A Fast and Efficient Picking Algorithm for Earthquake Early Warning Application Based on the Variance Piecewise Constant Models

Nicoletta D'Angelo[1(✉)], Giada Adelfio[1,2], Antonino D'Alessandro[2], and Marcello Chiodi[1,2]

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy
nicoletta.dangelo@unipa.it
[2] Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Abstract. An earthquake warning system, or earthquake early warning system, is a system of accelerometers, seismometers, communication, computers, and alarms that is devised for notifying adjoining regions of a substantial earthquake while it is in progress. This is not the same as earthquake prediction, which is currently incapable of producing decisive event warnings. The implementation of efficient and computationally simple picking algorithm is necessary for this purpose, as well as automatic picking of seismic phases for seismic surveillance and routine earthquake location for fast hypocenter determination. In this paper, a picking method, based on the detection of signals changes in variance, is proposed, taking advantage of a generalized linear model formulation of the investigated problem. An application to simulated data is provided.

Keywords: Earthquake Early Warning · Picking · Change-points

1 Introduction

The term "Earthquake Early Warning" (EEW) is used to describe real-time earthquake information systems that have the potential to provide a warning before significant ground shaking. Warning times range from a few seconds to a little more than a minute and are primarily a function of the distance of the user from the earthquake epicentre. In EEW system, generally, two approaches could be followed, called an onsite and regional warning. The principle of onsite or single-station warning [7] is to detect seismic energy at a single location and provide warning of coming ground shaking at the same location, i.e., detect the P-wave and predict the peak shaking. The regional warning makes use of a seismic network and typically combines information derived by several stations. Generally speaking, the EEW systems first detects earthquakes, in particular, P-wave first arrival and then transmits a useful warning. Given that the strongest

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12255, pp. 903–913, 2020.
https://doi.org/10.1007/978-3-030-58820-5_65

ground shaking usually arrives at the time of, or after, the S-wave arrival, using the P-wave to provide warning has the potential to increase the warning time everywhere, to reduce the radius of the blind zone, and to potentially provide a warning at the epicentre. It is therefore important for a robust EEW system, the implementation of efficient and computationally simple picking algorithm. Automatic picking of seismic phases is also important in seismic surveillance and routine earthquake location for fast hypocenter determination. To be suitable for both application, a false alarm must be avoided and time picking must be as accurate as possible.

In this work, we propose a new automatic picking algorithm suitable for the implementation of EEW and in seismic surveillance. The algorithm, based on changes in variance, is tested on synthetic seismograms.

This paper is structured as follows: first, in Sect. 2, an overview about the most widespread automatic picking algorithms is reported; secondly, the description of the proposed methodology is reported in Sect. 3; finally, in Sect. 4, the variance piecewise constant models are applied to seismic waveforms; the last section is devoted to conclusions and final remarks.

2 Automated Seismogram Onset Time Determination

For correct early warning, it is fundamental to recognize early the beginning of a seismogram employing fast and robust algorithms. This is important because the algorithm must run in real-time and false alarm or lost event should be avoided. 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 interpreted as the first P-phase arrival time, which is, of course, afflicted 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 source 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 networks, where the number of stations and the aperture of the network is not large. For regional and global networks this simple event detection algorithm has to be modified.

[12, Chapter 16] review the most widespread automatic picking algorithms and analyze their properties. Here we report a brief overview. It is also worth to notice that comparative works among different pickers have been carried out in literature [4,11,19].

[5,6] introduce the concept of characteristic function (CF), where the 'character' of the seismic trace is specified. The CF is obtained by one or several non-linear transformations of the seismogram and should increase abruptly at

the arrival time of a seismic wave. In addition to the calculation of the CF the next steps of a picking algorithm are the estimation of the arrival time from the CF and the quality estimation.

The Allen picker is a fast and robust algorithm, which also accounts for automatic quality assessment. However, as this algorithm is amplitude based only, it might miss emergent P-onsets. A comparative study by [11] 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 [8]. This algorithm is frequently applied, e.g. by 'Programmable Interactive Toolbox for Seismological Analysis' (PITSA, [17]) and the picking system MannekenPix [4].

In contrast to Allen's squared envelope function, this CF is sensitive to changes in amplitude, frequency as well as in phase.

The Baer and Kradolfer picker is also very fast and robust and quite user-friendly, as this algorithm only needs 4 input parameters. A shortcoming of this algorithm is the missing automated quality assessment. Several comparative studies [4,11,19] show how this picking algorithm tends to be somewhat late compared to manual P-picks.

When an earthquake signal occurs, the statistical properties of a seismogram change abruptly. Therefore, the measurement of statistical properties in a moving window are suitable for the determination of a CF and subsequent estimation of arrival times. 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 [10].

Though just amplitude-based, higher order statistics are quite sensitive even to emergent P-onsets. In combination with a sophisticated picking algorithm (e.g. [11]), which exploits the entire information provided by the determined CF, it yields excellent results. If precisely tuned, the automated quality assessment proposed by [11] gives similar weights as the analysts. However, choosing the parameters for this sophisticated algorithm is quite difficult and needs a great experience.

The so called autoregressive-Akaike-Information-Criterion-piker (AR-AIC) proposed by [19] is based on the work by [2,3,13] and [21].

A longer time series is divided into two locally stationary segments each modelled by an autoregressive (AR) process. The first segment represents noise, the second segment contains the signal. After estimating the two sets of AR parameters, two prediction errors are computed and then the minimum of the two-model Akaike-Information-Criterion (AIC) indicates the arrival time.

The AR-AIC picker 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.

3 Variance Piecewise Constant Models

[1] 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.

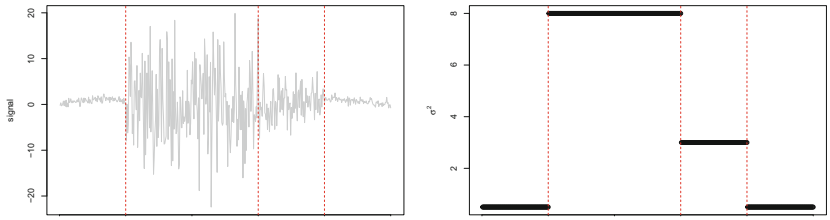


Fig. 1. An example of simulated signal and its corresponding variance with jump points

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 temr. 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

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

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 variance of a sequence of Gaussian random variables may be transformed equivalently to the case of testing for changes in 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 common smoothing procedure, e.g., fitting a cubic smoothing spline to the data. Following a suggestion in [20], 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 [16] for independent normally distributed 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

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

where $\theta_i = E[\sum_j^i s_j]$ and it has the advantage of an efficient estimating approach via the algorithm discussed in [14,15], fitting iteratively the generalized linear model:

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

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 [15]. 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 [9]. Thus, the fitted optimal model with $\hat{K} \leq K^*$ changepoints is selected by the generalized Bayesian Information Criterion (BIC_{C_n}), defined by:

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

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.

4 Application to Seismic Waveforms

The choice of C_n. The first issue concerns the value of C_n to used in the BIC_{C_n} criterion to select the changepoints. Therefore, by simulation, we assess the performance of different specifications of C_n. Assuming for simplicity $x_i = i, \quad i = 1, \ldots, 700$, the true variance used for the simulations is $\sigma_i^2 = 0.5 + 10I(i > 175)$, while the true mean is not specified, as the mean behaviour of the signal is assumed to be unknown. We consider BIC_{C_n} with different values of C_n and we report the results in Table 1.

Table 1. Empirical means and Mean Squared Error values of the detected number of changepoints over 500 runs and for $K_0 = 1$

	C_n			
	1	$\log(n)$	$\log\log(n)$	$2\log\log(n)$
Mean	0.94	0.96	0.96	0.94
mse	0.12	0.07	0.04	0.06

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.

The Simulated Signals. To test the performance of the algorithm, a dataset consisting of 100 waveforms over 60 s is simulated. The changepoints are three, equal to 10, 12, and 22 s. The first changepoint represents the arrival of the first P-wave, the second one represents the arrival time of the fist S-wave and the last one corresponds to the end of the seismic event [18].

The simulated dataset is shown in Fig. 2. Each signal has the same changepoints but different variances corresponding to each phase.

First, the algorithm is applied to the entire dataset. For each waveform, we obtain a different set of changepoints, estimated along with the signal. We report the output of the procedure applied to the first waveform of the dataset in Fig. 3.

The dashed red lines are placed in correspondence of the changepoints estimated through the algorithm. As we may notice, the proposed algorithm succeeds in the identification of the arrival of both the P-wave and S-wave. The straight lines are three changepoints closer to the true values (10, 12 and 22 s), and therefore they represent the changepoints, among all the estimated ones, that a further algorithm should pick as the three relevant changepoints.

In Table 2 the summary statistics of the three relevant changepoints estimated along the whole dataset are reported. We may see that the mean values are close to the true values and that the smallest mean squared error is the one of the first changepoint, that is, the arrival of the first P-wave. Indeed, the estimates of the other two changepoints are less precise. This effect can be related to the occurrence of the first changepoint in correspondence of the most abrupt change in variation of the signal, making it easier to detect.

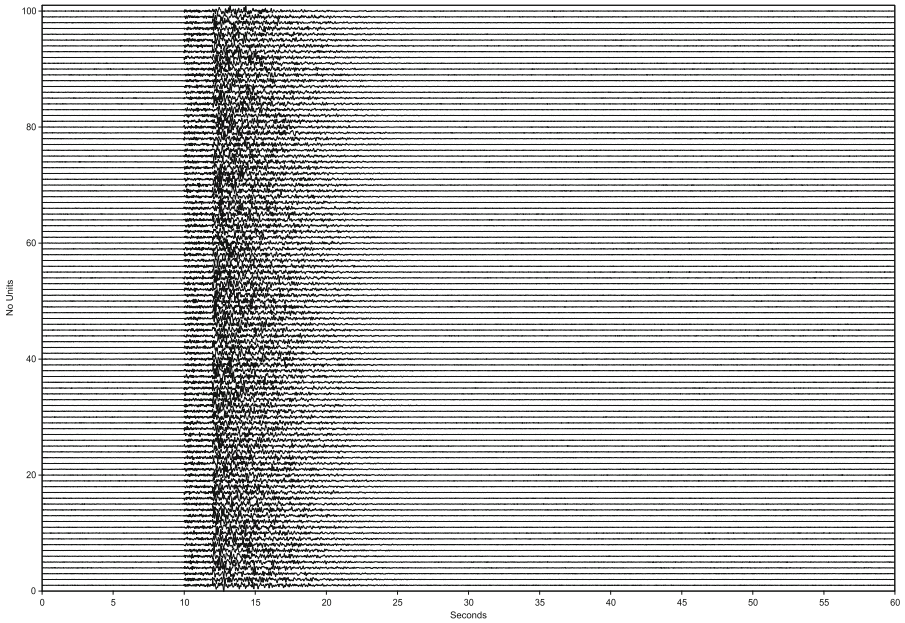


Fig. 2. Simulated dataset of 100 signals

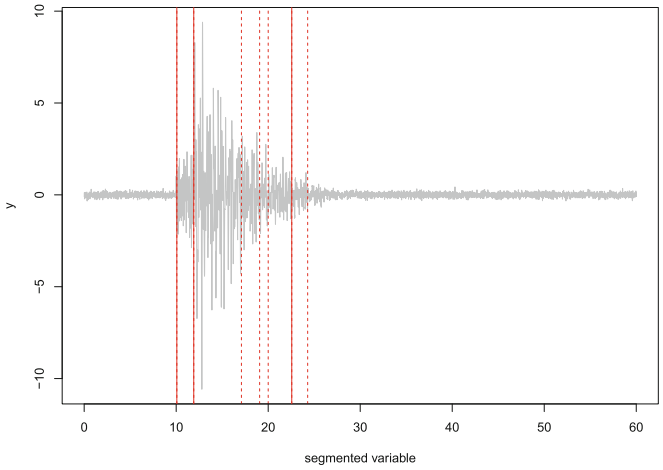


Fig. 3. Changepoints detection in the first seismic waveform of the simulated dataset (Color figure online)

Table 2. Empirical means and Mean Squared Error values of the three relevant changepoints over the 100 waveforms of the simulated dataset, obtained through the application of the algorithm

	ψ_1	ψ_2	ψ_3
Mean	9.90	11.79	22.09
mse	0.03	0.44	0.46

Simulated Trend. Then, to test the performance of the algorithm in different settings, further analysis is proposed.

The analyses are carried out assuming underlying trends equal to $\mu_1 = \sin \pi x$ and $\mu_2 = 0.4 \sin 100\pi x$, where x in equally spaced in the range $[1 : 6000]$. These are superimposed and added to the simulated signals. In Fig. 4 (left panel) the first waveform of the simulated dataset is shown, together with the simulated trends, plotted in red. Therefore, the algorithm can account for a mean behaviour of the observed signal, without any loss of precision. In particular, two different models are considered to take into account the mean behaviour of the signal. The first is linear and the second is obtained fitting a smooth spline with 100 knots. These are represented in Fig. 4 (right panel) in red and green, respectively.

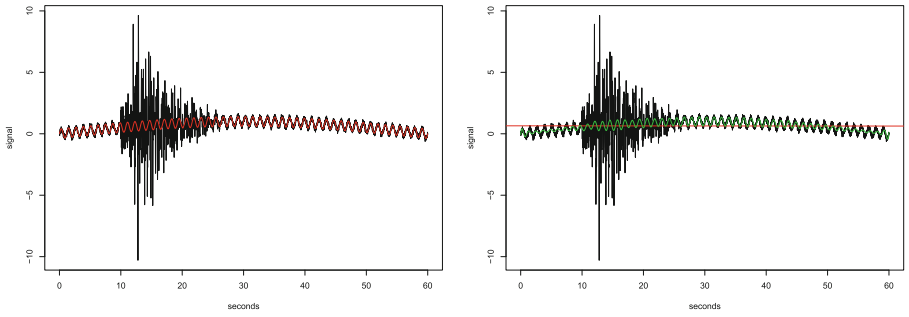


Fig. 4. Simulated data with trend (on the left) and two mean behaviours estimated to take into account the trend (on the right), byb a linear model in red and by splines in green, respectively. (Color figure online)

In Table 3 the results of the application along with the whole dataset, to which the mean behaviour is added, are shown.

We may notice that the estimates of the relevant changepoints are more precise when taking into account the mean behaviour of the signal through a smooth spline, rather than ignoring it. Indeed, as we may notice from Table 3, the mean values are closer to the true changepoints and the Mean Square Error Values are smaller.

Table 3. Empirical means and Mean Squared Error values of the three relevant changepoints estimated through the application of the algorithm to the original simulated data plus the trend

		ψ_1	ψ_2	ψ_3
linear	Mean	9.91	11.94	21.97
	mse	0.10	0.34	0.58
spline	Mean	9.90	11.75	22.08
	mse	0.03	0.15	0.47

Post Selection. Finally, we propose a further algorithm to detect, among the estimated changepoints, the three ones corresponding to the arrival of the first P-wave, the arrival of the first S-wave and the end of the seismic event. In particular, we compare the ratio between the variances of subsequent phases of the signal and select only the biggest three. As far as the first algorithm, we fit a smoothing spline to account for the mean behaviour of the signal (see Fig. 5 and Table 4). Of course, the post-selection adds uncertainty to the estimates obtained through the application of the algorithm, so the Mean Squared Error values obtained are the sum of the Mean Squared Errors from the true values plus the uncertainty due to the post-selection. Therefore, as the Mean Squared Error value referred to the first changepoint is equal to the one of the proposed procedure, this means that the first changepoint is always correctly selected by the post-selection algorithm.

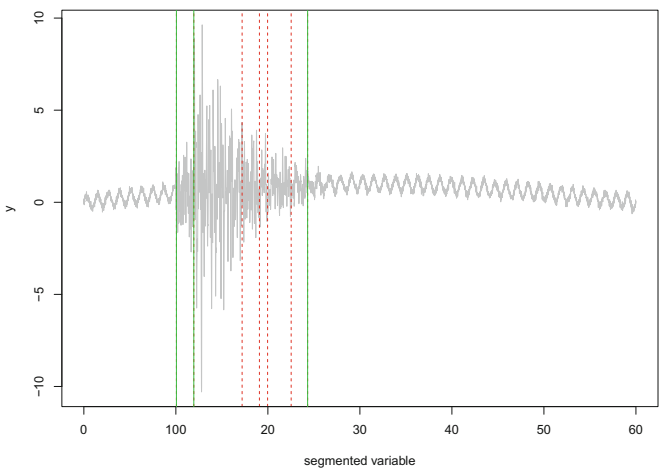


Fig. 5. Results of the algorithm applied to the simulated data with the trend estimated through a smoothing spline with 100 knots. Changepoints estimated by the algorithm are the red dashed lines while the those estimated by the post selection are in green. (Color figure online)

Table 4. Empirical means and Mean Squared Error values of the three relevant changepoints over the 100 waveforms of the simulated dataset, with the added trend, obtained through the application of the algorithm and the post-selection approach.

	ψ_1	ψ_2	ψ_3
Mean	9.90	12.29	21.39
mse	0.03	2.75	6.13

Overall, we may conclude that the algorithm succeeds most of the times in picking the first relevant changepoints, that is the one corresponding to the arrival of the first P-wave. Concerning the second changepoint, that is the one corresponding to the arrival of the first S-wave, this is more easily detected if the mean behaviour of the underlying trend is taken into account through a smoothing spline with the right choice of the number of knots. Indeed, some preliminary analyses, not reported here for brevity, show that a smoothing spline with too many knots tends to overfit the signal, hiding that variation of the signal that could make clearer the detection of the changepoints. Finally, we see that in all the considered setting, the last relevant changepoint, that corresponds to the end of the seismic event, is the most difficult to detect. This may related to the less abrupt change in variation at the end of the signal and, therefore, both the proposed algorithms and the post-selection one can not be accurate in this case.

5 Conclusions and Future Work

The proposed approach is a highly sophisticated algorithm based on changes in variation.

All the analysis carried out in the present work account for simulated data, showing that the proposed algorithm accounts for unknown trends of the signals. Future work will deal with real data, depending on the different characteristics of the analysed region that may affect the observed waveforms. Moreover, taking into account the trend of the signal, by fitting splines, is important to remove all those changes in the waveform that are not attributable to the seismic event.

Moreover, in future work, we will perform comparison between the proposed method and the other already existing algorithms, as the Allen picker, Baer and Kradolfer picker, higher order statistics and AR-AIC picker, to asses their general performance, also in terms of computational efficiency and robustness.

References

1. Adelfio, G.: Change-point detection for variance piecewise constant models. Commun. Stat. Simul. Computat. **41**(4), 437–448 (2012)
2. Akaike, H.: Markovian representation of stochastic processes by canonical variables. SIAM J. Control **13**(1), 162–173 (1975)

3. Akaike, H.: Autoregressive model fitting for control. In: Parzen, E., Tanabe, K., Kitagawa, G. (eds.) Selected Papers of Hirotugu Akaike. SSS, pp. 153–170. Springer, New York (1998). https://doi.org/10.1007/978-1-4612-1694-0_12
4. Aldersons, F.: Toward Three-dimensional Crustal Structure of the Dead Sea Region From Local Earthquake Tomography (2004)
5. Allen, R.: Automatic phase pickers: their present use and future prospects. Bull. Seismol. Soc. Am. **72**(6B), S225–S242 (1982)
6. Allen, R.V.: Automatic earthquake recognition and timing from single traces. Bull. Seismol. Soc. Am. **68**(5), 1521–1532 (1978)
7. Allen, R.M., Gasparini, P., Kamigaichi, O., Bose, M.: The status of earthquake early warning around the world: an introductory overview. Seismol. Res. Lett. **80**(5), 682–693 (2009)
8. Baer, M., Kradolfer, U.: An automatic phase picker for local and teleseismic events. Bull. Seismol. Soc. Am. **77**(4), 1437–1445 (1987)
9. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
10. Hartung, J., Elpelt, B., Klösener, K.H.: Statistik: Lehr-und Handbuch der angewandten Statistik. Walter de Gruyter GmbH & Co KG (2014)
11. Küperkoch, L., Meier, T., Lee, J., Friederich, W., Group, E.W.: Automated determination of p-phase arrival times at regional and local distances using higher order statistics. Geophys. J. Int. **181**(2), 1159–1170 (2010)
12. Küperkoch, L., Meier, T., Diehl, T.: Automated event and phase identification. In: New Manual of Seismological Observatory Practice 2 (NMSOP-2), pp. 1–52 (2012)
13. Morita, Y.: Automatic detection of onset time of seismic waves and its confidence interval using the autoregressive model fitting. Earthquake **37**, 281–293 (1984)
14. Muggeo, V.: Segmented: an R package to fit regression models with broken-line relationships. R NEWS **8**(1), 20–25 (2008)
15. Muggeo, V.M.: Estimating regression models with unknown break-points. Statist. Med. **22**(19), 3055–3071 (2003)
16. Muggeo, V.M., Adelfio, G.: Efficient change point detection for genomic sequences of continuous measurements. Bioinformatics **27**(2), 161–166 (2011)
17. Scherbaum, F., Johnson, J., Rietbrock, A.: Programmable Interactive Toolbox for Seismological Analysis (1999)
18. Schmidt, H., Tango, G.: Efficient global matrix approach to the computation of synthetic seismograms. Geophys. J. Int. **84**(2), 331–359 (1986)
19. Sleeman, R., Van Eck, T.: Robust automatic p-phase picking: an on-line implementation in the analysis of broadband seismogram recordings. Phys. Earth Planetary Interiors **113**(1–4), 265–275 (1999)
20. Smyth, G.K., Huele, A.F., Verbyla, A.P.: Exact and approximate REML for heteroscedastic regression. Stat. Model. **1**(3), 161–175 (2001)
21. Takanami, T., Kitagawa, G.: A new efficient procedure for the estimation of onset times of seismic waves. J. Phys. Earth **36**(6), 267–290 (1988)