IMPROVING AUTOMATIC LOCATION PROCEDURE BY WAVEFORMS SIMILARITY ANALYSIS: AN APPLICATION IN THE SOUTH WESTERN ALPS (ITALY)

M. Massa (1), G. Ferretti (1), D. Spallarossa (1) and C. Eva (1).

(1) Dipartimento per lo Studio del Territorio e delle sue Risorse, Università di Genova, Viale Benedetto XV, 5, 16132 Genova, Italy.

Abstract

The accuracy of automatic procedures for locating earthquakes is influenced by several factors such as errors in picking seismic phases, network geometry, modeling errors and velocity model uncertainties. The main purpose of this work is to improve the performances of the automatic procedure employed for the “quasi-real-time” location of seismic events in North Western Italy by developing an additional location algorithm based on the waveform similarity analysis. To detect “earthquake families” a cross-correlation technique was applied to a data set of seismic waveforms recorded in the period 1985-2002, in a small test area (1600 km²) located in the South Western Alps (Italy). Normalized cross correlation matrices were calculated using about 2700 seismic events, selected on the basis of the signal to noise ratio, manually picked and located by using the Hypoellipse code. The waveform similarity analysis, based on the bridging technique, allowed to group about 65% of the selected events into 80 earthquake families (multiplets) located inside the area considered. For each earthquake family a master event is selected, manually re-picked and re-located by using Hypoellipse code. Having chosen a reference station (STV) on the basis of a more completeness of the data set, an automatic procedure with the aim of cross-correlating new seismic recordings (automatically picked) to the waveforms of the events belonging to the detected families has been developed. If the new event is proved to belong to a family (on the basis of the cross-correlation values), its hypocenter co-ordinates are defined by location of the master event of the associated family. The performance of the proposed procedure is tested and demonstrated using a data set of 104 selected earthquakes recorded in the
period January 2003 - June 2004 and located in the test area. The automatic procedure is able to locate, associating events with the multiplets detected by the waveform similarity analysis, about 50 % of the test events, nearly independently from the accuracy of the automatic phase picker and without the biasing of the network geometry and of the velocity model uncertainties.

Introduction

The development of reliable and accurate automatic algorithm for processing seismic data and for “quasi-real-time” location of earthquakes represents one of the main objective of seismic warning systems. Recently several robust automatic procedures have been implemented for detecting, onset picking and identifying of signal phases in seismogram records (Allen, 1978; Earle and Shearer, 1994; Evans and Pitt, 1995; Dai and MacBeth, 1995; Patanè and Ferreri, 1999; Sleeman and van Eck, 1999). The appropriate choice and usage of an accurate automatic picker are essential for reliable hypocenter determination, fixing the quality of the phase readings. Nevertheless, in automatic and “quasi-real-time” management of earthquake locations, the accuracy of absolute hypocenter co-ordinates is also controlled by other factors such as network geometry, available phases and knowledge of the crustal structure (Pavlis 1986; Gomberg et al., 1990). In particular, in the case of very spatially limited microseismic activity, occurring in a heterogeneous medium, joint hypocenter determination (Douglas, 1967) often results in biased locations due to coupling of the above factors (Cattaneo et al., 1997; Cattaneo et al., 1999).

This paper points out the uncertainties of hypocenter co-ordinates proved by the automatic location procedure (automatic phase picking, based on Allen picker - Allen,1978 - coupled with Hypoellipse code - Lahr, 1979) and the possible improvements in “quasi-real-time” earthquake location, considering a test area located in the South Western Alps, Italy (Figure 1b). The purpose is to introduce in the automatic location procedure, adopted for monitoring the seismic activity of North-western Italy since 2005, an additional automatic method based on waveform similarity analysis in order to strongly improve the reliability and the accuracy of “quasi-real-time” location, mainly for low energy events (MI < 3.0).
In recent years, several studies have been carried out to assess the accuracy and the reliability of the procedure developed to locate events in North Western Italy (Cattaneo 1999; Cattaneo 1999, Cattaneo 1999; Spallarossa 2001). This area, monitored by the Regional Seismic Network of North Western Italy (hereinafter RSNI) since 1982 (Figure 1a), is characterized by strong structural heterogeneities and by a complex tectonic pattern (Moho depth ranging from about 10 km in the Ligurian sea to nearly 50 km beneath the Alps, Buness et al., 1990) that has induced significant lateral variations of the crustal and upper mantle velocity (Kissling et al., 1995; Kissling and Spakman, 1996; Parolai et al., 1997).

The automatic “quasi-real-time” location of earthquakes recorded by the RSNI network, performed using the well known Hypoellipse code (Lahr, 1979), is influenced and, often, biased by the accuracy of the velocity models employed, by the network geometry and by the quality of the automatic phase picking. As shown by the distribution of the horizontal error axes calculated for the events recorded in the test area during the period 1985-2002 (Figure 2), the location of the earthquakes is characterized by considerable uncertainties (horizontal errors up to 50 km) mainly along the SW-NE direction, pointing out that routine location procedure can lead to wrong hypocenter co-ordinates determination.

The method proposed in this paper, starting from the results of a waveform similarity analysis, allows to determine the hypocentral coordinates of seismic events by a simple association to a particular earthquake family (Tsujiura, 1983). To identify the earthquake families located in the test area, a technique based on the peak of the normalized cross-correlation function (Console and Di Giovanbattista, 1987; Planet and Cansi, 1988; Pechmann and Thorbjarnardottir, 1990; Deichmann and Garcia-Fernandez, 1992; Haase et al., 1995, Cattaneo et al., 1997), combined with the bridging technique (Cattaneo, 1999), has been applied. The automatic location procedure, described in detail in the following paragraphs, is tested by applying it to a significant number of earthquakes that occurred in the test area in the period January 2003 - June 2004 (Figure 1b).
Data, data processing and cluster analysis

When analyzing a low seismicity area, characterized by very clustered earthquakes, it is quite common to find a great number of events that show similar waveforms (also commonly referred to as a "doublet", for a single pair of events or a "multiplet", for a series). Geller and Mueller (1980) found that earthquakes with such characteristics are caused by very similar source mechanisms with hypocenters that lie within one quarter of a dominant wavelength from each other, representing rupturing of the same asperity. Such groups of events were called "earthquake families" by Tsujiura (1983), who identified this kind of swarm as energy release due to repeated slip on the same fault plane during a foreshocks-mainshock-aftershocks sequence. If we assume that each seismic source zone is characterized by a particular focal mechanism and a small spatial extent with respect to the ray path length, all the events generated by this source will show similar waveforms.

In this paper, in order to recognize earthquake families and, afterwards, to test the performances of the proposed automatic location procedure, two data sets, including earthquakes located in a small area (40 km x 40 km) of the South Western Alps (Figure 1b), have been collected: the first, composed by events recorded in the period 1985-2002 and manually re-picked, is used to collect the data base of multiplets and the second, composed by events recorded in the period January 2003 - June 2004 and automatically picked, is used to test the reliability of the proposed location procedure.

To detect similar events in the test area, the first data set made up of about 3000 seismic events (Figure 1b), with local magnitude up to 3.3 (Spallarossa et al., 2002), recorded by an one-component reference station (STV station, Figure 1b), are considered. Among several algorithms for waveform similarity characterization proposed in recent years (cross spectral techniques, Scherbaum and Wendler 1986, Got et al., 1994; pattern recognition, Joswig, 1995; cross-correlation analysis on P and S-wave checked independently, Maurer and Deichmann, 1995; fractal approach, Smalley et al., 1987; syntactic pattern recognition schemes Zhizhin et al., 1992; 1994), a technique based on the peak of the normalized cross-correlation function is applied. To ensure reliable outcomes, only the recordings characterized by a signal to noise ratio greater than 10 dB (about 2700 events, that is 90% of
the initial data set of waveforms) are used. Figure 3 shows the S/N ratio for the S-phases computed over windows 3 s wide for different frequency bands.

All available vertical waveforms recorded by the STV station have been organized in matrices and then compared in the search for waveform similarities. In order to perform similarity analysis a continuous signal of six seconds after the P onset has been considered. For recordings at an average hypocentral distance of 5 to 15 km, the comparison is, then, made on a signal more complicated than a simple first pulse. The first part of the coda, that is merely affected by the propagation and can be matched only by seismograms of rays that have traveled similar paths, is taken into account. Our cross-correlation window does not include the “regional” part of coda waves (Aki, 1969), expected at twice the S wave travel time (Rautian and Khalturin, 1978), at least for our data set for which the S-P times range from 1.0 to 4.0 seconds.

Since the selected waveforms may contain noise and since high frequency wiggles would be too difficult to compare, biasing the cross-correlation results, all waveform comparisons have been carried out in the frequency range 1-10 Hz. This filtering window, adopted after an analysis of the S/N ratio (figure 3), reduces natural noise without altering the seismograms too much. One example of the matrix of cross correlation coefficient at the STV station is shown in figure 4 for a subset of the data set.

The detection of multiplets is performed using a bridging technique to overcome bias when comparing waveforms differing from each other by more than one order of magnitude (Deichmann and Garcia-Fernandez, 1992). If two couples of events (A,B) and (B,C), exceeding the a priori selected correlation threshold, share a common quake (B), then all three events are attributed to the same family, even if the match between A and C is below the reference value for similarity (called threshold hereinafter). The bridging algorithm is based on the Equivalence Class approach (Press et al., 1988) and has already been applied to local earthquake data sets by Aster and Scott (1993) and Cattaneo et al. (1997; 1999). The potential and the success of this technique is estimated in Ferretti et al., 2005.

Finally, the cross-correlation threshold (index of waveform similarity) has been chosen by a two-step procedure:
- by analysis of the distribution of the cross-correlation coefficients versus each couple of events recorded by the one component reference station; where the plot shows a sudden flattering, and the trend deviates from a pure normal distribution (Maurer and Deichmann, 1995), that is where the threshold is chosen (Figure 5a).

- by the examination of the number of families and population of each group versus a different correlation threshold. The value of threshold to assign to STV station is evaluated to ensure both the maximum number of recognised families and the greatest number of events for each of them (Figure 5b).

Finally a cross correlation threshold of 0.80 (80% of waveform similarity) has been chosen. To strengthen the results derived considering the STV station, the waveform similarity analysis is also applied to the ENR one component station located in the considered area (Figure 1). In a small area with small distance between stations, families which are recognized by more than one receiver, indicate reliability in the definition of multiplets and represent an important constraint on the number of events belonging to them (Ferretti et al, 2005).

**Cluster analysis results**

By the waveform similarity procedure, described in the previous paragraph, about 1500 events belonging to the first data set (2,700 vertical waveforms selected on the basis of the signal to noise ratio), which includes recordings from 1985 to 2002, have been grouped into multiplets. In figure 6, the distribution of the events belonging to the 80 detected earthquake families, 72 of them recognized by both the STV and ERN stations, is plotted. The lower number of families detected by ENR station is related both to the worse efficiency of this receiver during the time period considered (about 2400 records versus 3000 of the STV) and the slightly worse signal to noise ratio biasing the records of this one component station.

From figure 7, the spatial distribution of the epicenters of the events belonging to the main multiplets recognized by the STV reference station (family 7, 201 events; family 12, 205 events; family 13, 57 events; family 56, 94 events) strongly correlate with the SW - NE trend of the horizontal error axes derived by the routine location procedure (figure 2).
Thus, it is questionable whether the spatial distribution of the epicenters, as defined by the routine locations, is really accurate and representative of the real position of the seismogenic structures activated in the area. The routine locations (figures 6 and 7) show a non negligible spatial dispersion in the SW-NE direction for the events belonging to some of the multiplets that cannot be justified by a level of similarity greater than 80 % (threshold of 0.80 used in the waveform similarity analysis) for the events of each family. As demonstrated in previous studies (Augliera et al., 1995), very high correlation values, computed using large temporal windows, imply very short inter distances between similar events.

Looking at distribution of the families, it is also possible to recognize several multiplets whose spatial extent partially overlaps within a quite limited area (i.e. families 7 and 12 in figure 7). In order to verify the correctness of our grouping, a visual inspection of all waveforms (figure 8) and a detailed analysis of each multiplet (considering polarities, S-P values, focal mechanisms) has been performed.

As a next step, a master event representative of each detected family has been selected using the following constraints, designed to identify the best located events:
- high signal to noise ratio calculated on both the STV and, if available, ENR stations (best quality of signals);
- greatest number of recorded phases (by the RSNI network);
- minimum azimuthal gap;
- mean value of magnitude compared to the other events belonging to the same family.

If possible, among the events satisfying these the previous constraints, those recorded by the dense network installed in the test area during the experiment ALPS3D (August 1996 - February 1997 - Groupe de Recherche Geofrance3D, 1997; Spallarossa et al., 2001) have been selected. The events selected by these constraints potentially ensure accurate and reliable location on the basis of number of phases, network geometry and azimuthal gap.

The 80 master events have been carefully manually re-picked and re-located by the Hypoellipse code (figure 9). The reliability of the location of the 80 masters has been tested by considering different 1D velocity models (Spallarossa et al., 2002) and by carrying out a 3D-grid search location (Scarfi et al., 2003; Massa et al., 2005). Figure 10 shows the root
mean square (rms) travel-time residuals for the master event of family number 12; residuals for 8000 points (i.e. trial fixed hypocentral positions) are shown. In this experiment, wide distributed low rms values indicate an unstable solution, whereas a tight distribution represents a stable solution. Looking at figure 10, the accuracy of the location of the event considered is confirmed by the small dimension of the area with rms < 0.1s on both horizontal and vertical sections. The same test, performed on the events representative of the most populated families, lead to similar results, demonstrating the high reliability of the location of each master event.

**Location algorithm and reliability tests**

Starting from the results of the waveform similarity analysis (formation of the 80 families and re-location of the selected master events), an additional automatic location procedure has been developed and tested. In figure 11 the flow diagram shows the proposed automatic location method, whose main steps are summarized below:

Step 1 - An events recorded by RSNI network is automatically picked (by a picking engine based on Allen picker - Allen, 1978) and preliminarily located (by Hypoellipse code - Lahr, 1979) inside the area considered in the waveform similarity analysis. Preliminary hypocentral coordinates and initial estimation of location errors are determined.

Step 2 - An automatic cross correlation procedure is implemented to compare the new recorded waveform with the seismograms of the events belonging to the detected family (i.e. 1500 events, 80 families from the period 1985 - 2002). The signals are pre-processed according to the parameters used for the waveform similarity analysis (performed by the reference station) by filtering in a frequency range of 1.0 - 10 Hz and taking 6 seconds of seismogram starting from the P-phase onset.

Step 3 - The analysis of the cross-correlation matrix is performed. Using a cross-correlation threshold of 0.80 (similarity index of 80%), we check whether the new event can be associated with one of the recognized families (i.e. the cross-correlation value between the new recording and one or more events belonging to a family is higher than 0.8) .
Step 4 - If step 3 is successful, the new event increases the population of the associated family and the hypocentral co-ordinates of the new event are defined by the master event hypocenter of the associated family.

Step 5 - If step 3 is not successful, an automatic cross-correlation between the new event and the “non-grouped data set” (i.e. the 1200 events not belonging to any family, see previous section) is performed.

Step 6 - The analysis of the cross-correlation values calculated between the new waveform and the earthquakes belonging to the “non-grouped data set” is performed. Taking into account a correlation threshold of 0.80 (similarity index of 80%), we check if a “new” earthquake family can be formed.

Step 7 - If step 6 is successful and, hence, the initial data set of multiplets increases (i.e. from 80 to 81 families) a master event related to the new family is selected, according to the constraints previous listed, manually re-picked and carefully re-located.

Step 8 - If step 6 is not successful, the new event is inserted in the “non-grouped data set” and the preliminary hypocentral co-ordinates (Step 1) defines the “quasi-real-time” location of the event.

The procedure is repeated when a new event recorded by the RSNI network and initially located inside the area considered in this study occurs. The proposed location procedure is tested by applying it to the data set of 104 events recorded in the period 2003 - 2004 and occurring in the area under study (figure 12). Each event, after an automatic picking, is processed by the proposed location algorithm. The results, shown in figure 13, indicate that the procedure based on waveform similarity analysis is able to locate about 50 events (50% of the 2003-2004 data set). The test events can be associated to the families with numbers 1, 7, 12 (figure 14), 25, 28, 30, 47, 58, 76, 78, 79 (steps 2 and 3) and, then, their locations can be automatically defined by the hypocentral co-ordinates of the respective master events (step 4). Among the non-grouped events (50% of the test data set), ten earthquakes are associated into two new seismic families (figure 13, light and dark gray stars) following steps 5 and 6. For the new clusters a master event is selected and re-located (step 7). The
remaining events (about 40) are finally collected in the initial “non-grouped” data set (step 8).

**Discussion and conclusion**

The main purpose of this paper is to propose an automatic algorithm based on the waveform similarity analysis for improving “quasi-real-time” location procedure in South-western Alps. The preliminary location obtained by applying automatic picking engine (based on Allen picker, Allen, 1978) and Hypoellipse code (Lahr, 1979) are often affected by considerable uncertainties (Figure 2). As stated above, the lack of a seismic station sited in the Po Plain (figure 1), with the exception of the ROTM station (only operational since 2002), strongly affects the location accuracy. The routine location procedure shows a distribution of similar events (cross-correlation value greater than 0.80) located in a relatively wide areas elongated NE-SW with errors greater than 5-15 km. The proposed methodology, which implemented the routine location procedure, allows to locate events by waveform similarity only, avoiding the biasing related to the uneven network geometry (great azimuthal gap), to the number of recording stations (just the availability of the STV recordings may be enough to locate recurring events) and to the performances of the automatic picker (just the P phase detection of the STV recordings is necessary).

Considering the parameters chosen to recognize earthquake families (window length, threshold value) and the low magnitude of the events (no extended sources are dealt with), the association of a group of similar events to a point source can be judged correct and reasonable. Hence, the reliability of the proposed procedure is only influenced by the completeness of the data set used for similarity analysis, by the accuracy of the similarity procedure used to determine the reference earthquake families and, finally, by the precision of location of the master events.

In the test site chosen in this paper, the low magnitude detection threshold of the network has assured the completeness of the catalogue since 1985 for magnitude down to 1.0. The 80 detected families potentially well characterize the seismotectonic features of the area, strengthening the location capability of the proposed procedure. The data set used to detect
reference families (1985-2002 data set) presumably collects events associated with most seismic sources located in the zone and, as demonstrated by tests, makes possible the automatic location (by means of our methodology) of at least 50% of newly occurring events. Moreover, by steps 5 and 6 (figure 11), the procedure makes it possible to increase the initial data set of reference families, by recognizing and adding new families, improving the location capability of the method (i.e. the new families recognized by analyzing the 2003 - 2004 catalogue).
References


Captions

Figure 1: (a) Map showing the seismicity that occurred in the Western Alps in the period 1985-2002. The white dotted line indicates the test area considered in this study. White squares and triangles indicate the 3- and 1-component stations of the RSNI network respectively. (b) Map showing the seismicity recorded by the STV (about 3000 events) and the ENR (about 2400 events) 1-C stations (gray triangles) in the test area (40 km x 40 km) for the period 1985-2002.

Figure 2: Axes of the estimated horizontal error ellipses calculated for the earthquakes plotted in figure 1b; the dotted gray line indicates the position of the test area. It is worth noting the SW-NE direction of the main error axes.

Figure 3: Histograms showing the signal to noise ratio calculated considering a portion of seismogram (3s window length) starting from the S phases onset; the results obtained by testing different band-pass filter are reported. The black lines indicate the cumulative curves.

Figure 4: Plot of normalized cross-correlation matrix for a subset (about 200 events) of the 2.700 selected vertical waveforms recorded by the STV station; the cross-correlation values obtained analyzing all possible pairs of signals (considering 6s window from the P-phase onset and filtered in a frequency band of 1.0-10 Hz), are shown. Shaded areas indicate seven “earthquake families” detected applying the bridging technique and considering a threshold value equal to 0.80.

Figure 5: (a) Distribution of the cross correlation values versus the number of pairs; in the zoom right panel note the heavy tail towards high coefficient values greater than 0.75. (b) Number of families recognized by the STV station versus the cross correlation threshold value (index of similarity); for each threshold value the total number of events grouped into families is also reported.
**Figure 6:** Plot of the 1500 events (black crosses) grouped into the 80 recognized earthquake families with respect to the whole selected seismicity (gray crosses) (2700 event). The “earthquake families” (multiplets) have been detected by the bridging technique, considering a minimum correlation threshold of 0.80.

**Figure 7:** Plot of the events belonging to the meaningful detected multiplets: family 7, 201 units (gray crosses), family 12, 205 units (black crosses), family 13, 57 units (gray triangles) and family 56, 94 units (gray stars).

**Figure 8:** Example of waveforms (family 1), filtered in a frequency range of 1.0-10 Hz, recorded by the STV station; the portion of seismogram (6s starting to the P-phase onset) considered in the waveform similarity analysis is shown (gray solid lines). The related amplitude Fourier spectra are also reported (right panels).

**Figure 9:** Plot of the epicentral co-ordinates of the 80 master events (gray stars) with respect to the location of the 1500 events (gray crosses) belonging to the detected earthquake families. The axes of the estimated horizontal error ellipses of the master events are also plotted (black crosses).

**Figure 10:** Plot of the horizontal (left panel) and vertical (right panel) sections of the rms residuals cube (side of 5 km) surrounding the hypocentral co-ordinates of re-located master event selected for family 12. See text for explanation.

**Figure 11:** Flow diagram illustrating the steps of the proposed automatic location procedure. The dotted black line indicates the pre-processing and the waveform similarity analysis described in the text.

**Figure 12:** Map showing 104 earthquakes (gray crosses), recorded by the STV station in the period January 2003 - June 2004, and used to test the location procedure described in figure
11. The axes of the horizontal error ellipses, as derived in step 1 (Figure 11), are also plotted (black crosses).

**Figure 13:** Plot of the 56 events (gray numbers) associated with a particular earthquake family by steps 2 and 3 (figure 11); the black numbers indicate the epicentral co-ordinates of the related master events that define the location of “new” events (step 4). Black and gray stars indicate the location of the events belonging to two new earthquake families recognized in the test area considering the period January 2003 - June 2004 (steps 6 and 7 of figure 11); the black circles indicate the two events selected as master events.

**Figure 14:** Plot of the 9 events (gray crosses) and of the associated horizontal error axis (black crosses) associated with the family 12 (step 2, 3 and 4 of figure 11). The gray star indicates the epicentral co-ordinates of the related master event. An example of waveforms of the associated events is reported in the left panel; the master event of the family and the portion of seismogram (6s window length) considered in the similarity analysis (solid gray lines) are indicated.
Figure 1
Figure 3
Figure 4
Figure 5
Figure 8
SINGLE STATION LOCATION ALGORITHM

WAVEFORM SIMILARITY ANALYSIS

Data set of 80 “earthquake families”.

“No grouped events” (about 1.200 waveforms).

Selection of a master event characteristic of each family.
The master events have been re-picked and re-located. Their hypocentral coordinates define the reference location for each multiplet.

STEP 1 - Recording of a new earthquake (as test events about 100 earthquakes recorded by STV 1C station in the period January 2003 - June 2004, have been considered).

STEP 2 - Automatic cross correlation procedure between the new recorded vertical waveforms and the 1500 events belonging to the “earthquake families” data set.

STEP 3 - Coupling (considering a correlation threshold > 0.80) between the new recorded event and an earthquake belonging to a family?

STEP 4 - The new event increases the population of the associated family. The hypocentral coordinates of the master of this considered family represent the location of the new event.

STEP 5 - Automatic cross correlation procedure between the new recorded waveform and the events belonging to the “no grouped data set”.

STEP 6 - Couple becomes a family (at least 3 events)?

STEP 7 - Increase the number of families

STEP 8 - The “no grouped data set” increases

Figure 11
Figure 14