A new dissimilarity measure for clustering seismic signals

Francesco Benvegna[$], Antonino D'Alessandro[+], Giosuè Lo Bosco[$], Dario Luzio[*], Luca Pinello[$], Domenico Tegolo[$]

francesco.benvegna@unipa.it, antonino.dalessandro@ingv.it, giosue.lobosco@unipa.it, dario.luzio@unipa.it, pinello@unipa.it, domenico.tegolo@unipa.it

[$]Dipartimento di Matematica e Informatica
via Archirafi 34, 90123 Palermo, Italy

[*]Dipartimento di Fisica e Chimica della terra
Via Archirafi 36, 90123 Palermo, Italy

[+]Istituto Nazionale di Geofisica e Vulcanologia
Centro Nazionale Terremoti, Italy

Abstract. Hypocenter and focal mechanism of an earthquake can be determined by the analysis of signals, named waveforms, related to the wave field produced 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 work 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.

1 Introduction

In seismically active areas often occurred earthquakes that produce very similar waveforms (multiplets). A high level of similarity between the waveforms is a clear indication of events generated in a small seismogenetic volume, with similar source mechanisms. These events can be associated with both tectonic [1, 2] and volcanic activity [4]. Based on the similarity between complete seismograms of microearthquakes occurred on the San Andreas Fault, Geller and Mueller deduced that their hypocenters can't be distant from each other by more than a quarter of the dominant wavelength [3].

The definition of the new dissimilarity was inspired by a simple observation: a seismic signal is characterized by the overlapping of several wave trains (seismic

phases) which, because of their different travel path, arrive at the recording point at different times. They relate to both body waves and surface waves. The body waves concerning the irrotational component of the displacement field (P) propagate faster than those concerning the solenoidal one (S) and even more than the surface or guided waves.

The common hypocentral location methods, based on P and S phases arrival times inversion, are generally not accurate enough for a reliable relative location of very close hypocenters (hypocentral spacing much smaller than the typical distance between the stations of the seismic network) and modeling of the focal mechanisms distribution in the source region. To determine differential arrival times with high accuracy, techniques exploiting waveform similarities have been proposed [1, 7].

The dissimilarity functions based on signal cross correlation have been used to measure the difference degree between seismic events [8, 9] and to provide more precise estimates of the differences in arrival times of P and S phases of similar events [10, 11]. A new challenge needs to identify, some groups containing similar signals with respect to a predetermined criterion, in a large set of three-component signals.

Clustering technique and the related algorithms can be adopted to face that challenge. In the general case, cluster analysis play a central role in the design of data analysis systems [12]. Moreover, clustering allows analysts to discover the nature of the data for further analysis. Dissimilarity (similarity) functions are a fundamental ingredient of clustering procedures, and their discrimination ability can be measured by means of clustering validation indices [13]. Clustering validation indices can be divided into internal and external ones: the former gives a reliable indication of how well a partitioning solution captures the inherent separation of the data into clusters, the latter measures how well a clustering solution agrees with the *gold solution* for a given data set [15]. A gold solution of a generic dataset, can be also inferred by analyzing the data, i.e., by the use of internal knowledge via data analysis tools such as clustering algorithms.

A basic consideration about cross-correlation and/or cross-spectral dissimilarities, is that they are effective in forming subsets of similar events just if earthquakes included in each set are very close in space, magnitude and focal parameters domains and the waveforms recorded have a good signal to noise ratio. Another of its drawback is the computation time, that will necessary affect the adopted clustering algorithm. This is a very important point since the development of dense seismic networks, with 3 components broadband sensors, permit to collect a lot of seismological data that should be processed by clustering techniques.

In this paper we propose a new dissimilarity measure able to catch difference in shapes between waveforms. It has been used in conjunction with a hierarchical clustering algorithm and applied to a dataset of earthquakes waveforms and to another dataset of signals generated by bursts, both recorded by an Ocean Bottom Seismometers with Hydrophone (OBS/H) deployed in the southern Tyrrhenian sea. We compared its discrimination ability with that of a cross correlation

based dissimilarity. Results show the effectiveness of using the proposed dissimilarity, in terms of cluster homogeneity validation index and computational time.

2 Dissimilarities definitions

In this section the two dissimilarity measures used in this work are described. The first one is the classical *cross correlation dissimilarity*, the other one is the new designed mesure called *cumulative shape dissimilarity*.

We recall that the *cross correlation* between two vectors x_1 and x_2, both of length n, is so defined

$$R_{x_1,x_2}(k) = \begin{cases} \sum_{i=0}^{n-k-1}(x_1(i+k)-\mu_{x_1}) \times (x_2(i)-\mu_{x_2}) & \text{if } k \geq 0 \\ R_{x_2,x_1}(-k) & \text{otherwise} \end{cases}$$

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

$$\delta_R(x_1, x_2) = 1 - \frac{1}{\sigma_x \sigma_y} \max_{k=1,..,2n-1} R_{x_1,x_2}(k-n). \tag{1}$$

Where σ_x and σ_y are the standard deviations of x_1 and 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. In fact, it is ineffective in forming subsets of similar events if earthquakes included in each set are not very close in space, magnitude and focal parameters domain, and noise is present in the recorded signal. Moreover, for a signal of length n its computational time is $O(n^2)$. 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, especially the first P and S arrival times, are less sensitive to the travel path and clearly 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[1]) signals which contain the two types of body waves: P wave and S wave. Both waves have a 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;

[1] 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.

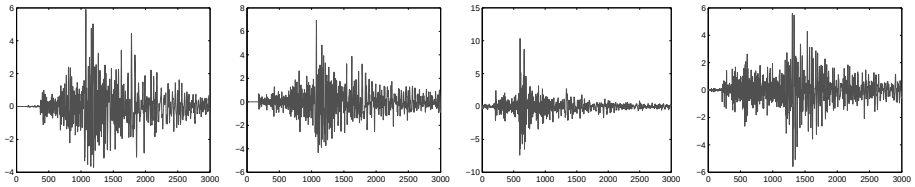
– it should be capable of detecting where two wave shapes are similar regardless of magnitude.

The first two properties, can be satisfied by a dissimilarity acting 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 plot 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_1 and x_2 both of the same length n, and let s_1 and s_2 be their cumulative sums $s_i(k) = \frac{\sum_{r=1}^{k} x_i^2(r)}{\sum_{r=1}^{n} x_i^2(r)}$ $(i = 1, 2)$, we can calculate their absolute difference $sd(k) = |s_1(k) - s_2(k)|$. Finally, the new proposed dissimilarity, called *cumulative shape dissimilarity* δ_s is defined as:

$$\delta_s(x_1, x_2) = \sum_k \frac{|sd(k+1) - sd(k)|}{max_j\,|sd(j+1) - sd(j)|}. \tag{2}$$

Note that δ_s represents the sum of the derivative of the difference between the cumulative sums of x_1 and x_2. In figure 1 we report 4 examples of signal, in figure 2 their cumulative sums and the pairwise dissimilarities. Finally, in figure 3 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 $O(n)$.



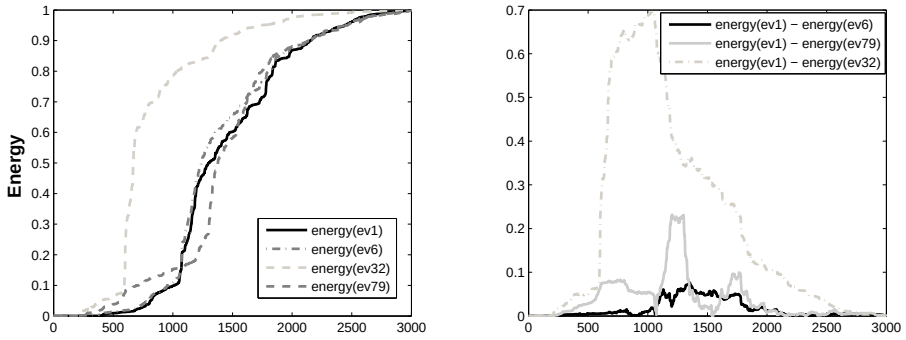
(a) (b) (c) (d)

Fig. 1. (a) event 1 (b) event 6 (c) event 32 (d) event 79

3 Evaluation of a dissimilarity measure

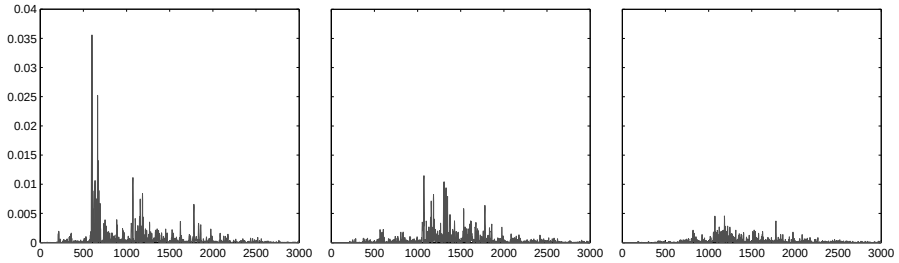
In order to evaluate the performance of a dissimilarity, we have adopted three different indices. Two of them are related to the partitioning inducted by a clustering algorithm which make use of the dissimilarity, while the other one does not consider any partitioning information.

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 [15],



(a) (b)

Fig. 2. (a) cumulative energy of the events; (b) difference between cumulative energies



(a) (b) (c)

Fig. 3. Derivative at sample point i of the difference between cumulative energies ($|sd(i+1) - sd(i)|$) (a) event 1 - event 32 (b) event 1 - event 79 (c) event 1 - event 6

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)` internal indices [15] of a partitioning $\mathcal{P}$ produced by a clustering algorithm by using the dissimilarity δ, whose formulas are here reported:

$$H = \frac{1}{|X|} \sum_{i=1}^{t} \sum_{x \in p_i} 1 - \delta(x, \mu_i) \tag{3}$$

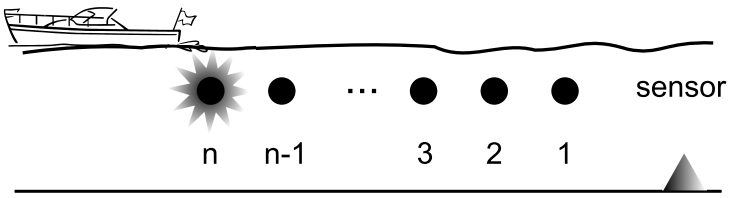


Fig. 4. Plan of the bursts experiment

$$S = \frac{1}{\sum_{i \neq j} |p_i||p_j|} \sum_{i \neq j} |p_i||p_j| \delta(\mu_i, \mu_j) \tag{4}$$

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 so called 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 $\mathcal{P}$. External indices are usually defined via a $r \times t$ contingency table T, where T_{ij} represents the number of items in both c_i and p_j, $1 \leq i \leq r$ and $1 \leq j \leq t$. For our experiment we have used the `Adjusted Rand index`[14].

$$R_A = \frac{\sum_{i,j} \binom{T_{ij}}{2} - \frac{[\sum_i \binom{T_{i\cdot}}{2} \sum_j \binom{T_{\cdot j}}{2}]}{\binom{N}{2}}}{\frac{1}{2}[\sum_i \binom{T_{i\cdot}}{2} + \sum_j \binom{T_{\cdot j}}{2}] - \frac{[\sum_i \binom{T_{i\cdot}}{2} \sum_j \binom{T_{\cdot j}}{2}]}{\binom{N}{2}}} \tag{5}$$

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, it is also possible to use an a priori information different form the gold solution. In the following, we will define a new index, called `Dissimilarity Optimality index` which make use of the sort of data items.

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:

$$do = \sum_{i=1}^{n} \frac{|i - j - 1|}{N - 2} \text{ with } j = \underset{1 \leq k \leq N, k \neq i}{\operatorname{argmin}} \delta(x_i, x_k) \tag{6}$$

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

4 Experimental Results

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 [17]. 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 datalogger 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 on 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 [16]. Among this microevents, 95 of them 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.
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. 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 shows the arrangement of the experiment. The acquired data define what is here named as *bursts dataset*, that 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.

4.1 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

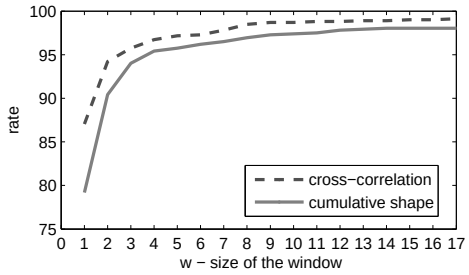


Fig. 5. Diagram of coverage proximity for w between 1 and 17

part has an higher signal to noise ratio as explained in section 2. The performance of dissimilarities on this dataset has been measured by using the Dissimilarity Optimality index. This is due to the fact that the conducted experiment involves 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.

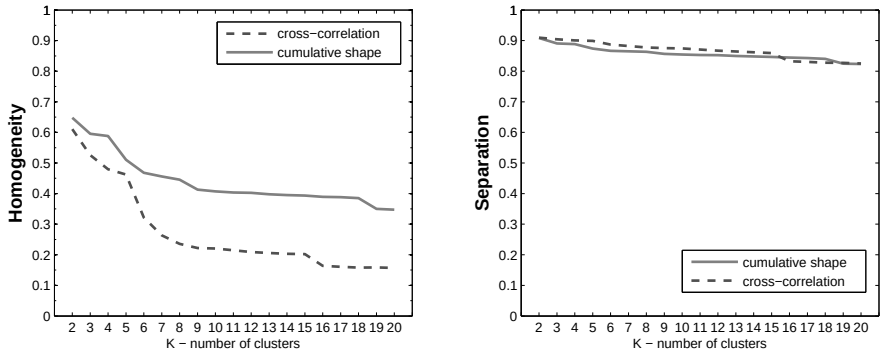
Both values are very close to 0 and their difference is very small.

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 5 shows its computation for w ranging from 1 until 17.

The results (see figure 5) 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 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 the expert taking into consideration both its 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 [17]. This 4 clusters, had finally been splitted 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



(a) (b)

Fig. 6. Internal indices for the considered dissimilarities: (a) Homogeneity; (b) Separation

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 6(a,b) 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.

5 Conclusion and future work

In this paper, a new dissimilarity measure between seismic signals called cumulative shape dissimilarity has been proposed. A number of tests have been done on two different dataset of earthquake events. The former is characterized by synthetic signal without gold solution in spite of the latter that, due to its real nature, have a gold solution proposed by an expert providing 9 cluster with a variable number of elements. 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). 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%). 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.

References

1. Scherbaum F., J. Wendler, Cross spectral analysis of Swabian Jura (SW Germany) threecomponent microearthquake recordings, J. Geophys., 60, 157-166, 1986.
2. Console R., Di Giovambattista R., Local earthquake relative location by digital records, Phys.Earth Planet. Inter. 47, 43-49, 1987.
3. Geller R.J. and Mueller, C.S., Four similar earthquakes in central California, Geophys. Res. Lett. 7, 821824, 1980.
4. Got J.L., M. Frechet, F.W. Klein, Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea, J. Geophys. Res. 99, 15, 375-386, 1994.
5. Aster R.C., Scott J., Comprehensive Characterization of Waveform Similarity in Microeartquake data sets, Bulletin of the Seismological Society of America, Vol. 83, No. 4, pp. 1307-1314, 1993.
6. Maurer H.R., Deichmann N., Microearthquake cluster detection based on waveform similarities with an application to the western Swiss Alps. Geoph. J. Int., 123, 588-600, 1995
7. Deichmann N., M. Garcia-Fernandez. Rupture geometry from high-precision relative hypocenter locations of microearthquake clusters, Geophys. J. Int. 110, 501-517, 1992.
8. Mezcua J., Rueda J., Earthquake relative location based on waveform similarity. Tectonophysics, 233, 253-263, 1994
9. Menke W., Using waveform similarity to constrain earthquake locations. Bull.Seismol. Soc. Am., 89, 1143-1146, 1999.
10. Gillard D., Rubin A.M., Okubom P., Highly concentrated seismicity caused by deformation of Kilauea's depp magma system. Nature, 384, 343-346, 1996.
11. Phillips W.S., House L.S., Feheler J., Detailed joint structure in a geothermal reservoir from studies of induced microearthquake studies. Journal of Geophysical Research, 102, 745-763, 1997.
12. Jain A. K., Murty M. N., Flynn P. J., Data clustering: a review, ACM Comput. Surv., 31(3), pp. 264–323, 1999.
13. Giancarlo R., Lo Bosco G., Pinello L., Distance Functions, Clustering Algorithms and Microarray Data Analysis, Learning and Intelligent Optimization 2010, Lecture Notes in Computer Science, 6073, 125–138, 2010.
14. Hubert, L. and Arabie, P., Comparing partitions, Journal of Classification, 2, pp. 193–218, 1985
15. Shamir R., Sharan R., Algorithmic approaches to clustering gene expression data, Current Topics in Computational Biology, 269–299, 2002.
16. D'Alessandro A., Luzio D., D'Anna G., Mangano G., Panepinto S., Single station location of small-magnitude seismic events recorded by OBS in the Ionian Sea. Geophysical Research Abstracts, EGU General Assembly, Vienna, Austria, 12, EGU2010-8840, 2010.
17. Giunta G., Luzio D., Tondi E., De Luca L., Giorgiani A., D'Anna G., Renda P., Cello G., Nigro F., Vitale M.,The Palermo (Siciliy) seismic cluster of Septermber 2002, in the seismotectonic framework of the Tyrrhenian Sea-Sicily border area, Ann. of Geoph., 47(6), 1755–1770, 2004