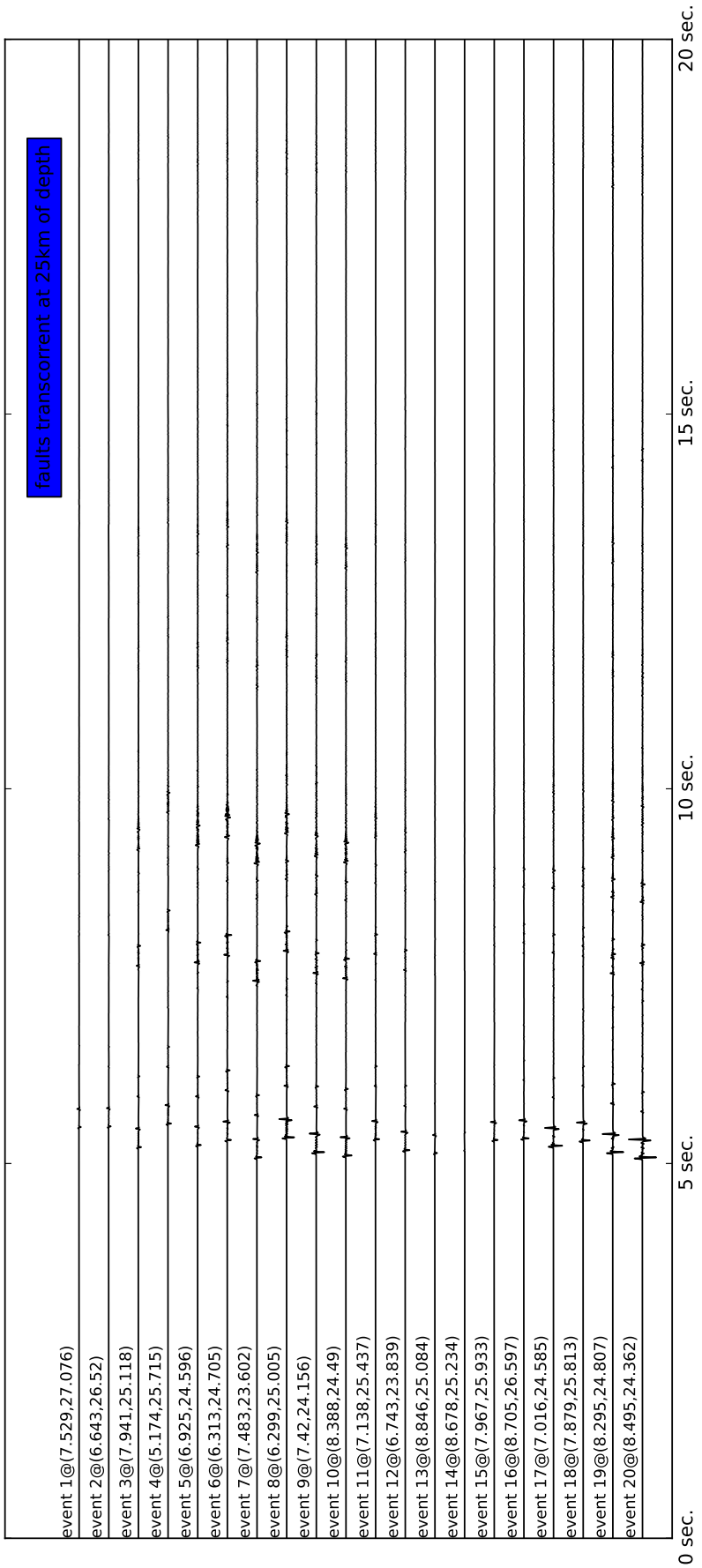


Figure A.12. Simulated dataset with strike fault source at 25km of depth

Appendix B

Computing Infrastructure

To perform simulation and test we used a computing infrastructure based on an high performance cluster. High-performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably and quickly. The term applies especially to systems that function above a teraflop or 1012 floating-point operations per second. The term HPC is occasionally used as a synonym for supercomputing, although technically a supercomputer is a system that performs at or near the currently highest operational rate for computers. Some supercomputers work at more than a petaflop or 1015 floating-point operations per second.

The most common users of HPC systems are scientific researchers, engineers and academic institutions. Some government agencies, particularly the military, also rely on HPC for complex applications. High-performance systems often use custom-made components in addition to so-called commodity components. As demand for processing power and speed grows, HPC will likely interest businesses of all sizes, particularly for transaction processing and data warehouses.

The used infrastructure is based on 14 blade with 8-core processors. Each unit was used to compute one simulation at time. The connection used among the blades is the Gigabit Ethernet. The high level of parallelism was reached by the use of software based on MPI (Message Passing Interface). It is a standardized and portable message-passing system designed by a group of researchers from academia and industry to function on a wide variety of parallel computers. The standard defines the syntax and semantics of a core of library routines useful to a wide range of users writing portable message-passing programs in Fortran 77 or the C programming language. Several well-tested and efficient implementations of MPI include some that are free and in the public domain. These fostered the development of a parallel software industry, and there encouraged development of portable and scalable large-scale parallel applications.

The software E3D was configured and compiled to be used with MPI. Through this system and library we generated one signal in about 17 hours.

Bibliography

[AAHS06] Stephen J. Arrowsmith, Marie D. Arrowsmith, Michael A. H. Hedlin, and Brian Stump. Discrimination of delay-fired mine blasts in wyoming using an automatic time-frequency discriminant. In James C. Hayes, Pamela G. Doctor, Tom R. Heimbigner, Charles W. Hubbard, Lars J. Kangas, Paul E. Keller, Justin I. Mcintyre, Brian T. Schrom, and Reynold Suarez, editors, *28th Seismic Research Review: Ground-Based Nuclear Explosion Monitoring Technologies APPLICATION OF ARTIFICIAL NEURAL NETWORK MODELING TO THE ANALYSIS OF THE AUTOMATED RADIOXENON SAMPLER-ANALYZER STATE OF HEALTH SENSORS*, 2006.

[All78] R.V. Allen. Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America*, 68:1521–2532, 1978.

[AR02] K. Aki and P.G. Richards. *Quantitative Seismology: Theory and Methods*. Geology (University Science Books).: Seismology. Univ Science Books, 2002.

[BF01] Ahmed Badawy and Ali K Abdel Fattah. Source parameters and fault plane determinations of the 28 december 1999 northeastern cairo earthquakes. *Tectonophysics*, 343:63 – 77, 2001.

[BFMS07] S. Barani, G. Ferretti, M. Massa, and D. Spallarossa. The waveform similarity approach to identify dependent events in instrumental seismic catalogues. *Geophysical Journal International*, 168(1):100–108+, 2007.

[BK00] C. Bai and B. L. N. Kennett. Automatic phase detection and identification by full use of a single three component broadband seismogram. *Bulletin of the Seismological Society of America*, 90:187–198, 2000.

[BL97] Michael J. A. Berry and Gordon S. Linoff. *Data Mining Techniques. For Marketing, Sales, and Customer Support*. Wiley, 1997.

[Bor12] Peter Borman, editor. *New Manual of Seismological Observatory Practice*. GFZ German Research Centre for Geosciences, 2012.

[BWT12] Helena Buurman, Michael E. West, and Glenn Thompson. The seismicity of the 2009 redoubt eruption. *Journal of Volcanology and Geothermal Research*, 2012.

[Cor07] Marcella Corduas. Dissimilarity criteria for time series data mining. Learning material, 2007.

[DKB] Tobias Diehl, Edi Kissling, and Peter Bormann. Tutorial for consistent phase picking at local to regional distances.

[DLD+10] Antonino D'Alessandro, Dario Luzio, Giuseppe D'Anna, Giorgio Mangano, and Stefano Panepinto. Single station location of small-magnitude seismic events recorded by obs in the ionian sea, 2010.

[DM97] Hengchang Dai and Colin MacBeth. The application of back-propagation neural network to automatic picking seismic arrivals from single-component recordings. *JOURNAL OR GEGPHYSICfi RESEARCH*, 102:105–115, 1997.

[ELLS11] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis, Fifth Edition*. Wiley, 5th edition, 2011.

[GLT+09] G. Giunta, D. Luzio, E. Tondi, L. De Luca, A. Giorgianni, G. D'Anna, P Renda, G. Cello, F. Nigro, and K. Vitale. The palermo (sicily) seismic cluster of september 2002, in the seismotectonic framework of the tyrrhenian sea-sicily border area. *Annals of Geophysics*, 47(6), 2009.

[GN06] David N. Green and JÃČrgen Neuberg. Waveform classification of volcanic low-frequency earthquake swarms and its implication at soufriÃČre hills volcano, montserrat. *Journal of Volcanology and Geothermal Research*, 153:51 – 63, 2006.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[HBV02] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part i. *SIGMOD Rec.*, 31(2):40–45, June 2002.

[HO10] Jens Havskov and Lars Ottemoller. *Routine Data Processing in Earthquake Seismology*. Springer, 2010.

[JCM12] J.P. Jones, R. Carniel, and S.D. Malone. Subband decomposition and reconstruction of continuous volcanic tremor. *Journal of Volcanology and Geothermal Research*, pages 98 – 115, 2012.

[KP12] Dane Ketner and John Power. Characterization of seismic events during the 2009 eruption of redoubt volcano, alaska. *Journal of Volcanology and Geothermal Research*, pages –, 2012.

[KPV07] E. Kokinou, C. Panagiotakis, and F. Vallianatos. Earthquake/noise discrimination and estimation of p-s phases based on wave characteristics. *Bulletin of Ihe Geological Sociely of Greece*, 2007.

[KR87] L. Kaufman and P. Rousseeuw. *Clustering by Means of Medoids*. Reports of the Faculty of Mathematics and Informatics. Delft University of Technology. Fac., Univ., 1987.

[LCW10] Yingmin Li, Huiguo Chen, and Zheqian Wu. Dynamic time warping distance method for similarity test of multipoint ground motion field. *Mathematical Problems in Engineering*, 2010.

[Lev88] Alan R. Levander. Fourth-order finite-difference p-sv seismograms. *Geophysics*, 53(11):1425–1436, November 1988.

[LH93] S. Larsen and D. Harris. Seismic wave propagation through a low-velocity nuclear rubble zone. Technical report, Lawrence Livermore National Lab., CA., 1993.

[Mad76] Raul Madariaga. Dynamics of an expanding circular fault. *Bulletin of the Seismological Society of America*, 66(3):639–666, 1976.

[MFB+11] Frédérick Massin, Valérie Ferrazzini, Patrick Bachélery, Alexandre Nercessian, Zacharie Duputel, and Thomas Staudacher. Structures and evolution of the plumbing system of piton de la fournaise volcano inferred from clustering of 2007 eruptive cycle seismicity. *Journal of Volcanology and Geothermal Research*, 202:96 – 106, 2011.

[MMP11] Robert Myhill, Dan McKenzie, and Keith Priestley. The distribution of earthquake multiplets beneath the southwest pacific. *Earth and Planetary Science Letters*, 301:87 – 97, 2011.

[MTI09] T. Miwa, A. Toramaru, and M. Iguchi. Correlations of volcanic ash texture with explosion earthquakes from vulcanian eruptions at sakurajima volcano, japan. *Journal of Volcanology and Geothermal Research*, 184:473 – 486, 2009.

[OAEGODCD06] Mauricio Orozco-Alzate, Marcelo Enrique Garcia-Ocampo, Robert P. W. Duin, and Cesar German Castellanos-Dominguez. Dissimilarity-based classification of seismic signals at nevado del ruiz volcano. In *II Latin American Congress of Seismology*, page 1–16, Bogota, Colombia, August 2006.

[PBB+12] Germán A. Prieto, Gregory C. Beroza, Sarah A. Barrett, Gabriel A. López, and Manuel Florez. Earthquake nests as natural laboratories for the study of intermediate-depth earthquake mechanics. *Tectonophysics*, pages 42 – 56, 2012.

[PMDOA+10] Diana Porro-Munoz, Robert P. W. Duin, Mauricio Orozco-Alzate, Isneri Talavera, and John Makario Londono-Bonilla. Classifying three-way seismic volcanic data by dissimilarity representation. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 814–817, Washington, DC, USA, 2010. IEEE Computer Society.

[Ran71] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):pp. 846–850, 1971.

[SC01] C.D. Stephens and B.A. Chouet. Evolution of the december 14, 1989 precursory long-period event swarm at redoubt volcano, alaska. *Journal of Volcanology and Geothermal Research*, 109:133 – 148, 2001.

[Sen08] Pavel Senin. Dynamic Time Warping Algorithm Review. Technical Report CSDL-08-04, Department of Information and Computer Sciences, University of Hawaii, Honolulu, Hawaii 96822, December 2008.

[SS01] Ron Shamir and Roded Sharan. Algorithmic approaches to clustering gene expression data. In *Current Topics in Computational Biology*, pages 269–300. MIT Press, 2001.

[Ste77] Samuel W. Stewart. Real-time detection and location of local seismic events in central california. *Bulletin of the Seismological Society of America*, 67, 1977.

[TK08] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition, Fourth Edition*. Academic Press, 4th edition, 2008.

[TWS10] Weston Thelen, Michael West, and Sergey Senyukov. Seismic characterization of the fall 2007 eruptive sequence at bezymianny volcano, russia. *Journal of Volcanology and Geothermal Research*, 194(4):201 – 213, 2010.

[UTS+08] Kodo Umakoshi, N. Takamura, N. Shinzato, K. Uchida, N. Matsuwo, and H. Shimizu. Seismicity associated with the 1991-1995 dome growth at unzen volcano. *Journal of Volcanology and Geothermal Research*, 175(1):91–99, jul 2008.

[Vir86] J. Virieux. P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics*, 51:889, April 1986.

[VSL12] Maurizio Vassallo, Claudio Satriano, and Anthony Lomax. Automatic picker developments and optimization: A strategy for improving the performances of automatic phase pickers. *Seismological Research Letters*, 83:541–554, 2012.

[WAY+98] Mitchell Withers, Richard Aster, Christopher Young, Judy Beiriger, Mark Harris, Susan Moore, and Julian Trujillo. 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, February 1998.

[ZYD+12] Hussam Eldein Zaineh, Hiroaki Yamanaka, Rawaa Dakkak, Ahlam Khalil, and Mohamad Daoud. Estimation of shallow s-wave velocity structure in damascus city, syria, using microtremor exploration. *Soil Dynamics and Earthquake Engineering*, 39(0):88 – 99, 2012.