Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-022-02218-x

ORIGINAL PAPER

A new picking algorithm based on the variance piecewise constant models

Nicoletta D'Angelo[1] · Andrea Di Benedetto[2] · Giada Adelfio[1,3] · Antonino D'Alessandro[3] · Marcello Chiodi[1,3]

Accepted: 17 February 2022

Abstract

In this paper, we propose a novel picking algorithm for the automatic P- and S-waves onset time determination. Our algorithm is based on the variance piecewise constant models of the earthquake waveforms. The effectiveness and robustness of our picking algorithm are tested both on synthetic seismograms and real data. We simulate seismic events with different magnitudes (between 2 and 5) recorded at different epicentral distances (between 10 and 250 km). For the application to real data, we analyse waveforms from the seismic sequence of L'Aquila (Italy), in 2009. The obtained results are compared with those obtained by the application of the classic STA/LTA picking algorithm. Although the two algorithms lead to similar results in the simulated scenarios, the proposed algorithm results in greater flexibility and automation capacity, as shown in the real data analysis. Indeed, our proposed algorithm does not require testing and optimization phases, resulting potentially very useful in earthquakes routine analysis for novel seismic networks or in regions whose earthquakes characteristics are unknown.

Keywords Earthquake early warning · Picking · Change-points · Variance piecewise constant models · Arrival times

1 Introduction

Earthquakes may be generated by fracture processes in the Earth's crust, causing a partial release of the elastic strain energy stored by tectonic processes. The released energy is partially propagated away from its source as a wave-field. There are three basic types of seismic waves—P-waves (also known as primary waves, traveling at the greatest velocity through the Earth), S-waves (transverse waves also known as secondary waves, slower than P-waves) and surface waves (similar in nature to water waves and travel just under the Earth's surface). P-waves and S-waves are sometimes collectively called body waves. The spatial sampling of the wave-field and recorded by a seismic network are the waveforms represented by seismograms. A correct registration and detection by the seismic station of the arrival of the first P-wave, as well as other relevant phases of the seismic event, is crucial for understanding the nature of the generating event (Adelfio et al. 2012). An earthquake monitoring network is a set of seismic stations (accelerometer and velocimeters) suitably distributed over the territory capable of detecting the occurrence of an earthquake. In addition to sensors capable of measuring the shaking generated by the earthquake, a seismic network includes data transmission and processing systems capable of determining in the shortest possible time the location of an earthquake (hypocenter) and its magnitude. When a seismic network is very efficient, i.e. able to automatically and quickly detect an earthquake, it can be used as an early warning tool. Both in the case of use for seismic monitoring and for early warning, the first step to be faced is the correct detection of the seismic event, the correct estimate of the arrival times of the main seismic phases. Given the great growth of seismic networks and the large amount of data that is collected during seismic sequences, the

✉ Nicoletta D'Angelo
nicoletta.dangelo@unipa.it

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy

[2] Dipartimento di Matematica e Informatica, Università degli Studi di Palermo, Palermo, Italy

[3] Istituto Nazionale di Geofisica e Vulcanologia (INGV), Palermo, Italy

Published online: 11 April 2022

development of automatic picking algorithms capable of carrying out a precise and reliable identification of the arrival times of seismic phases has become increasingly important. These algorithms must be at the same robust but not computationally complex so that they can be executed in real-time and also used for early warning purposes. An accurate picking can allow a precise hypocentral localization; moreover, high-quality data can be used for tomographic reconstructions of the subsoil.

More in detail, as stated above, first arrival times on seismograms coincides with the arrival of the first P-wave. The time of the phase-detection $\hat{T}_i$ at a station i is interpreted as the first P-phase arrival time, which is, of course, affected with an error ϵ_i. $\hat{T}_i$ may be written as $\hat{T}_i = T_0 + t_i + \epsilon_i$, where T_0 is the earthquake origin time and t_i is the travel time of a P-wave to station i. The coincidence trigger detects an event if for any combination of a minimum number of stations (typically three or four) the condition $|\hat{T}_i - \hat{T}_j| \leq \epsilon$ is met. ϵ is the maximum allowed difference between trigger times at neighbouring stations. This coincidence trigger works satisfying for local or regional networks, where the inter-distance among the seismic stations is not large. For global networks, this simple event detection algorithm has to be modified. Küperkoch et al. (2012) review the most widespread automatic picking algorithms. Comparative works among different pickers have been carried out in literature (Sleeman and Van Eck 1999; Aldersons 2004; Küperkoch et al. 2010). Here we briefly outline the most known in the literature. Allen (1978, 1982) introduce the concept of characteristic function (CF), obtained by one or several non-linear transformations of the seismogram and should increase abruptly at the arrival time of a seismic wave. This allows both to estimate the arrival time from the CF and assess the quality estimation. The Allen's picker is a fast and robust algorithm, which also accounts for automatic quality assessment. However, since this algorithm is just based on the amplitude information, it might miss emergent P-onsets. A comparative study by Küperkoch et al. (2010) shows that this algorithm tends to pick somewhat early compared to what an analyst would pick.

Another widely used picking algorithm is the one proposed by Baer and Kradolfer (1987). This algorithm is frequently applied, e.g. by 'Programmable Interactive Toolbox for Seismological Analysis' [PITSA, Scherbaum (1992)] and the picking system MannekenPix (Aldersons 2004). In contrast to the Allen's squared envelope function, this CF is sensitive to changes in amplitude, frequency and phase.

The Baer and Kradolfer's picker is also very fast and robust and quite user-friendly, needing just four input parameters. A shortcoming of this algorithm is the missing automated quality assessment. Several comparative studies (Sleeman and Van Eck 1999; Aldersons 2004; Küperkoch et al. 2010) show how this picking algorithm tends to be somewhat late compared to manual P-picks.

The statistical properties of the seismogram might be characterized by its distribution density function and by parameters like variance, skewness and kurtosis. The latter two are parameters of higher order statistics (HOS) and are defined by Hartung et al. (2014). Though just amplitude-based, higher order statistics are quite sensitive to emergent P-onsets. In combination with a sophisticated picking algorithm [e.g. Küperkoch et al. (2010)], which exploits the entire information provided by the determined CF, it yields excellent results. If precisely tuned, the automated quality assessment proposed by Küperkoch et al. (2010) gives similar weights as the analysts. However, choosing the parameters for this sophisticated algorithm is quite difficult and needs a great experience.

Finally, the so called autoregressive-Akaike-Information-Criterion-piker (AR-AIC) proposed by Sleeman and Van Eck (1999) is based on the work by Akaike (1975, 1998), Morita (1984) and Takanami and Kitagawa (1988). It is a highly more sophisticated algorithm based on information theory. The algorithm is computationally quite expensive and hence much slower than the other reviewed pickers.

In this paper, we advocate the usage of the algorithm proposed in Adelfio (2012) for the automated seismogram onset time determination. This considers the case of changepoint detection procedure for changes in variation, assuming that the variance function can be described by a piecewise constant function with segments delimited by unknown changepoints. It is worth to notice that there exists a wide literature about changes in mean in a Gaussian model (Chernoff and Zacks 1964; Gardner 1969; Hawkins 1992; Worsley 1979), as well as the problem of variance change-point detection, mostly focusing on autoregressive time-series models (Wichern et al. 1976; Wang and Wang 2006; Zhao et al. 2010).

In D'Angelo et al. (2020), a new automatic picking algorithm, based on the proposal of Adelfio (2012) and suitable for the implementation of an automatic seismic surveillance system, is proposed and tested on a set of 100 synthetic seismograms, showing that the model is always able to correctly detect the arrival of the first P-wave, as well as other relevant phases of the seismic event, such as the arrival of the first S-wave and the end of the seismic event. These simulated waveforms all presented the same true values of arrival times but different underlying noise.

In D'Angelo et al. (2021) the performance of the proposed algorithm is tested on a set of simulated waveforms as generated by seismic events with different characteristics, such as the magnitude, and with different scenarios of detection, namely with different epicentral distances from the nearest seismic station that first recorded the event. This allows assessing the performance of the algorithm with respect to the different characteristics of both the seismic event and the detection scenario, to identify the most suitable scenario for the application of our algorithm. Those preliminary experiments show that the algorithm performs well in identifying the arrival times of the first P- and S-waves. In particular, the arrival time of the first P-waves is detected more easily than the arrival time of the first S-waves. This is a relevant result because the arrival time of the first P-wave represents the beginning of the seismic event. Furthermore, it is noticed that the post-selection algorithm is not always able to correctly identify the relevant changepoints among the first estimated subset of possible values.

Following these results, in this paper, we aim to present our proposed algorithm's methodology, suitable for the automatic identification of the two relevant phrases in a seismic waveform: the arrival times of the P- and S-waves. To assess the algorithm's performance in different scenarios, we simulate a new richer dataset of waveforms with different magnitudes and epicentral distances. Moreover, to show the advantages of our approach, we compare our results with that obtained, applying a standard Short Time Average over Long Time Average (STA/LTA) algorithm (Allen 1978).

These two algorithms lead to similar results in terms of performance. However, the proposed algorithm is characterized by greater flexibility and automation capacity, as it does not require testing and optimization phases. This peculiarity makes it potentially very useful in earthquakes routine analysis in the case of novel seismic networks, in particular in those areas where earthquake characteristics are unknown. Indeed, features like the window width, threshold and characteristic function may depend on the recording network and on the application. The proposed algorithm just requires to set the maximum number of potential changepoints, denoted as K^*. This, may influence the computational time, as the larger is K^*, the more is the time to estimate the corresponding changepoints, and most of all, the time to compare the set of the reduced models by the used *lars* procedure, for finding the best changepoints. Furthermore, our proposal provides an automatic detection of the arrival time of the P- and S-waves, and therefore, no intervention is needed by the researcher to identify the arrivals. Finally, the proposed algorithm can be easily modified to allow the identification of further seismic phases, such as the end of the seismic event.

All the developed codes are available from the authors.

The structure of the paper is as follow: Sect. 2 presents the new picking algorithm; Sect. 3 reports the testing of the algorithm on a dataset of simulated waveforms; an application to real data is presented in Sect. 4; finally, Sect. 5 contains the conclusions and future works.

2 Methodology: variance piecewise constant models

This section proposes a new methodology for automatic picking of arrival times based on the theory of the variance piecewise constant models.

Adelfio (2012) considers the case of changepoint detection procedure for changes in variation, assuming that the variance function can be described by a piecewise constant function with segments delimited by unknown changepoints.

Let y_i be the outcome and x_i be the observed sample, for $i = 1, 2, \ldots, n$ occasions. Let us assume that $y_i = \mu_i + \epsilon_i$, where μ_i is for instance a sinusoidal function representing the observed signal and $\epsilon_i \sim N(0, \sigma_i^2)$ is an error term. In this context, σ_i^2 is a variance function approximated by a piecewise constant regression function with $K_0 + 1$ segments. An example is shown in Fig. 1.

For simplicity, the model for changes in variance after the k^*th observation is

$$y_i = \begin{cases} \mu_i + \lambda\epsilon_i & 1 \le i \le k^* \\ \mu_i + \tilde{\lambda}\epsilon_i & k^* \le i \le n \end{cases}$$

with λ, $\tilde{\lambda}$, and k^* unknown and

$$\begin{cases} H_0: & \lambda = \tilde{\lambda} \\ H_1: & \lambda \neq \tilde{\lambda} \end{cases}$$

Taking advantage of a generalized linear model formulation of the investigated problem, the test for stepwise changes in the variance of a sequence of Gaussian random variables may be transformed equivalently to the case of testing for changes in the mean of the squared residuals from an estimated linear model that accounts for the mean behaviour of the observed signal. The estimation of the mean signal $\hat{\mu}$ can be carried out by using a standard smoothing procedure, e.g., fitting a cubic smoothing spline to the data. Following a suggestion in Smyth et al. (2001), a gamma generalized linear model (GLM) is fitted with a log-link function, with response given by the squared studentized residuals $s_i = (y_i - \hat{y}_i)^2 / w_i$, with $\hat{y} = \hat{\mu}$ and weights $w_i = 1 - h_i$, where h_i is the ith diagonal element of the hat matrix H. According to this approach, testing H_0 against H_1 means that we are looking for a change in the mean of the residuals from a fitted linear model.

The proposed approach can be considered as a wider version of the *cumSeg* models proposed in Muggeo and Adelfio (2011) for independent normally distributed

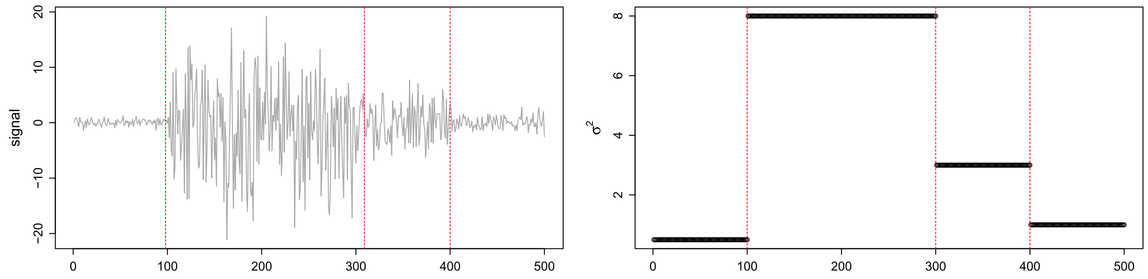


Fig. 1 An example of simulated signal and its corresponding variance with jump points. The red dashed lines indicate the true changepoints

observations with constant variance and piecewise constant means to detect multiple changepoints in the mean of the gene expression levels in genomic sequences by the least-squares approach. The authors assume that the datum $y_i, \forall i$ is defined as the sum of the signal μ_i and noise $\epsilon_i \sim N(0, \sigma_i^2)$ and that μ_i is approximated by a piecewise constant regression function with $K_0 + 1$ segments, that is:

$$y_i = \beta_1 + \delta_1 I(x_i > \psi_1) + \ldots + \delta_{K_0} I(x_i > \psi_{K_0}) + \epsilon_i.$$

Here, $I(\cdot)$ is the indicator function, such that $I(x) = 1$ is x is true, $\boldsymbol{\psi}$ represents the K_0 locations of the changes on the observed phenomenon, β_1 is the mean level for $x_i < \psi_1$, and $\boldsymbol{\delta}$ is the vector of the differences in the mean levels at the change points. The authors proceed to take the cumulative sums of the jump-points model to get a convenient modelling expression that faces the discontinuities at the changepoints ψ_k assuming a piecewise linear or segmented relationship. Therefore, looking for changes in variance, the model is specified as

$$g(\theta_i) = \beta_1 x_i + \delta_1 (x_i - \psi_1)_+ + \ldots + \delta_{K_0}(x_i - \psi_{K_0})_+ \qquad (1)$$

where the term $(x_i - \psi_k)_+$ for the changepoint k is defined as $(x_i - \psi_k) I(x_i > \psi_k)$, and $\theta_i = E[\sum_j^i s_j]$. This model specification has the advantage of an efficient estimating approach via the algorithm discussed in Muggeo (2003, 2008), fitting iteratively the generalized linear model:

$$g(\theta_i) = \beta_1 x_i + \sum_k \delta_k \tilde{U}_{ik} + \sum_k \gamma_k \tilde{V}_{ik}^{-}, \qquad (2)$$

where $\tilde{U}_{ik} = (x_i - \tilde{\psi}_k)_+$, $\tilde{V}_{ik}^{-} = -I(x_i > \tilde{\psi}_k)$. The parameters β_1 and $\boldsymbol{\delta}$ are the same of Eq. (1), while the γ are the working coefficients useful for the estimation procedure Muggeo (2003). At each step the working model in Eq. (2) is fitted and new estimates of the changepoints are obtained via

$$\hat{\psi}_k = \tilde{\psi}_k + \frac{\hat{\gamma}_k}{\hat{\delta}_k}$$

iterating the process up to convergence. $K^*(<K)$ values are returned, producing the fitted model

$$g(\hat{\theta}_i^*) = \hat{\beta}_1 + \hat{\delta} V_{i1} + \ldots + \hat{\delta}_{K^*} V_{iK^*},$$

where $V_{ik} = I(x_i > \hat{\psi}_k)$ for $k = 1, 2, \ldots, K^*$ and the squared residuals are modelled as the response of a gamma GLM with logarithmic link function. Selecting the number of significant changepoints means selecting the significant variables among $V_1, \ldots, V_k$, where K^* is the number of estimated changepoints from model (1). The author solves the model selection problem by using the *lars* algorithm by Efron et al. (2004). Thus, the optimal fitted model with $\hat{K}^* < K^*$ changepoints, is selected by the generalized Bayesian Information Criterion (BIC_{C_n}), that is:

$$BIC_{C_n} = -2 \log L + edf \log(n) C_n$$

where L is the likelihood function, *edf* is the actual model dimension quantified by the number of estimated parameters (including the intercept, the δ and ψ vectors), and C_n is a known constant. The vector of the corresponding selected changepoints is denoted by $\hat{\boldsymbol{\psi}}^*$.

The first issue concerns the value of C_n to be used in the BIC_{C_n} criterion to select the changepoints. In D'Angelo et al. (2020), by simulation, the performance of different specifications of C_n is assessed and, among the different examined specifications of C_n, simulations reveal that $C_n = \log \log n$ has the best performance. Thus, we use this value for the provided analysis.

2.1 The proposed algorithm: changepost

Based on the above methodology, we propose a further algorithm (denoted as *changepost*) to detect, among the estimated changepoints, the two corresponding to the arrival of the first P-wave, and the arrival of the first S-wave. Formally, we define the relevant changepoints to be identified as the true arrival times of the first P- and S-waves, denoted by ψ_1 and ψ_2, respectively (i.e. $K_0 = 2$). In particular, we compare the ratio between the variances of the subsequent phases identified by the $\hat{K}^*$ changepoints $\hat{\boldsymbol{\psi}}^*$ estimated by the main algorithm. The two relevant changepoints are selected as the two ones in correspondence to the two biggest variance ratios. The pseudo-code comes in Algorithm 1.

Algorithm 1 *changepost*

```
Input: $\hat{\boldsymbol{\psi}}^* = \{\hat{\psi}_1^*, \ldots, \hat{\psi}_{\hat{K}^*}^*\}$; $s_i = (y_i - \hat{y}_i)^2/w_i$; $S = \{s_1, \ldots, s_n\}$ ; $start = 1$; $end = n$
Output: $\hat{\boldsymbol{\psi}}$

1: for $(i \;\; in \;\; 1 : \hat{K}^*)$ do
2:     if $(i == 1)$ then
3:         $ratio[i] \leftarrow \frac{\texttt{var}(S[start:\hat{\psi}_i^*])}{\texttt{var}(S[start:\hat{\psi}_{i+1}^*])}$
4:     end if
5:     if $(1 < i < \hat{K}^*)$ then
6:         $ratio[i] \leftarrow \frac{\texttt{var}(S[\hat{\psi}_{i-1}^*:\hat{\psi}_i^*])}{\texttt{var}(S[\hat{\psi}_{i-1}^*:\hat{\psi}_{i+1}^*])}$
7:     end if
8:     if $(i == \hat{K}^*)$ then
9:         $ratio[i] \leftarrow \frac{\texttt{var}(S[\hat{\psi}_{i-1}^*:\hat{\psi}_i^*])}{\texttt{var}(S[\hat{\psi}_{i-1}^*:end])}$
10:     end if
11: end for
12: $ratio[ratio < 1] \leftarrow (ratio[ratio < 1])^{-1}$
13: $\hat{\boldsymbol{\psi}} \leftarrow \hat{\boldsymbol{\psi}}^*[\texttt{which}(\texttt{rank}(ratio) > \hat{K}^* - 2)]$
```

As shown in steps (4), (7), and (10), for each estimated changepoint $\hat{\psi}_i^*$, we compute the ratio between the variance of the interval delimited by the $\hat{\psi}_{i-1}^*$ and $\hat{\psi}_i^*$, and the variance of the interval between $\hat{\psi}_{i-1}^*$ and $\hat{\psi}_{i+1}^*$. Then, in step (14), the two highest ratios suggest which are the corresponding changepoints $\hat{\psi}_1$ and $\hat{\psi}_2$ leading to the most relevant changes in variance.

Figure 2 depicts the changepoints indenfied by *changepost*, on the simulated data of Fig. 1. As the variance of the real signal is $\sigma_i^2 = 0.5 + 8I(i > .2) + 3I(i > .6) + I(i > .8)$, it is evident that *changepost* correctly identifies the changepoints corresponding to the most abrupt changes in the variance.

3 Simulations: evalutating the performance of the algorithm *changepost*

In this section, the proposed picking algorithm for the automatic seismogram onset time determination (simply denoted as *changepost*) is tested on a dataset of simulated waveforms. Simulated seismograms are used to have the maximum control about the arrival times of the P- and S-phases on the waveforms. This aspect is of fundamental importance for correct validation of the algorithm, impossible with experimental seismograms. Indeed, when using real data, i.e. experimental seismograms, it is not possible to know with certainty the arrival times of the seismic phases. Experimental seismograms are recordings of ground motion, or seismic waves, generated at several kilometers in depth and distance. The identification of the arrival times of the seismic phases on experimental seismograms, or the best picking of the P and S waves, is carried out manually by expert seismologists. However, even an expert seismologist may introduce errors and uncertainties in the picking phase. Thus, for controlling the precision and accuracy of an automatic picking procedure, the best practice is to start from simulated data, with well known seismic phases arrival times.

We aim at capturing the performance variations due to some characteristics of both the seismic event and its detection, which in turn affect some characteristics of the

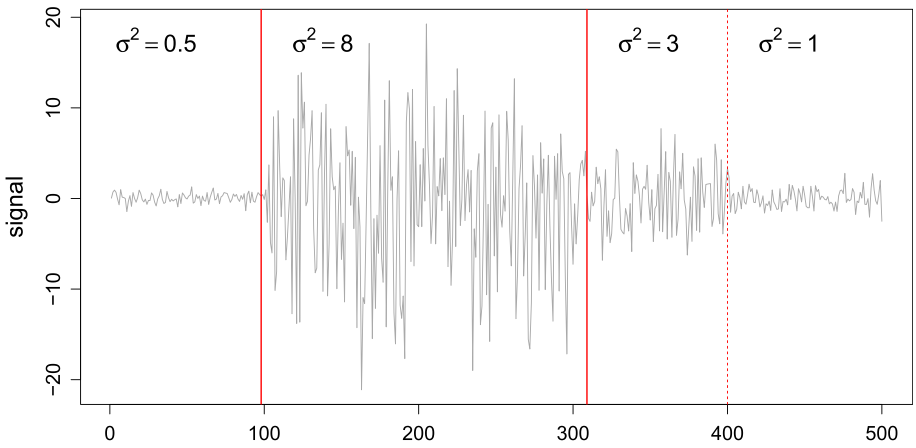


Fig. 2 Simulated data of Fig. 1. The red dashed lines indicate the changepoints identified by the main algorithm, described in Sect. 2. The red straight lines indicate the ones further identified by *changepost*

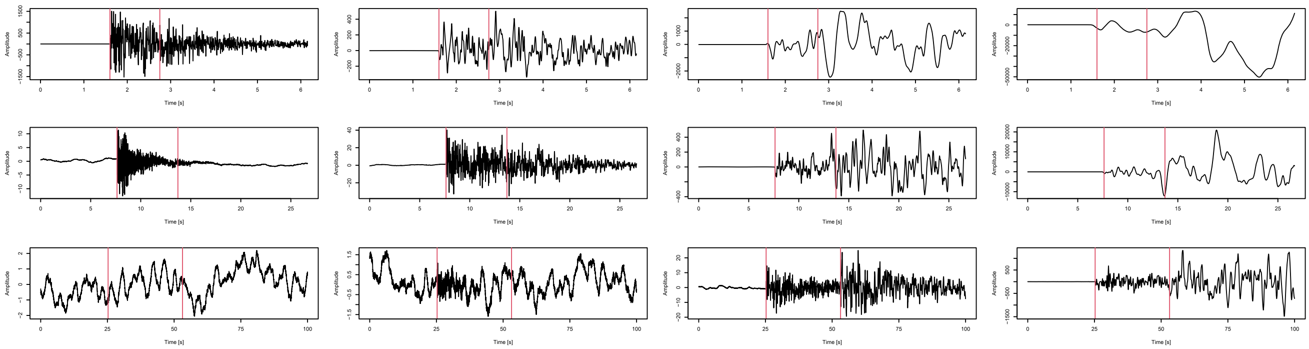


Fig. 3 A simulated waveform for each scenario and true arrival times. *From left to right* increasing magnitude levels. *From top to bottom* increasing distance from the nearest seismic station

waveforms. Therefore, seismic events with different magnitude are simulated, assuming different distances from the nearest seismic station.

Our tests allow highlighting the most general scenarios for the algorithm. Waveforms generated by earthquakes of small magnitude often have energy comparable to the background noise and allow to validate the functioning of the algorithm in case of a low signal-to-noise ratio. Waveforms with different magnitudes and epicentral distances can also differ greatly in terms of frequency content. Events with small in magnitude and with small epicentral distances are generally rich in other frequencies; on the contrary, events with high magnitude and great epicentral distance are also rich in low frequencies. Variations in the epicentral distance also affect the nature of the seismic phases P and S.

The seismic phases, and more generally the shape of the seismomgrams, depend on the epicentral distance. At very small epicentral distances (from a few kilometers to a few tens of kilometers) the seismic waves travel inside the upper crust. The seismic phases coming first to the surface are not undergone to refraction and reflection phenomena; they can be considered as direct waves. At greater distances instead (typically over a 100 km), the seismic phases first emerging on the surface are refracted critically from the upper mantle (Mohorovich discontinuity). Therefore, using three epicentral distances (10, 250 and 50 km), simulations involve the recording of three different types of earthquakes, corresponding respectively to: local events, whose first arrival seismic phases are direct waves, regional events, in which the first seismic phases are seismic waves critically refracted by the upper mantle, and transitional events.

3.1 The simulation setup

The waveforms are simulated as coming from seismic events with different characteristics, referring to altogether 12 scenarios, one for each combination of the following:

- Three distances from the nearest station that recorded the seismic event: 10, 50 and 250 km;
- Four magnitudes: 2, 3, 4 and 5.

For each scenario, 100 waveforms are simulated, all assumed to have impulsive onset of P- and S-waves and standard seismic noise. Moreover, the synthetic signals are generated with a sampling rate of 200 samples per second.

The seismic waveforms are simulated using the deterministic hybrid approach proposed by Mourhatch and Krishnan (2020). In detail, the low-frequency content (limited to a frequency of 0.5 Hz) of the ground motion is generated from a kinematic source model using the open-source seismic wave-propagation package SPECFEM3D (Komatitsch and Tromp 1999; Komatitsch et al. 2004), that implements the spectral-element method, incorporating the regional 3-D wave-speed structure of the earth. Following Mourhatch and Krishnan (2020), low-frequency synthetic SPECFEM3D seismograms are combined with high-frequency seismograms generated using a variant of the classical EGF (Empirical Green's Function) approach.

For each seismic event, we generate all three components of motion (i.e. North–South, East–West and Vertical). In our analysis, we only report the results for the Vertical component, for the sake of brevity. In Fig. 3, an example of waveforms (vertical movement component) for each of the considered scenarios with highlighted the true P- and S arrival times is shown.

3.2 Simulation results for *changepost*

Table 1 reports the empirical means (m) and Mean Squared Error values (s) of the two relevant changepoints estimated by the proposed algorithm over the 100 waveforms coming synthetic seismic events, for the four different Magnitudes and three epicentral distances. For each epicentral distance, we assume different true arrival times (in blue). Along with the mean and the mean squared error values, we also

Table 1 Empirical means (m) and Mean Squared Error values (s) of the two relevant changepoints detected by *changepost*, over the 100 waveforms of each simulated dataset, with four different magnitudes and three epicentral distances

M		ψ_1	ψ_2	NA%
		10 km		
True		41.6	42.75	
2	m	41.415	43.101	7
	s	0.107	2.256	
3	m	41.514	42.708	1
	s	0.017	1.604	
4	m	41.496	42.278	2
	s	0.012	1.355	
5	m	41.382	45.751	1
	s	0.055	1.081	
M		ψ_1	ψ_2	NA%
		50 km		
True		47.63	53.714	
2	m	47.592	53.394	0
	s	0.002	5.168	
3	m	47.453	57.686	0
	s	0.240	49.011	
4	m	47.229	53.194	0
	s	0.789	19.630	
5	m	47.118	55.537	0
	s	0.376	45.036	
M		ψ_1	ψ_2	NA%
		250 km		
True		75.26	103.15	
2	m	–	–	100
	s	–	–	
3	m	–	–	100
	s	–	–	
4	m	75.115	105.626	0
	s	7.904	80.144	
5	m	74.192	103.592	1
	s	1.322	434.865	

compute the percentage of waveforms where no changepoint is estimated.

Overall we may notice that, as expected, the *changepost* algorithm performs the best as the distance from the nearest seismic station that recorded the event decreases and as the magnitude of the seismic event increases. This is the case with the best signal to noise ratio. Indeed, the NA values are most likely to occur when the magnitude is small and the distance is large, that is basically when the P- and S-waves have comparable or lower energy with respect to the background noise, that is indiscernible from that. In such cases, the arrival times can not be estimated correctly. The scenarios in which the distance from the nearest seismic stations is 50 km is the one reporting no NAs, regardless of the magnitude level. Nevertheless, this does not represent the best picking scenario, as the uncertainty of the estimates is larger than the performance in the 10 km scenario.

3.3 Comparison with STA/LTA

In this paragraph, we compare the *changepost* picking algorithm, based on the variance piecewise constant models, introduced in Sect. 2, with the Short-Term Average/Long-Term Average (STA/LTA) method (Allen 1978).

The STA/LTA method is the simplest and most commonly picking technique used in earthquake seismology. The STA/LTA method computes the ratio of the continuously computed average energy (generally the waveforms envelope, the absolute amplitude, or other characteristic functions) of a recorded trace in two synchronous moving-time windows: a Short-Term window and a Long-Term window (STA/LTA ratio). The short-time window permits to highlight sudden amplitude changes in the signal, while the long time one estimates the current average of the seismic noise. Therefore, the STA/LTA ratio allows highlighting variations in energy in the signal with respect to the background noise. These energy variations can be identified by setting thresholds: when the STA/LTA ratio exceeds a certain threshold, the arrival of a seismic phase is identified. The output of the STA/LTA algorithm is the characteristic function E_k, defined as:

$$E_k = x_k^2 + (x_k')^2 + C$$

where, x_k is the seismic trace, x_k' is its derivative and C is an empirical weighting constant.

This method is undoubtedly computationally efficient, and its variants are widely used for the picking of seismic phases. However, it needs a calibration phase to identify both the best length of the STA and the LTA and the best threshold level. The optimal STA width depends on the frequency content of the seismic event and, therefore, on its magnitude and epicentral distance. Similarly, the width of the LTA should also be chosen according to the noise characteristics. The trigger threshold is also very important: values that are too high can lead to the failure to identify the arrival of the seismic phases, values that are too low can provide false identifications. This method can therefore be inaccurate in the case of a low signal to noise ratio.

After several optimization tests for each earthquake class, we set the parameters reported in Table 2 for the comparison. Once the parameters are set, we run the tests,

Table 2 STA/LTA settings

STA/LTA parameters (s)	10 km	50 km	250 km
STA window length	0.1	0.5	2.5
LTA window length	1	5	25
Threshold trigger on	5.0	5.0	5.0
Threshold trigger off	2.5	2.5	2.5

Table 3 Empirical means (m) and Mean Squared Error values (s) of the two relevant arrival times estimated by the STA/LTA algorithm over the 100 waveforms of each simulated dataset, with four different magnitudes and three distances

M		ψ_1	ψ_2	NA%
		10 km		
True		41.6	42.75	
2	m	41.605	–	0
	s	0.000	–	
3	m	41.589	42.272	0
	s	0.003	0.413	
4	m	41.573	42.853	0
	s	0.000	0.140	
5	m	41.514	42.740	0
	s	0.009	0.291	
		50 km		
True		47.63	53.714	
2	m	47.682	49.148	1
	s	0.018	35.847	
3	m	47.620	47.642	0
	s	0.034	36.873	
4	m	47.617	53.632	0
	s	0.006	0.430	
5	m	47.589	53.472	0
	s	0.006	2.193	
		250 km		
True		75.26	103.15	
2	m	78.275	129.137	82
	s	9.145	989.601	
3	m	82.935	129.166	90
	s	58.905	922.363	
4	m	75.809	106.600	0
	s	0.655	13.357	
5	m	75.351	107.340	0
	s	0.033	21.231	

Table 4 Percentage of the changepoints estimated by the *changepost* algorithm, lying within a ±0.2 interval around the true value

	M	ψ_1	NA%	ψ_2	NA%
10 km	2	0.540	8	0.000	100
	3	0.740	1	0.020	96
	4	0.650	2	0.070	86
	5	0.200	69	0.115	77
50 km	2	0.500	0	0.020	96
	3	0.500	0	0.010	98
	4	0.300	43	0.020	96
	5	0.165	67	0.005	99
250 km	2	–	100	–	100
	3	–	100	–	100
	4	0.130	74	0.145	71
	5	0.010	98	0.010	98

obtaining the results reported in Table 3. We have noticed that the number of the picked arrival times is generally lower using the STA/LTA, with respect to the *changepost* algorithm. Then, a further note is needed. Before computing Mean, MSE and NA%, we set the picking's STA/LTA results in this way: for each seismic event, we check if there are zero picking, one or more than one picking. If zero picking is observed, we only increment the NA%; instead, if we find one or more than one picking, the closest picking to the real picking is determined. Just after finding all picking values for all events, we compute Mean and MSE for picked events, and NA%, as for the proposed algorithm. This specification is due since there are a number of cases where the STA/LTA algorithm picks a unique arrival time (see the first scenario—S-waves—in Table 3).

Nevertheless, if the number of the estimated picked arrival times $\hat{K}^*$ is large, the probability of having a picking close to the true one increases, resulting in smaller mean squared error values and then, influencing the results as mentioned above. Therefore, we define a probability index, computed for each waveform q: let $\hat{\boldsymbol{\psi}}^*_q$ be the vector of the $\hat{K}^*_q > 1$ estimated changepoints, the probability index is defined as follows:

$$\frac{I(\hat{\boldsymbol{\psi}}^*{}_q \leq true \pm pick)}{\hat{K}^*_q}, \tag{3}$$

where $I(\cdot)$ is the indicator function, such that $I(x) = 1$ if x is true, i.e. counting how many times the changepoints $\hat{\boldsymbol{\psi}}^*_q$ estimated for the waveform q fall inside the interval $\pm pick$ around the true arrival time (which varies with the scenario considered).

In Tables 4 and 5 we report the computed probability index (3), for the proposed picking algorithm and for the

Table 5 Percentage of the changepoints estimated by the STA/LTA algorithm, lying within a ±0.2 interval around the true value

	M	ψ_1	NA%	ψ_2	NA%
10 km	2	1.000	0	0.000	100
	3	0.935	1	0.020	96
	4	0.433	0	0.253	38
	5	0.551	0	0.150	63
50 km	2	0.805	12	0.000	100
	3	0.935	5	0.000	100
	4	0.512	1	0.128	74
	5	0.464	0	0.112	73
250 km	2	0.000	100	0.000	100
	3	0.000	100	0.000	100
	4	0.170	81	0.000	100
	5	0.456	14	0.000	100

STA/LTA algorithm, respectively. In particular, we set the pick to 0.2 s, and compute (3) for both the P- and S-arrivals, separately. Then, those values are averaged with respect to the 100 waveforms of each scenario, and the percentage of NAs is reported, to take into account both the waveforms where no changepoint is estimated (i.e. NAs in Tables 1 and 3) and those cases in which $I(\hat{K}_q^* \leq true \pm pick) = 0$, i.e. no estimated changepoint for that specific waveform falls into the interval.

From Tables 4 and 5, we may notice that STA/LTA outperforms *changepost* in picking the P-Phases times. Almost in the all S-Phases, instead, STA/LTA provides the highest NA%: the scenario with distance 10 km and magnitude 2 is the only one where *changepost* provides 100% of NAs. Moreover, for the scenario 250 km distance and magnitude 2, *changepost* does not find any changepoint and then provides the highest NA%; this percentage gradually decrease as the magnitude increases.

When the distance is 50 km, even in lower magnitudes, the *changepost* algorithm provides a lower percentage of NA than STA/LTA. Overall, *changepost* outperforms STA/LTA in the S-Phases picking. Otherwise, the STA/LTA picking time is better for the P-Phases, being more precise in terms of tenths of a second.

A comment on the computational cost is in order. Indeed, computation time is crucial in automated seismogram onset time determination, mostly accounting for its implications in seismic monitoring and in earthquake early warning systems. Even tough we have assessed that *changepost* is quite slower than STA/LTA, the computational time of the former does not represent a limitation. The only setting influencing the computational time is K^*, that is the maximum number of changepoints to be detected: the larger K^*, the higher the computational time, but also the more the estimated changepoints. Therefore, since *changepost* is able to process hundreds of waveforms within minutes, the researcher could even consider to reproduce the analyses with different values of K^*, depending on the available time and the complexity of the waveforms. Therefore, even tough STA/LTA has lower computational time, the automation of *changepost* counterbalance its higher computational cost.

4 Application to real data

The seismic events selected for showing an application to real data belong to the seismic sequence of L'Aquila (Italy), in 2009. The complete database is made up of 80 seismic events recorded by 12 stations with three components (identify the component of motion to which they refer: Up–Down, North–South and East–West), for a total of 2880 waveforms. They all exhibit a magnitude between 3 and 4.1. The waveforms were sampled at 100 Hz, and the length total of each of the waveforms is 100 s (10,000 samples). To increase the signal ratio noise, a bandpass filter was applied in frequency band between 0.1 and 35 Hz. Such an operation was necessary to eliminate frequencies related to electronic and anthropic noise clearly not part of seismic signals. Also, the waveforms have been normalized with respect to the maximum amplitude.

Figures 4 and 5 contain five waveforms selected to show the results, and the arrival times identified by *changepost* and STA/LTA, respectively. For the proposed *changepost* procedure, K^* is set to 10. In Fig. 4, the red dashed lines represent all the $\hat{K}^*$ change-points detected by the main algorithm, while the solid red ones identify the two selected change-points, representing the arrival times of the P- and S-waves, respectively. Figure 5 depicts the results of the application of STA/LTA: the solid red lines indicate the estimated arrival times. As evident from the two figures, while *changepost* is always able to identify two arrival times (most likely to represent the arrivals of the P- and S-waves), the STA/LTA algorithm either succeeds in identifying only very early arrival times (very likely to be arrival times of the P-waves) or completely fails to identify any arrival time.

This is an expected result for the seismic events considered in this experiment, in particular using the STA/LTA settings reported in Table 2. Better results for the STA/LTA algorithm, comparable with those just showed with synthetic seismograms, could be obtained only after a few rounds of optimization of the triggering parameters. These results confirm the conclusion drawn by simulation study, that is the high flexibility of the proposed *changepost* algorithm. Indeed, it does not require neither testing

Fig. 4 Vertical components of three seismic events detected by the seismic station BSSO—Busso (Italy) during the L'Aquila seismic sequence occurred in 2009. *Red dashed lines* all the $\hat{K}^*$ change-points estimated by the main algorithm. *Red straight lines* the $\hat{K} = 2$ change-points, among the $\hat{K}^*$ estimated ones, most likely to represent the true arrival times of the P- and S-waves

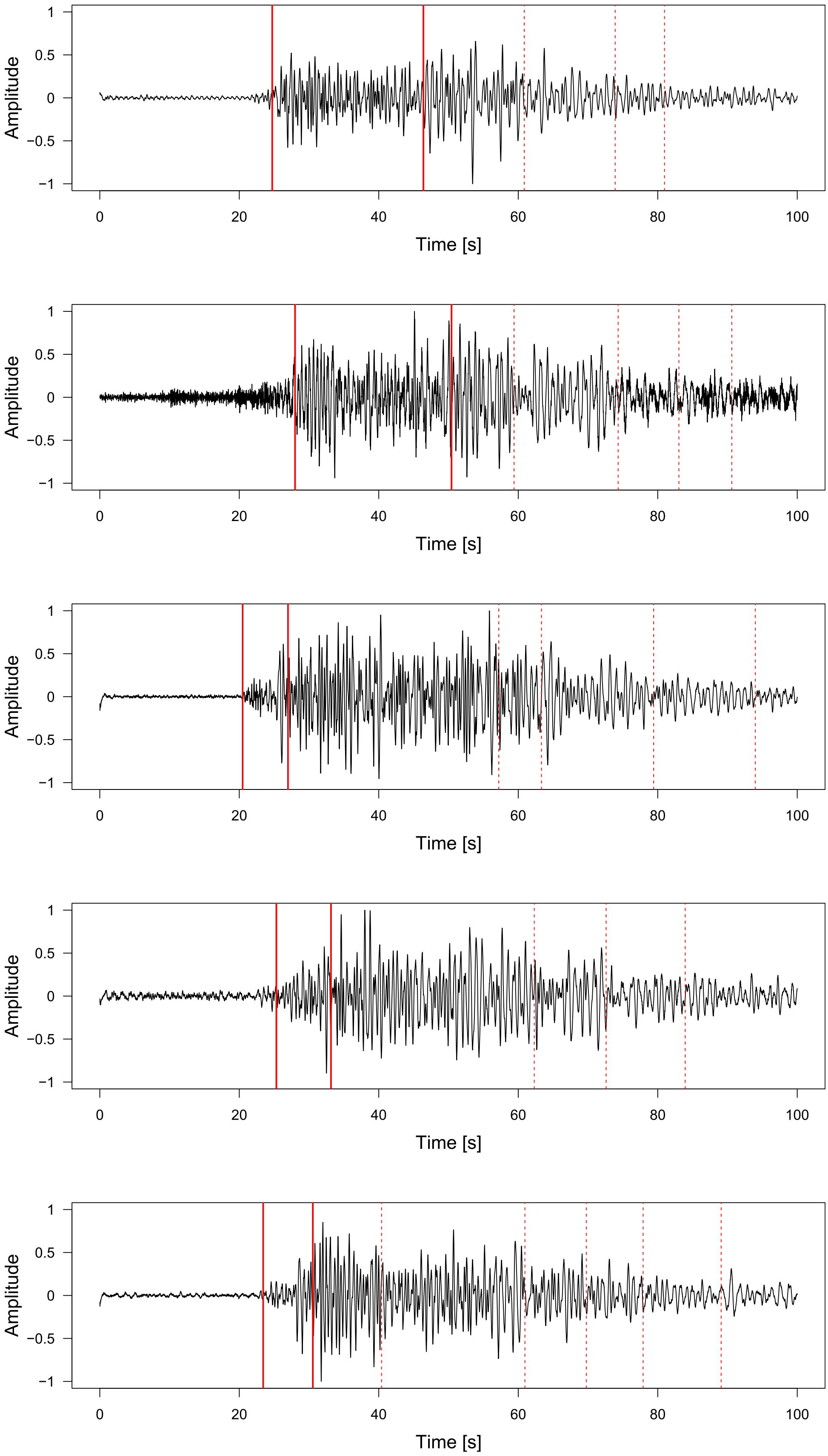
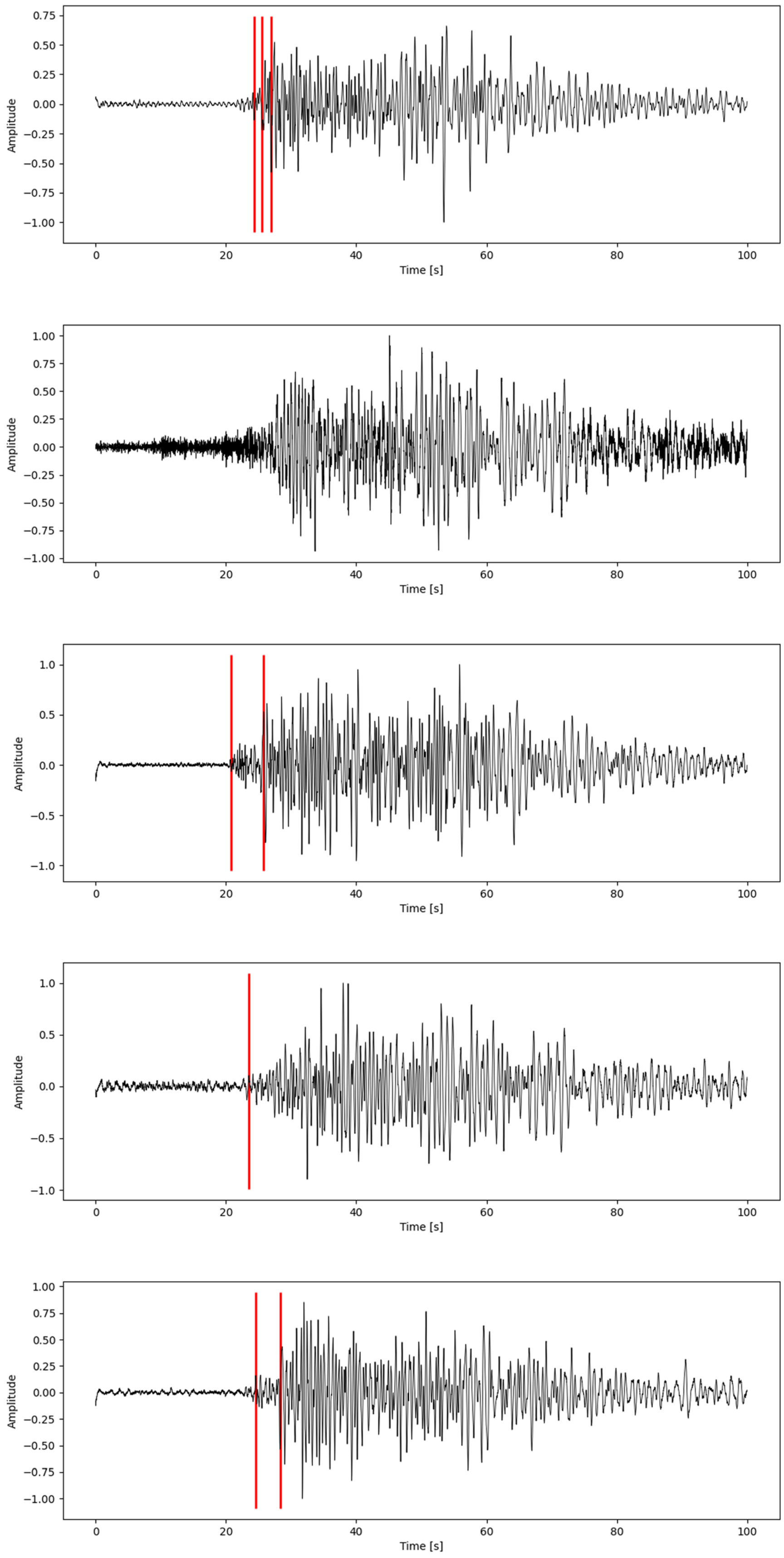


Fig. 5 Vertical components of three seismic events detected by the seismic station BSSO—Busso (Italy) during the L’Aquila seismic sequence occurred in 2009. *Red straight lines* the arrival times identified by STA/LTA



nor optimization phase, and its accuracy is almost independent of the analyzed dataset.

5 Conclusions

The precise and quick determination of the arrival times of the main seismic phases is of fundamental importance for seismic surveillance and routine earthquake hypocenter determination. Clearly, to be suitable also for early warning, a picking algorithm must be computationally efficient, avoid false alarms, and time picking must be as accurate as possible.

With these premises, in this work, we proposed a novel picking algorithm for the automatic P- and S-waves onset time determination. The *changepost* algorithm is based on the variance piecewise constant models. The effectiveness and robustness of our picking algorithm were tested on synthetic seismograms. In order to make the STA/LTA algorithm work correctly, many tests are necessary to optimize the processing parameters. These parameters (window width, threshold, characteristic function) are clearly a function of the type of seismicity recorded by the network and must be optimized from time to time.

If compared to the well-established STA/LTA offline picking algorithm, the *changepost* opens a promising path. Indeed, *changepost* is entirely automatic, meaning that no choice of any parameters is needed to run the algorithm. This feature can be particularly important when, for example, the characteristics of the seismicity of a given area are not well known or when a new seismic monitoring network is set up. The only prior setting regards the maximum number of changepoints to be detected: the larger the number, the more the resulting estimated changepoints, but also the higher the computational time. Furthermore, *changepost* provides automatically the arrival time of the P- and S-waves, and therefore, no intervention is needed by the researcher to identify the arrivals among those possibly triggered.

Changepost can be easily modified to allow the identification of further seismic phases, such as the end of the seismic event. Certainly, interesting results can be obtained by applying the same technique to transforms of the original signal (integrated signal, derivative, frequency filtered, etc.). These future developments that we are foreseeing could certainly improve the performance of the proposed algorithm.

Funding The authors have not disclosed any funding. This work was supported by 'FFR 2021 GIADA ADELFIO'.

Data availability Data are available from the corresponding author.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethical approval Not applicable.

References

Adelfio G (2012) Change-point detection for variance piecewise constant models. Commun Stat Simul Comput 41(4):437–448

Adelfio G, Chiodi M, D'Alessandro A, Luzio D, D'Anna G, Mangano G (2012) Simultaneous seismic wave clustering and registration. Comput Geosci 44:60–69

Akaike H (1975) Markovian representation of stochastic processes by canonical variables. SIAM J Control 13(1):162–173

Akaike H (1998) Autoregressive model fitting for control. In: Selected papers of Hirotugu Akaike. Springer, Cham, pp 153–170

Aldersons F (2004) Toward three-dimensional crustal structure of the Dead Sea region from local earthquake tomography. PhD thesis

Allen RV (1978) Automatic earthquake recognition and timing from single traces. Bull Seismol Soc Am 68(5):1521–1532

Allen R (1982) Automatic phase pickers: their present use and future prospects. Bull Seismol Soc Am 72(6B):S225–S242

Baer M, Kradolfer U (1987) An automatic phase picker for local and teleseismic events. Bull Seismol Soc Am 77(4):1437–1445

Chernoff H, Zacks S (1964) Estimating the current mean of a normal distribution which is subjected to changes in time. Ann Math Stat 35(3):999–1018

D'Angelo N, Adelfio G, D'Alessandro A, Chiodi M (2020) A fast and efficient picking algorithm for earthquake early warning application based on the variance piecewise constant models. In: International conference on computational science and its applications. Springer, Cham, pp 903–913

D'Angelo N, Adelfio G, D'Alessandro A, Chiodi M (2021) Evaluating the performance of a new picking algorithm based on the variance piecewise constant models. In: 50th meeting of the Italian Statistical Society

Efron B, Hastie T, Johnstone I, Tibshirani R et al (2004) Least angle regression. Ann Stat 32(2):407–499

Gardner L (1969) On detecting changes in the mean of normal variates. Ann Math Stat 40(1):116–126

Hartung J, Elpelt B, Klösener K-H (2014) Statistik: Lehr-und Handbuch der angewandten Statistik. Walter de Gruyter GmbH & Co KG, Berlin

Hawkins D (1992) Detecting shifts in functions of multivariate location and covariance parameters. J Stat Plan Inference 33(2):233–244

Komatitsch D, Tromp J (1999) Introduction to the spectral element method for three-dimensional seismic wave propagation. Geophys J Int 139(3):806–822

Komatitsch D, Liu Q, Tromp J, Suss P, Stidham C, Shaw JH (2004) Simulations of ground motion in the Los Angeles basin based upon the spectral-element method. Bull Seismol Soc Am 94(1):187–206

Küperkoch L, Meier T, Lee J, Friederich W, Group EW (2010) Automated determination of P-phase arrival times at regional and local distances using higher order statistics. Geophys J Int 181(2):1159–1170

Küperkoch L, Meier T, Diehl T (2012) Automated event and phase identification. In: New manual of seismological observatory practice 2 (NMSOP-2). pp 1–52

Morita Y (1984) Automatic detection of onset time of seismic waves and its confidence interval using the autoregressive model fitting. Earthquake 37:281–293

Mourhatch R, Krishnan S (2020) Simulation of broadband ground motion by superposing high-frequency empirical Green's function synthetics on low-frequency spectral-element synthetics. Geosciences 10(9):339

Muggeo VM (2003) Estimating regression models with unknown break-points. Stat Med 22(19):3055–3071

Muggeo V (2008) Segmented: an R package to fit regression models with broken-line relationships. R News 8(1):20–25

Muggeo VM, Adelfio G (2011) Efficient change point detection for genomic sequences of continuous measurements. Bioinformatics 27(2):161–166

Scherbaum F (1992) PITSA user guide (Programmable interactive toolbox for seismic analysis). IASPEI Software Library, 5

Sleeman R, Van Eck T (1999) Robust automatic P-phase picking: an on-line implementation in the analysis of broadband seismogram recordings. Phys Earth Planet Inter 113(1–4):265–275

Smyth GK, Huele AF, Verbyla AP (2001) Exact and approximate REML for heteroscedastic regression. Stat Model 1(3):161–175

Takanami T, Kitagawa G (1988) A new efficient procedure for the estimation of onset times of seismic waves. J Phys Earth 36(6):267–290

Wang L, Wang J (2006) Change-of-variance problem for linear processes with long memory. Stat Pap 47(2):279

Wichern DW, Miller RB, Hsu D-A (1976) Changes of variance in first-order autoregressive time series models-with an application. J R Stat Soc Ser C (Applied Statistics) 25(3):248–256

Worsley K (1979) On the likelihood ratio test for a shift in location of normal populations. J Am Stat Assoc 74(366a):365–367

Zhao W, Tian Z, Xia Z (2010) Ratio test for variance change point in linear process with long memory. Stat Pap 51(2):397–407