

A multi-channel approach for automatic microseismic event localization using RANSAC-based arrival time event clustering (RATEC)

Lijun Zhu^{*1}, Entao Liu^{†1}, and James H. McClellan^{‡1}

¹CeGP at Georgia Institute of Technology

February 17, 2017

Abstract

In the presence of background noise and interference, arrival times picked from a surface microseismic data set usually include a number of false picks which lead to uncertainty in location estimation. To eliminate false picks and improve the accuracy of location estimates, we develop a classification algorithm (RATEC) that clusters picked arrival times into event groups based on random sampling and fitting moveout curves that approximate hyperbolas. Arrival times far from the fitted hyperbolas are classified as false picks and removed from the data set prior to location estimation. Simulations of synthetic data for a 1-D linear array show that RATEC is robust under different noise conditions and generally applicable to various types of media. By generalizing the underlying moveout model, RATEC is extended to the case of a 2-D surface monitoring array. The effectiveness of event location for the 2-D case is demonstrated using a data set collected by a 5200-element dense 2-D array deployed for microearthquake monitoring.

Keywords— passive seismic, sensor array, time picking, classification, multi-channel

1 Introduction

Locations of microseismic events provide important information about conditions in a reservoir during hydraulic fracturing (Duncan 2005). If direct arrivals

*lijunzhu90@gmail.com

†liuentao@gmail.com

‡jim.mcclellan@ece.gatech.edu

can be picked at individual receivers then a geometric calculation such as triangulation can be used to determine the location of an event. This is a common practice in earthquake seismology (Zhang and Thurber 2003). For microseismic events such a method is viable for borehole data (Maxwell *et al.* 2010), but despite its simplicity and low computation cost, it has been argued that microseismic events are not reliably detected on individual traces from surface data (Duncan and Eisner 2010). Alternatively, geophysicists from exploration seismology have developed stacking methods and migration based methods that do not rely on picked arrival times. A travel-time migration based method was implemented by Duncan (2005), which stacks waveforms using calculated travel times, and later expanded by Zhebel and Eisner (2015) to include moment tensor inversion. Along with other semblance-based methods (Tan *et al.* 2014; František *et al.* 2015) these methods discretize a monitoring region and use grid search to find the optimal location that yields the best stacking result under an assumed velocity model. Although attempts have been made to accelerate the exhaustive search for the best location by global optimization such as differential evolution (Zhu, Liu and McClellan 2015) and particle swarm (Luu, Noble and Gesret 2016), these methods are, in general, slow on large monitoring regions with high spatial resolution requirements. Gajewski and Tessmer (2005) used reverse time migration (RTM) to find event locations which was later generalized by Artman, Podladtchikov and Witten (2010) and improved by Nakata and Beroza (2016) through exploring different imaging conditions. Recent development of full-wave inversion (FWI) has inspired full-wave based methods (Witten and Shragge 2016; Sharan *et al.* 2016) however, like the RTM based methods, the finite-difference (FD) modeling they rely on is computationally intensive and can be slower than the grid search program in travel-time based methods. Widespread development of hydraulic fracturing produces increasing amounts of data from passive monitoring which, in turn, demands a more efficient processing scheme for surface arrays that can be deployed without drilling monitoring wells. We will show that the issue of low SNR in surface-array recordings can be overcome so that arrival-time based methods become attractive because of their low computation cost.

A recent study by Akram and Eaton (2016) summarized and compared the most common arrival time picking methods on a single trace, such as short-term over long-term average ratio (STA/LTA) (Allen 1978; Earle and Shearer 1994), a modified energy ratio (MER) (Han, Wong and Bancroft 2009), a modified form of Coppens' method (MCM) (Sabbione and Velis 2010), and Akaike's information criterion (AIC) (Takanami and Kitagawa 1991), to name a few. A common theme in all of these methods is the use of processing to increase the significance of detected peaks and minimize the number of false picks from noisy data which leads to bad event location estimation. Ideally, when all the picks are perfect a moveout curve can be computed to fit exactly through the arrival-time picks. Alternatively, Zhu, Liu and McClellan (2016) took advantage of the fact that a group of receivers with both good and bad picks still *contains a subset of picks that follow an expected trend* of arrival times when events are present. By fitting a moveout curve through subsets of picked arrival times using ran-

dom sample consensus (RANSAC) (Fischler and Bolles 1981), Zhu *et al.* (2016) were able to recover the true arrivals in the presence of a large amount of false picks under low SNR conditions. In this paper, we continue the study of applying RANSAC for picked arrival times and improve our method described in (Zhu *et al.* 2016) for realistic seismic monitoring scenarios. To reduce the false picks in time picking results and thus improve the event location estimation, we propose a RANSAC-based arrival time clustering (RATEC) method as a pre-processing step that groups true picked times into different events and identifies false picks. To demonstrate the accuracy of RATEC for 1-D receiver arrays, synthetic simulations are conducted in homogeneous, layered and non-layered isotropic media. The proposed scheme is also validated through a natural earthquake dataset collected on a 5200-element 2-D surface network in Long Beach, CA. All cases show that the RATEC framework is accurate and robust under low SNR conditions and applicable to a variety of different monitoring setups.

2 Motivation

To eliminate false picks generated by a time picking algorithm due to background noise, a classifier for true event picks is necessary. Such a classifier needs to learn the pattern of a seismic event from all arrival-time picks and apply a rule to cluster the picks into two groups: true event and false picks. It also needs to be robust enough to accommodate the variety of patterns shown by different events. Since the true first-arrival times of any isolated seismic event result in a predictable moveout curve on a monitoring receiver array, a parametric model for valid moveouts can be used to build a classifier for true picks of an actual seismic event.

Moveout curves have been studied extensively in seismology by Dix (1955) and Dellinger, Muir and Karrenbach (1993) (See Appendix A). For homogeneous media, it is simply a hyperbola. Dix (1955) proved that moveout curves can also be modeled as hyperbolas for isotropic layered media when the receivers have small offsets relative to an event epicenter. He also gave explicit parametric equations for such curves (as a rotated hyperbola) when a tilting layer is present. Dellinger *et al.* (1993) showed that for TI (transverse isotropic) media, an elliptic parametric model can be used to approximate the expected moveout curves. Since horizontal variation in velocity is relatively small in microseismic monitoring, a hyperbola can be used to approximate the arrival time moveout curve for event in non-layered media as well. To sum up, for a surface monitoring receiver array in microseismic monitoring, a quadratic parametric model exists for a moveout curve observed on receiver array from a valid seismic event. Thus, the problem of finding the true picks for a seismic event can be solved by fitting a parametric model using picked arrival times. For simplicity, we only consider isotropic media with short offsets in this study which corresponds to a hyperbola. Ellipses share the same quadratic model as hyperbolas, but with different parameter requirements.

However, due to poor SNR on surface monitoring arrays, there are many

false picks that are far from true event moveout curves which we refer to as outliers. Least squares curve fitting uses as many data points as possible to minimize the amount of misfit error. Although it is the optimal solution under a Gaussian random noise assumption, it fails dramatically in the presence of outliers. We, instead, adopt a random sampling scheme that repeatedly uses a minimum number of data points to fit a curve, and selects the best curve (hypothesized model) as the one with the most data points close to it, i.e., with the maximum number of inliers.

This random sampling scheme was first proposed by Fischler and Bolles (1981) as RANSAC and then improved by many others (Stewart 1995; Torr and Zisserman 2000; Chum and Matas 2002; Tordoff and Murray 2005; Chum and Matas 2005). It has also been extended to fit multiple models simultaneously (Wang and Suter 2004; Toldo and Fusiello 2008). Based on the fitted hyperbolic model, the picked arrival times are, in fact, clustered into event groups and non-event groups. Such clustering not only separates picked arrival times into different phases (e.g., P-wave and S-wave phases), but also improves the accuracy of localization results by eliminating false picks due to noise.

3 Method

3.1 RANSAC overview

Despite many variations and adaptations of RANSAC (Choi, Kim and Yu 2009), there are essentially two steps per iteration (hypothesize-and-test) which will be repeated to yield the best fit to the data:

- *Hypothesize*: A **minimal** sample subset (MinSet, denoted as Ω_M^k) is randomly selected from the dataset and the **unique** model parameters (\mathbf{p}^k) are computed for Ω_M^k .
- *Test*: Elements in the dataset (Ω_D) are evaluated to determine which ones can be labeled as *inliers*, i.e., consistent with the hypothesized model in the sense that the distance from the model’s moveout curve is less than some prescribed value (δ). The set of all such inliers is called a consensus set (ConSet, denoted as Ω_C^k).

Note that $\Omega_M^k \subset \Omega_C^k \subset \Omega_D$. A set Ω_M^k consists of only the minimal number of samples required to uniquely determine a model, e.g., two samples for a line and three for a circle. The more elements in Ω_C^k , the better the model we have obtained for the k^{th} hypothesis.

Notice that each RANSAC iteration requires very little computation and there exists a unique solution for each chosen MinSet. In this way, we can afford using a large number of iterations consistently to perfect a hypothesized model. Fischler and Bolles (1981) give a statistical analysis of the required iteration number of RANSAC process with inlier ratio of u (ratio between number of inliers and total number of data points). The number of iterations \hat{N} to

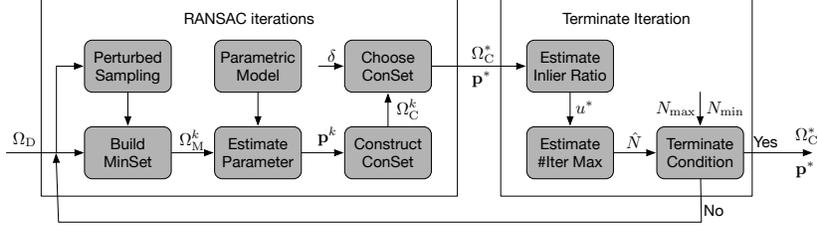


Figure 1: Flow chart of RANSAC process within each iteration. After many iterations the case with the biggest ConSet is declared the best parameter vector \mathbf{p}^* .

guarantee that at least one Ω_M will only contain true picks with probability p is

$$\hat{N} = \frac{\log(1-p)}{\log(1-u^m)}, \quad (1)$$

where m is the size of a MinSet, which is 5 for a hyperbola and 9 for a hyperboloid. For example, when 50% of all picks are close to a hyperbola ($u = 0.5, m = 5$), we can guarantee a 99% chance of finding an outlier-free Ω_M ($p = 0.99$), if we run $\hat{N} = 145$ iterations. For a 2-D surface array, $m = 9$, and then $\hat{N} = 2356$. Although \hat{N} may be large, the core operations of deriving the hyperbola parameters and testing its validity are extremely fast so thousands of trials are reasonable.

The RANSAC process is summarized as the flow chart in Figure 1. After the k^{th} iteration, the current best ConSet Ω_C^* is updated with the k^{th} ConSet if Ω_C^k has more inliers. Then Ω_C^* is used to estimate the current best inlier ratio u^* . Based on equation (1), the number of iterations required \hat{N} can be updated (Tordoff and Murray 2005). The current \hat{N} is also compared against preset minimum and maximum values N_{\max} and N_{\min} . Once the termination condition is satisfied, the best ConSet Ω_C^* and model parameter \mathbf{p}^* will be returned; otherwise, the iteration loop will continue.

3.2 Parameter estimation

The proposed method uses a quadratic model to estimate the parameters of a hyperbolic curve, which takes the following form:

$$\mathcal{P}(x, y; \mathbf{p}) = [x \quad y] \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + [d \quad e] \begin{bmatrix} x \\ y \end{bmatrix} + f = 0 \quad (2)$$

or, equivalently,

$$\mathcal{P}(x, y; \mathbf{p}) = ax^2 + bxy + cy^2 + dx + ey + f = 0, \quad (3)$$

where \mathbf{p} has six real elements $\boldsymbol{\theta} = (a, \dots, f)$, but there are actually only five free parameters, since one of the nonzero elements can be always normalized to 1. When the determinant

$$\Delta_1 = 4 \begin{vmatrix} a & b/2 & d/2 \\ b/2 & c & e/2 \\ d/2 & e/2 & f \end{vmatrix} \quad (4)$$

is nonzero ($\Delta_1 \neq 0$), equation (3) defines a non-degenerate conic section. To verify it is hyperbola, we must then check a second determinant

$$\Delta_2 = 4 \begin{vmatrix} a & b/2 \\ b/2 & c \end{vmatrix} = b^2 - 4ac. \quad (5)$$

When $\Delta_2 > 0$, equation (3) defines a hyperbola.

Given a set of n arrival time picks, (x_i, y_i) for $i = 1, \dots, n$, we form a $n \times 6$ data matrix \mathbf{D}_n and a 6×1 coefficient vector \mathbf{p} , such that $\mathbf{D}_n \mathbf{p}$ is the model $\mathcal{P}(x, y; \mathbf{p})$ evaluated at the time picks. With measurement error, there will be a nonzero residual \mathbf{r} as follows:

$$\begin{bmatrix} x_1^2 & x_1 y_1 & y_1^2 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^2 & x_n y_n & y_n^2 & x_n & y_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \Leftrightarrow \mathbf{D}_n \mathbf{p} = \mathbf{r}. \quad (6)$$

We use only five picks to uniquely determine \mathbf{p} as discussed previously. If all picks in Ω_M are true picks, the residual term \mathbf{r} is usually negligible. Then we solve the linear system $\mathbf{D}_5 \mathbf{p} = \mathbf{0}$, which is effectively finding the null space of \mathbf{D}_5 . From the singular value decomposition (SVD) of \mathbf{D}_5 , it is easy to see that the last right singular vector $\mathbf{v}_6 \in \text{null}(\mathbf{D}_5)$. Comparing with the pseudo-inverse method used in (Zhu *et al.* 2016), the solution \mathbf{v}_6 adopted here is not guaranteed to have the minimal L_2 norm. However, we avoid the numerical stability problems when the matrix $\mathbf{D}_n \mathbf{D}_n^T$ is ill-conditioned. To process the large number of candidate MinSets Ω_M efficiently, we do a quality control (QC) of \mathbf{p} by checking determinants, $\Delta_1 \neq 0$ and $\Delta_2 > 0$, before proceeding to the more computationally demanding test step that computes the distance between Ω_D and the hypothesized RANSAC model to obtain a ConSet Ω_c .

3.3 Add perturbation to MinSets

With the presence of measurement noise, the null space method has a tendency to fit the wrong type of curve, namely a parabola or an ellipse, which is then eliminated by the parameter QC step. In noisy cases when the percentage of inliers is low, this problem may cause RANSAC to select a suboptimal parameter vector whose curve passes through some outliers as well as true inliers, as shown in Figure 2.

To overcome this tendency when fitting quadratic models, a constrained least squares approach (Fitzgibbon, Pilu and Fisher 1999; O’Leary and Zsombor-Murray 2004) has been developed to force the fitted curves to be hyperbolas (and ellipses) by solving a generalized eigen system determined by a constraint matrix. Although this method works for general hyperbolic curve fitting problems, its strategy runs against RANSAC’s MinSet assumption, i.e., using a random set with the minimum number of points. Least squares employs as many data points as possible in order to minimize the distance between the data and the optimal model. However, the more points required in the MinSet,

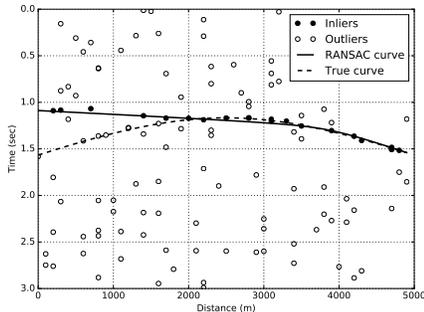


Figure 2:

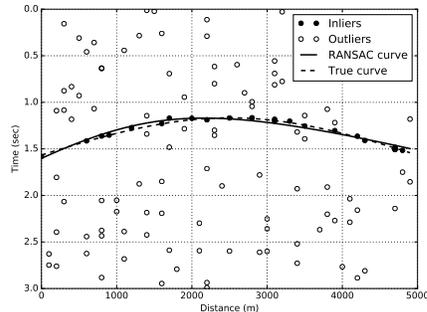


Figure 3:

Figure 4: RANSAC classification of noisy data: (a) suboptimal fitting results from direct null space method, (b) optimal fitting results from null space method with perturbations.

the smaller the possibility that a MinSet will be outlier-free. This dilemma restricts the ability of the constrained least-squares method to find a proper model when the SNR is low.

Fortunately, the RANSAC process has the luxury of running a very large number of fast iterations to find the optimal model. By randomly perturbing the time picks in Ω_M^k to produce a few additional MinSets and running more iterations, we can combat the effect of measurement noise. For example, after adding a small perturbation to the picked arrival time in Figure 4a, the correct moveout curve are selected in Figure 4b. Although we perturbed the picked arrival times to get the model coefficients, the final output of inliers and the location estimation are conducted on the original “unperturbed” data.

3.4 Processing pipeline

The overall processing pipeline of RATEC is summarized in Figure 5. Arrival times are picked on pre-processed data to extract event features out of seismic traces as time pick pairs (\mathbf{x}, t) . In this study, we use the widely adopted short-term over long-term average ratio (STA/LTA) method to generate a characteristic function for each seismic trace. However, any valid time picking method, such as those included in (Akram and Eaton 2016), can replace STA/LTA depending on the specific SNR condition. Peak detection is conducted on characteristic functions to generate (\mathbf{x}, t) pairs which are then clustered by RANSAC to select true picks that correspond to a valid event moveout. These clustered picks can be fed into other location estimation programs such as double difference (Waldhauser and Ellsworth 2000; Zhang and Thurber 2003). When no prior knowledge of the velocity model is available, we provide a moveout curve fitting based event location estimator assuming a homogeneous medium.

Notice that this is a highly flexible framework in which multiple methods can be used for each block to optimize the performance for different datasets. Figure 5 provides a generic approach to demonstrate the accuracy and robustness of

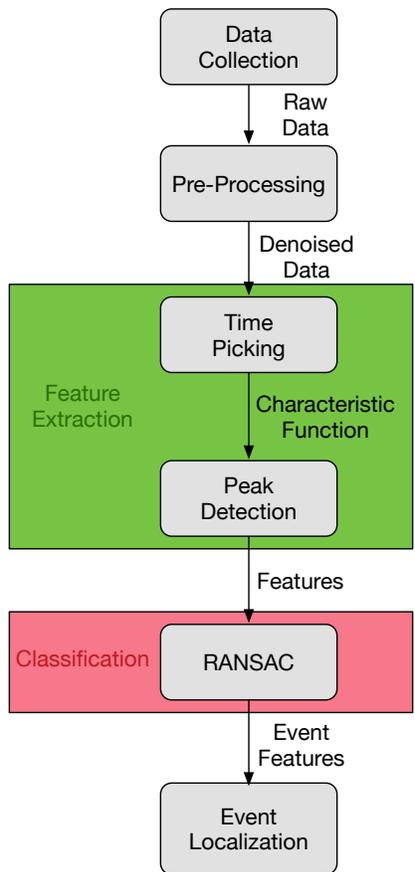


Figure 5: Processing pipeline of proposed RATEEC method: peak detection can be customized to adapt to the RANSAC framework; classification based on RANSAC eliminates false picks when estimating the moveout curve for an event.

RATEEC; however, it can be customized to specific needs and easily incorporated into any time picking based processing workflow.

4 Pre-process seismic data for RANSAC classification

The input data can be pre-processed to facilitate peak detection and help RANSAC better fit the moveout curve. The strategy is to encourage more time peaks by including as many weak events as possible while not introducing too many false picks. Here we give an example pre-processing method that takes advantage of RANSAC’s ability to eliminate outliers while including more weak picks that might be related to a true event.

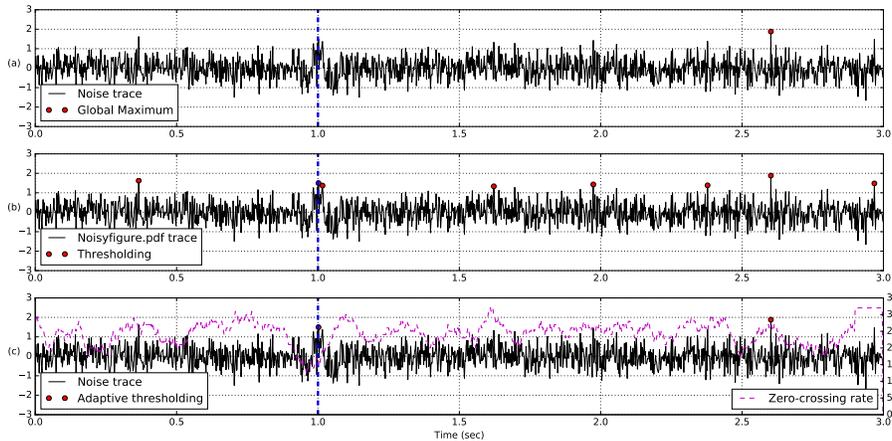


Figure 6: Detection methods for single event case on noise trace with PSNR = 6 dB (true peak at $t = 1.0$ sec): (a) global maximum; (b) thresholding at 70% of global maximum; and (c) adaptive thresholding at 95% of global maximum weighted by the local zero-crossing rate. Adaptive thresholding picks true arrival time cluster with only one false pick.

4.1 Guided peak detector

A straightforward approach to peak detection in characteristic functions is to find the global maximum on each trace. However, such peak locations can easily be affected by background noise as shown in Figure 6a, and smaller events are overlooked when multiple events are present. Likewise, locating peaks by local maxima is adversely affected by background noise since a noise signal tends to have a large number of local peaks.

One way to include more time picks is to use thresholding at a fraction of the maximum value. Shown in Figure 6b, setting the threshold to 70% of the maximum value yields more picks. This method succeeds in finding more arrival times (both true and false), but puts a heavy burden on the following classification block if too many of these peaks are false ones. A better way to include more time picks is to make a rough estimate of where the real signal lies based on a signal attribute such as the local zero-crossing rate defined below:

$$r_{zc}(\tau) = \frac{1}{T} \sum_{t=\tau-T/2}^{\tau+T/2} \mathbf{1}(s_t s_{t-1} < 0), \quad (7)$$

where $\mathbf{1}(s_t s_{t-1} < 0)$ counts sign changes in $\mathbf{s}_t = \text{sgn}(s(t))$, and T is the interval over which $r_{zc}(\tau)$ is computed. The local zero-crossing rate should be low when the signal is present. Using 95% of the global maximum threshold on the characteristic function weighted by $\sqrt{r_{zc}(\tau)}$, Figure 6c shows that this guided peak detector successfully picks only the real arrival time peak and the global maximum (due to severe background noise).

4.2 Merging close picks

With background random noise, there may be multiple peaks clustered around a true event pick. This not only leads to more computational cost in later localization algorithms but also introduces uncertainty into event locations. Such a problem can be solved by merging close picks into one pick. A common practice in manual picking is to use the starting point of a pick cluster as the event arrival time. This is reasonable as the peak detector usually picks both the arrival signal (first break) and its coda wave (points that follow which form a cluster of time picks). However, it not only requires more computation to search for closely located peaks but also can be misled by a false pick that slightly leads the true picks. It can soon become tricky to set the correct parameter for how close the picks need to be to each other for a merge.

We consider using Gaussian smoothing which is widely used for edge detection in image processing (Basu 2002). Gaussian smoothing helps in reducing details (adjacent small peaks) within the characteristic function and attenuating insignificant local peaks due to noise. It convolves the response function with a Gaussian function defined below:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}, \quad (8)$$

where σ is the standard deviation which can be set as the dominant duration of a wavelet (0.1 sec in this case). Although Gaussian smoothing cannot eliminate all the close false picks, it can help mitigate such errors. Moreover, the RANSAC process is less sensitive to inlier distance selection after Gaussian smoothing.

5 Seismic Examples

In this section, we explore the ability of the proposed RATEC method in a more realistic scenario of seismic processing. In the first example, a Ricker wavelet is manually delayed with moveout from a homogeneous medium assumption to demonstrate the essence of the proposed method. The second example uses a recorded seismic trace consisting of P-wave and S-wave phases which are then manually delayed according to a layered velocity model to simulate data from an array. This is a typical scenario in microseismic surface monitoring and we demonstrate that RATEC is able to extract both P-wave and S-wave phases and group them into event clusters. In the third example, we explore the problem when the layered model assumption is violated by using the Marmousi2 velocity model to generate the testing data with a finite-difference time-domain (FDTD) simulation. In the final example, we demonstrate that RATEC can be easily extended to the case of a 2-D surface monitoring array by validating it on a 5200-element 2-D dense array deployed for earthquake monitoring in Long Beach, CA.

5.1 Ricker wavelet in homogeneous media

For Figure 7a, a 25-element linear array with nominal spacing of 200 m is deployed on the surface (i.e., 5 km aperture) to monitor a deep event 2 km below the array center. The receiver locations are perturbed by additive white Gaussian noise (AWGN) with $\sigma = 50$ m to simulate receiver offsets in the field away from uniformly spaced locations due to unavoidable physical restrictions in the field. Note that such perturbations effectively create a nonuniform linear array. The raw data section is shown in Figure 7a with the true moveout curve marked with a blue dashed line.

The source wavelet employed here is a Ricker wavelet, and the medium is assumed to be homogeneous with a velocity of 3 km/s. AWGN with peak signal-to-noise ratio (PSNR) of 6 dB is added to simulate random background noise. Because PSNR is not affected by the trace length, it is used to measure the noise level throughout the paper. Its definition is as follows:

$$\text{PSNR} = 20 \log_{10} \frac{\max(|s_i(t)|)}{\sigma}, \quad (9)$$

where $s_i(t)$ is the signal at the i^{th} receiver, and σ is the standard deviation of the AWGN.

After applying the STA/LTA method on each trace, we use the zero-crossing guided peak detector and peak merging to find the candidate arrival time picks. These picks are passed to a classification block to be grouped into event (inlier) and non-event (outlier) clusters. When zoomed in around the moveout curve as shown in Figure 7b, a small deviation between the fitted curve and the true moveout curve is observed; however, all picks close to the true moveout curve are successfully clustered into the event/inlier group.

Once clustering and correction are complete in the previous steps, the improved picked arrival times in this example can be used to locate events. There are many existing event localization methods that use picked arrival times, such as Geiger’s method (Geiger 1912) and the double-difference method (Waldhauser and Ellsworth 2000; Zhang and Thurber 2003). The corrected arrival times can be used by these methods with known velocity models to improve the location estimation.

When the velocity model is unknown, we can assume a homogeneous medium in order to compute predicted arrival times from possible source locations. Based on the inliers given by RATEC, we can minimize a nonlinear objective function that measures the sum of squared errors between the RATEC classified inlier picks and the predicted arrival times

$$\epsilon = \sum_{i=1}^n (t_i - t_i^p(\mathbf{x}, T_0, v))^2 \quad (10)$$

where t_i is the RATEC pick at the i^{th} receiver and t_i^p is the predicted arrival time which is a hyperbola that depends on the event location \mathbf{x} , origin time T_0 , and homogeneous velocity v . The minimizer of equation (10) gives the event location and event origin time, and the medium velocity simultaneously.

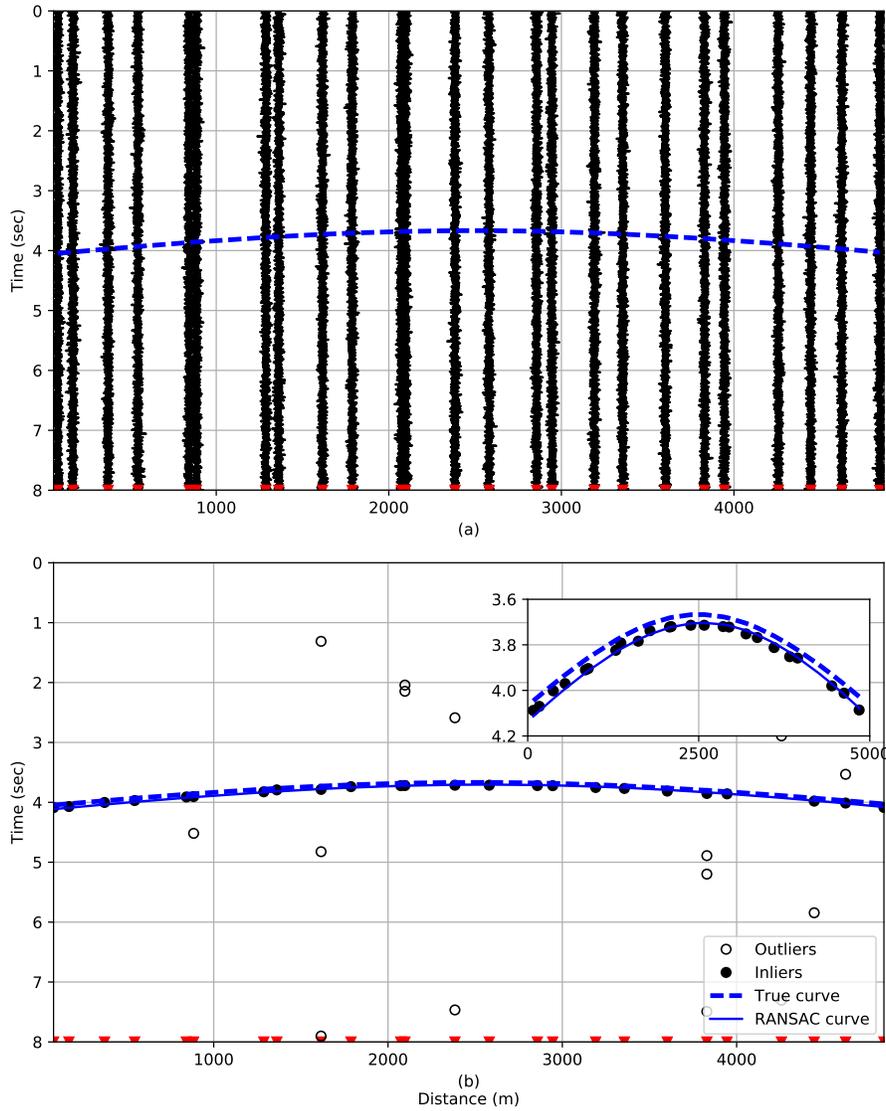
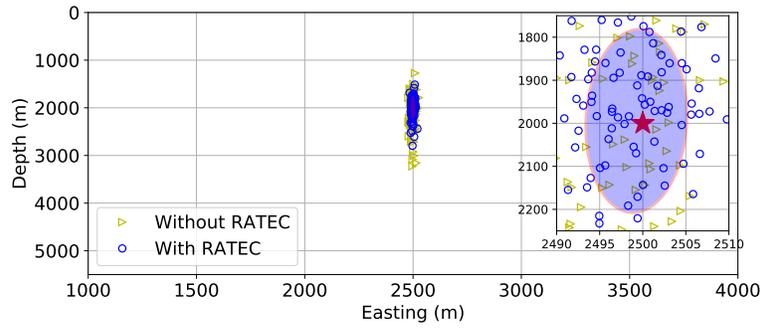
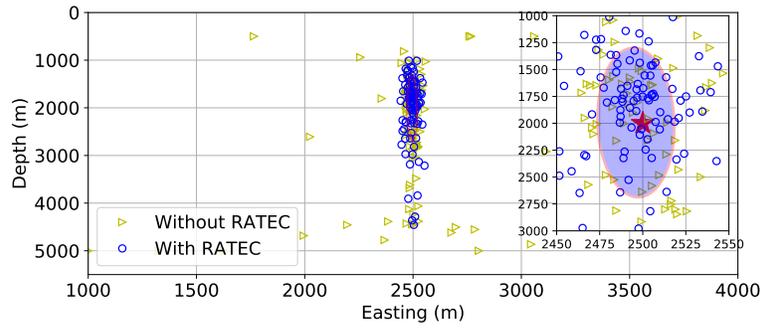


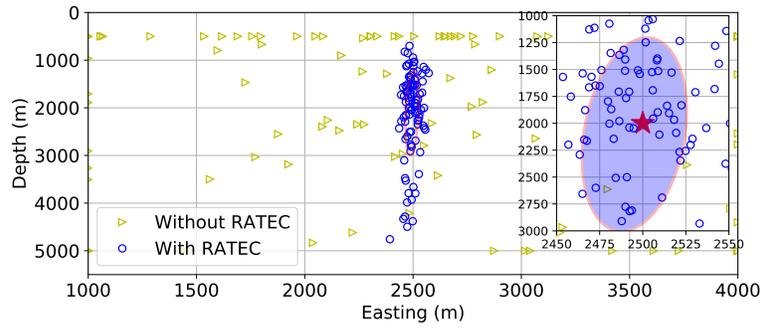
Figure 7: Simple example of time picking: (a) raw traces of a 25-element nonuniform linear array with an aperture of 5 km. Source waveform is a Ricker wavelet of 10 Hz central frequency and the PSNR = 6 dB. Blue dashed curve indicates the true moveout. (b) Arrival times picked on each trace clustered into inlier (solid dots) and outlier (empty dots) groups. Insert in the top right shows a zoomed-in view along the vertical (time) axis.



(a)



(b)



(c)

Figure 8: Location results with and without the RATEC scheme using 100 Monte Carlo experiments under three different background PSNR noise levels: (a) 20 dB, (b) 8 dB, and (c) 6 dB. Blue ellipses (in the insets) show the one-sigma confidence interval of the event location estimator. Note the changing axis limits for the insets. The red star indicates the minimizer for the noiseless case which overlaps the true event location at (2500, 2000) m.

To validate the accuracy of this localization scheme, 1000 Monte Carlo experiments are conducted to compare the results with and without the RATEC scheme under different noise levels. Only the first 100 data points, which is sufficient to capture the location estimation distribution as a point cloud, are shown in Figure 8 to avoid crowding. For low background noise, PSNR = 20 dB, both methods obtain the event location accurately around the true event location (2500, 2000) m as shown in Figure 8a. The uncertainty in depth is mostly a result of the array geometry which has poor resolution in the direction perpendicular to the linear array. The location results with RATEC have a more compact distribution around the true location. There are more blue dots than yellow triangles within the one-sigma confidence interval indicated by the blue ellipse. The location estimates without RATEC start to fall apart when the PSNR is around 8 dB as shown in Figure 8b. Although most of the yellow triangles are still around the true location region with a larger spread, there are a significant number of location estimates far away from the event region. On the other hand, the results with RATEC show a consistent distribution around the true event region in Figure 8b. Under severe noise as shown previously in Figure 7 with 6 dB PSNR, the location results without RATEC become completely unreliable while those with RATEC still give very good estimates. In Figure 8c about 50% of the blue dots, but only one yellow triangle, lie within the confidence ellipse.

The accuracy of the location estimate with and without RATEC measured in root-mean-square error (RMSE) for the complete 1000 Monte Carlo experiments are shown in Table 1. With the RATEC correction, the location estimate in easting is improved significantly. The error in depth is much larger but is reduced by applying the RATEC correction.

Table 1: RMSE of RATEC localization results from 1000 Monte Carlo experiments.

RMSE	Easting			Depth		
	20 dB	8 dB	6 dB	20 dB	8 dB	6 dB
Without RATEC	8.88	312.21	911.79	386.96	1273.19	1914.61
With RATEC	6.09	25.49	65.73	199.43	869.78	1025.52

5.2 Recorded seismic trace in layered model

Here, the seismic trace shown in Figure 9a is used as a source signal. After manually picking the P-wave and S-wave phases, shown in Figure 9b and 9c respectively, the P and S phases are delayed separately according to their travel time (T) computed from a layered model against horizontal offset (x) using the parametric equation (11) given by Dix (1955). A detailed explanation can be

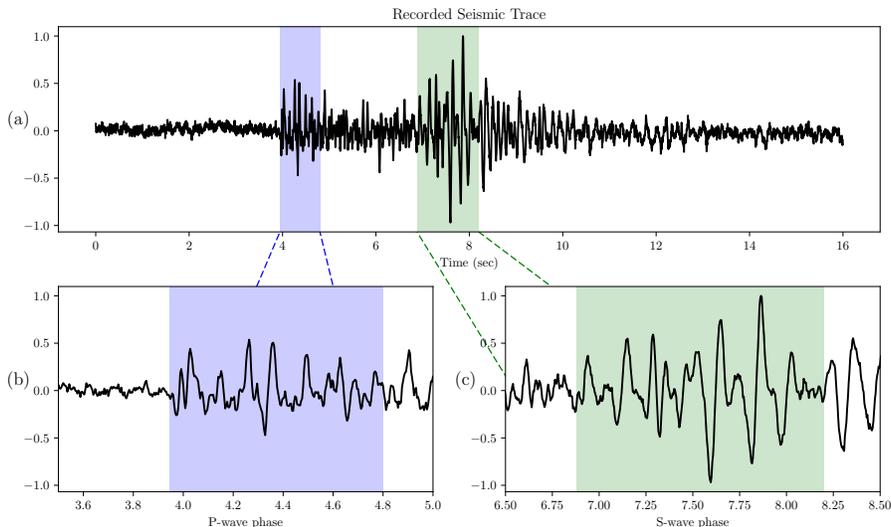


Figure 9: (a) Recorded seismic trace with P-wave (blue) and S-wave (green) phases for layered model simulation, (b) manual pick of P-wave phase, and (c) S-wave phase.

found in Appendix A.

$$\begin{cases} x = \sum_k \frac{ph_k v_k}{\sqrt{1 - (v_k p)^2}} \\ T = \sum_k \frac{h_k}{v_k \sqrt{1 - (v_k p)^2}} \end{cases} \quad (11)$$

where p is the ray parameter that is constant among all layers, h_k and v_k are layer thickness and layer velocity which defines a layered velocity model. Unlike the hyperbolic approximation discussed before, this equation is mathematically valid even when $x \rightarrow \infty$; however, the direct wave may not necessarily be the first arrival wave when x is large. In addition, for large x cases, the SNR condition may be too bad for a valid localization problem. Thus, all examples here are conducted for small x (less than 5 km).

The layered velocity model used in this example as shown in Figure 10a is taken from Marmousi2 elastic velocity model (Martin, Wiley and Marfurt 2006). The top water layer in Marmousi2 is removed and event source is located around 2.5 km deep. The same nonuniform surface array as in Section 5.1 is used for monitoring underground seismic events occurring at the center of the array.

With noise at 10 dB PSNR, the P-wave and S-wave arrivals are not obvious in the raw data shown in Figure 10b. With a spectrogram the dominant frequency of the arrival event is estimated to be 10 Hz, so a low-pass filter with cutoff frequency at 20 Hz is used as pre-processing. Both P-wave and S-wave arrivals are observed in Figure 10c after low-pass filtering. The result of applying the RATEC method is shown in Figure 10d, where moveout curves were generated

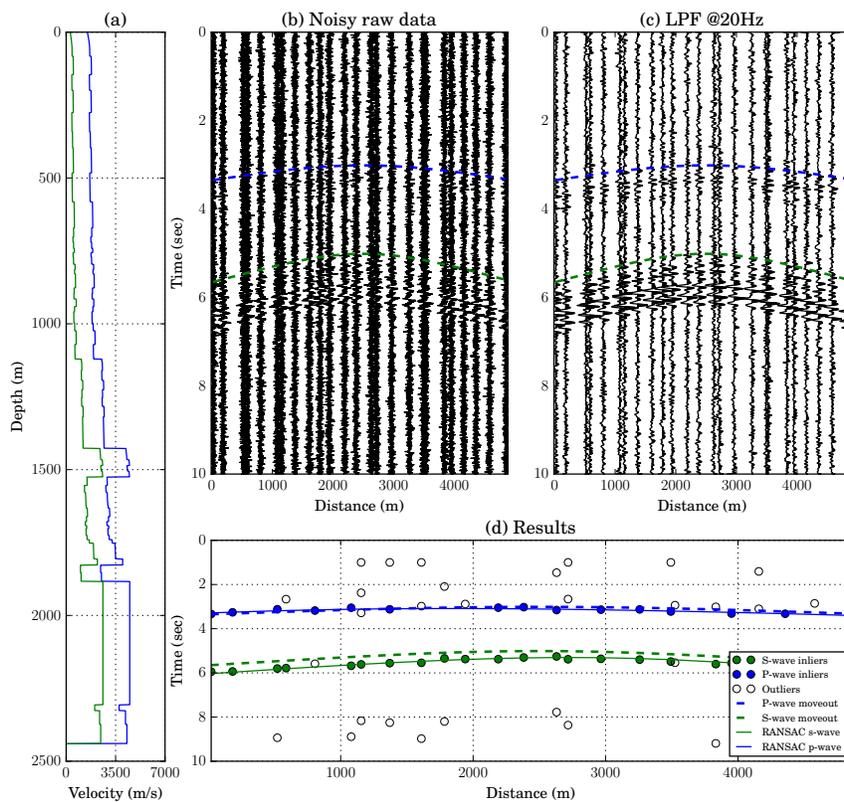


Figure 10: Layered velocity model example using recorded seismic trace with P-wave and S-wave phases: (a) 1-D velocity model from Marmousi2; (b) noisy raw data with PSNR = 10 dB with respect to the S-wave peak; (c) 20 Hz low-passed data; (d) fitted moveout curve and classification results comparing to true P-wave and S-wave moveout curves.

by fitting the classified and corrected arrival time picks. The proposed method is used iteratively in this example to extract all possible event phases: after one moveout curve is detected and identified, its outliers are used as the input for the next iteration to search for more curves until there are not enough time picks to successfully define a moveout curve. Here, both P-wave (blue) and S-wave (green) phases are identified in this example with most of the true arrival times labeled correctly.

5.3 Ricker wavelet in non-layered media example

Although RATEC is based on a layered velocity model assumption, it is robust enough to handle non-layered model to some extent. In Figure 11a, the acoustic Marmousi model is used to introduce horizontal variation in the velocity model. A finite-difference time-domain based numerical simulation is used to generate the receiver data shown in Figure 11b with 10 dB PSNR of AWGN. Since each trace has different peak value, which is common in a real seismic scenario, the PSNR defined here uses the global peak of all the traces. Receivers in the layered region (0 ~ 1000 m) tend to have better SNR than those in the non-layered region (1000 ~ 1500 m).

After applying the RATEC scheme, Figure 11c shows the results of curve prediction and arrival time labels. Even though the true moveout is not exactly a hyperbola, the RATEC method is able to label all the true arrival times given a reasonable tolerance distance. Zoomed in around the layered region, good prediction and perfect labeling are observed in Figure 11d. Notice that there now exists larger offsets between picked and true arrival times. Figure 11e shows the results in the non-layered region where the SNR is worse. Despite the fact that many picks in that region are false picks, RANSAC is able to eliminate most of the picks far away from the true moveout curve and label the true time picks correctly.

5.4 Surface extension on Microearthquake data

RATEC can be easily extended to a surface array by changing the underlying hyperbolic curve model to a hyperboloid surface model. Similar to equation (2), a hyperboloid surface can be defined using a 3-D quadratic equation which takes the following general form:

$$\mathcal{P}(x, y, z; \mathbf{p}) = [x \ y \ z] \begin{bmatrix} a & b/2 & d/2 \\ b/2 & c & e/2 \\ d/2 & e/2 & f \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + [g \ h \ i] \begin{bmatrix} x \\ y \\ z \end{bmatrix} + j = 0. \quad (12)$$

Using the same RATEC scheme, we can adapt the framework to hyperboloid surface fitting by finding the parameter vector $\mathbf{p} = (a, b, \dots, j)$ in a 10-dimensional space. Although this may seem to be a much larger parameter space, it adds little burden on search process as RANSAC searches only Ω_M rather than complete parameter space.

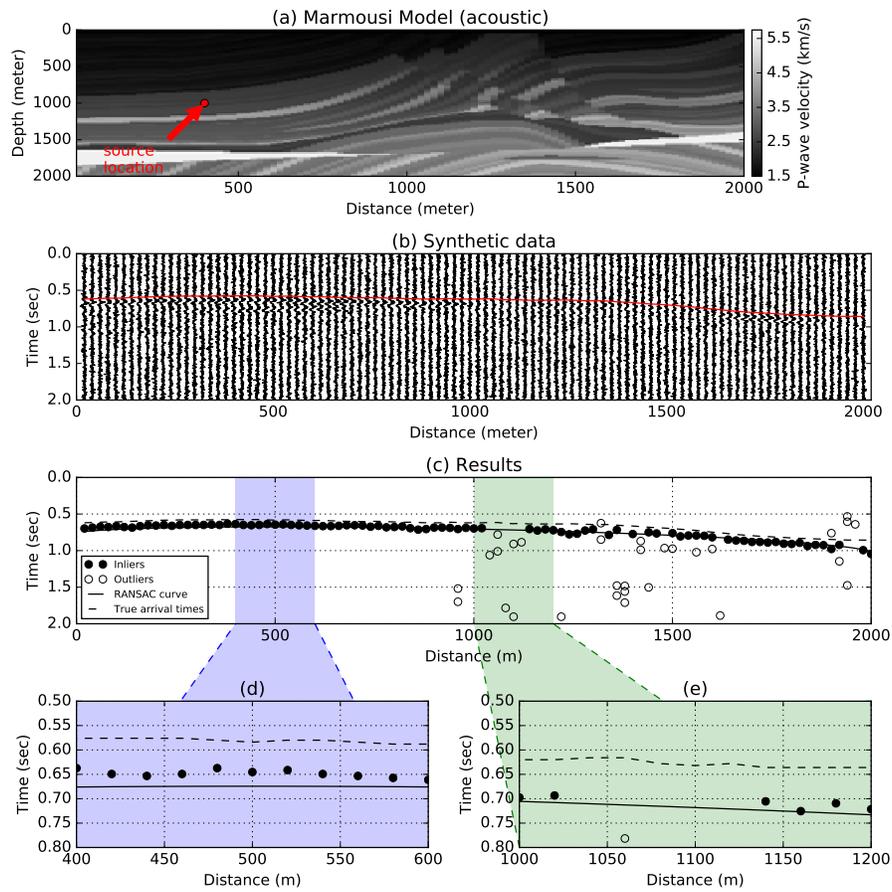


Figure 11: Marmousi model example under 10 dB PSNR: (a) velocity model (red dot indicates source location), (b) synthetic data (red line indicates true arrival times), (c) RATEC results, (d) zoomed-in results between 400 m and 600 m and (e) zoomed-in results between 1000 m and 1200 m.

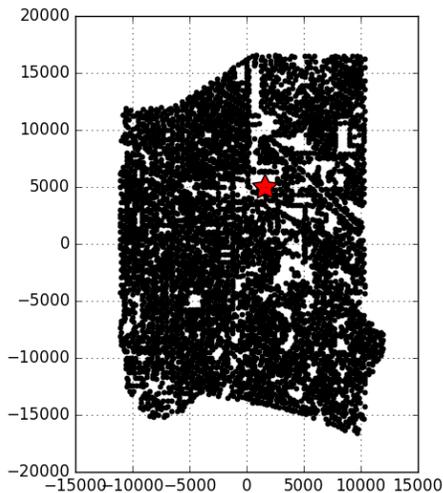


Figure 12: Top view of the sensor array with located event indicated by red star.

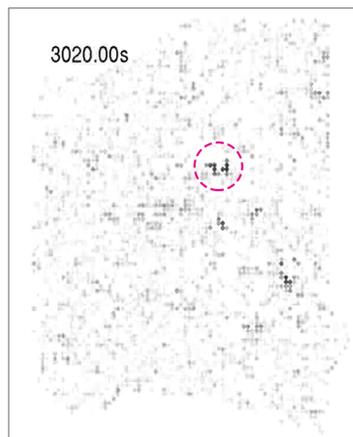


Figure 13: Snapshot of the seismic dataset at time $t = 3020.00$ s; the visible event lies inside the red circle.

The proposed method was then tested on a data set of 50 sec collected by the Long Beach nodal array in southern California which contains 5200 sensors. The top view of the sensor array is shown in Figure 12. Prior to applying RATEC scheme, no reliable location estimation can be given by picked arrival times due to a large number of false picks as shown in Figure 14. We can use the picked arrival time to locate event using a homogeneous medium assumption since there is no velocity model known prior to this experiment. Based on the true picks given by RATEC, this seismic event is recognized as a surface event whose location is shown by its epicenter marked by the red star in Figure 12. In order to verify our result, we schematically show the corresponding snapshot on the sensor array in Figure 13. The gray-scale of the dots indicates the clipped signal amplitude on the corresponding sensor. The red circle in Figure 13 confirms that in the inverted time and location using the classified true picks, there is indeed a weak event that is barely visible in the array. Moreover, the work log shows that there is a surface source in the estimated area but the local earthquake catalog has no record of earthquakes during the event time. In Figure 14 we show the time picking results that contain a large number of false picks. The best-fitted hyperboloid surface from 3-D RATEC is shown as the red surface. On a laptop, RATEC takes just 31 sec to finish the classification process, which is sufficient for real-time processing (note that the recording duration is 50 sec).

5.5 Parameter selection

Although it may seem that there are many parameters to be set for the RATEC method, they are actually tied to just one parameter that can be estimated from the data itself: f_{dom} , the dominant frequency of the source wavelet.

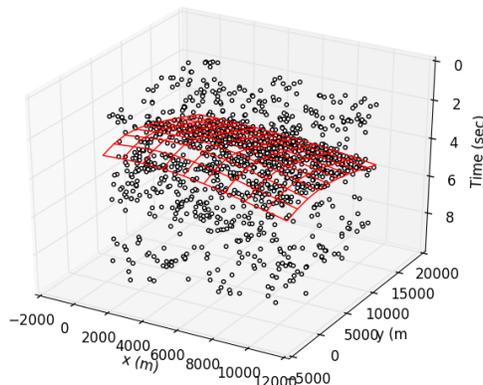


Figure 14: Top view of the picks (\circ) from the 2-D sensor array with fitted surface in red.

Table 2: Parameters used in all simulations.

Parameter	Recommended selection
Dominant frequency (f_{dom})	10 Hz
Dominant period (T_{dom})	$1/f_{\text{dom}}$ sec
Cut-off frequency in LPF	$2f_{\text{dom}}$
Short-term window (STW)	$0.5T_{\text{dom}}$
Long-term window (LTW)	$5T_{\text{dom}}$
Zero-crossing rate window length (T)	LTW
Gaussian smooth function sigma	STW
Threshold Distant (ThresDist)	$0.5T_{\text{dom}}$
RANSAC additive noise sigma	ThresDist / 2

All parameters used in the above simulation examples are summarized in Table 2 and their recommended relationship to f_{dom} is listed as well. These parameters work well with Ricker wavelet based simulations since the dominant frequency is a valid measurement of wavelet length (dominant period T_{dom}). However, sometimes the true wavelet length is longer than the dominant period (T_{dom}), e.g., a seismic trace with dispersion. In this case, we fix the cut-off frequency at 20 Hz but make T_{dom} longer to alleviate the problem.

6 Conclusion

In this paper, we tackle the problem of event location estimation from arrival times by fitting a parametric model and proposed an RANSAC-based fitting method (RATEC) to classify picked arrival times and detect possible events. RATEC discriminates true event arrival times from false picks by associating them with some reasonable moveout curves. Tests with synthetic data show that RATEC performs well for a 1-D linear array under layered medium assumption, as well for non-layered media and in the presence of dispersion. RATEC is

also expandable to the case of 2-D surface arrays by replacing the underlying hyperbolic curve model with a hyperboloid surface model. The effectiveness of event location for the 2-D case is demonstrated in a 5200-element dense 2-D array for earthquake monitoring at Long Beach, CA.

Acknowledgement

This work is supported by the Center for Energy and Geo Processing at Georgia Tech and King Fahd University of Petroleum and Minerals. We are grateful to Prof. Zhigang Peng and Zefeng Li for helpful discussions and the analysis of microearthquake data. The seismic data analyzed in this study are owned by Signal Hill Petroleum, Inc. and acquired by NodalSeismic LLC. We thank NodalSeismic LLC for making the one-week data available in this study.

References

- Akram J. and Eaton D. W. 2016. A review and appraisal of arrival-time picking methods for downhole microseismic data. *Geophysics* **81**, (2), KS71–KS91.
- Allen R. V. 1978. Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America* **68**, (5), 1521–1532.
- Artman B., Podladtchikov I. and Witten B. 2010. Source location using time-reverse imaging. *Geophysical Prospecting* **58**, (5), 861–873.
- Basu M. 2002. Gaussian-based edge-detection methods-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **32**, (3), 252–260.
- Choi S., Kim T. and Yu W. 2009. Performance evaluation of ransac family. *Journal of Computer Vision* **24**, (3), 271–300.
- Chum O. and Matas J. 2002. Randomized RANSAC with Td, d test. *Proc. British Machine Vision Conference* **2**, 448–457.
- Chum O. and Matas J. 2005. Matching with PROSAC-progressive sample consensus. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* **1**, 220–226.
- Dellinger J., Muir F. and Karrenbach M. 1993. Anelliptic approximations for TI media. *Journal of Seismic Exploration* **2**, (1), 23–40.
- Dix C. H. 1955. Seismic velocities from surface measurements. *Geophysics* **20**, (1), 68–86.
- Duncan P. M. 2005. Is there a future for passive seismic? *First Break* **23**, (6), 77–80.
- Duncan P. M. and Eisner L. 2010. Reservoir characterization using surface microseismic monitoring. *Geophysics* **75**, (5), 75A139–75A146.
- Earle P. S. and Shearer P. M. 1994. Characterization of global seismograms using an automatic-picking algorithm. *Bulletin of the Seismological Society of America* **84**, 366–376.
- Fischler M. A. and Bolles R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* **24**, (6), 381–395.
- Fitzgibbon A., Pilu M. and Fisher R. B. 1999. Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**, (5), 476–480.
- František S., Valenta J., Anikiev D. and Eisner L. 2015. Semblance for microseismic event detection. *Geophysical Journal International* **201**, (3), 1362–1369.

- Gajewski D. and Tessmer E. 2005. Reverse modelling for seismic event characterization. *Geophysical Journal International* **163**, (1), 276–284.
- Geiger L. 1912. Probability method for the determination of earthquake epicenters from the arrival time only. *Bull. St. Louis Univ* **8**, (1), 56–71.
- Han L., Wong J. and Bancroft J. 2009. Time picking and random noise reduction on microseismic data. *CREWES Research Reports* **21**, 1–13.
- Luu K., Noble M. and Gesret A. 2016. A competitive particle swarm optimization for nonlinear first arrival travelttime tomography. *SEG International Exposition and 86th Annual Meeting*, 2740–2744.
- Martin G. S., Wiley R. and Marfurt K. J. 2006. Marmousi2: An elastic upgrade for Marmousi. *The Leading Edge* **25**, (2), 156–166.
- Maxwell S., Rutledge J., Jones R. and Fehler M. 2010. Petroleum reservoir characterization using downhole microseismic monitoring. *Geophysics* **75**, (5), 75A129–75A137.
- Nakata N. and Beroza G. C. 2016. Reverse time migration for microseismic sources using the geometric mean as an imaging condition. *Geophysics* **81**, (2), KS51–KS60.
- O’Leary P. and Zsombor-Murray P. 2004. Direct and specific least-square fitting of hyperbola and ellipses. *Journal of Electronic Imaging* **13**, (3), 492–503.
- Sabbione J. I. and Velis D. 2010. Automatic first-breaks picking: New strategies and algorithms. *Geophysics* **75**, (4), V67–V76.
- Sharan S., Herrmann F., Wang R. and Leeuwen T. V. 2016. Sparsity-promoting joint microseismic source collocation and source-time function estimation. *SEG International Exposition and 86th Annual Meeting*, 2574–2579.
- Stewart C. V. 1995. MINPRAN: a new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**, (10), 925–938.
- Takanami T. and Kitagawa G. 1991. Estimation of the arrival times of seismic waves by multivariate time series model. *Annals of the Institute of Statistical mathematics* **43**, (3), 407–433.
- Toldo R. and Fusiello A. 2008. Robust multiple structures estimation with j-linkage. *10th European Conference on Computer Vision*, 537–547.
- Tordoff B. J. and Murray D. W. 2005. Guided-MLESAC: faster image transform estimation by using matching priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, (10), 1523–1535.

- Torr P. H. and Zisserman A. 2000. MLESAC: a new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding* **78**, (1), 138–156.
- Wang H. and Suter D. 2004. Robust adaptive-scale parametric model estimation for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, (11), 1459–1474.
- Witten B. and Shragge J. 2016. Full-wavefield tomography for seismic monitoring *SEG International Exposition and 86th Annual Meeting*, 2513–2517.
- Tan Y., He C., Deng P. and Cao N. 2014. Automatic microseismic event detection based on multi-channel semblance coefficient. *CPS/SEG Beijing 2014 International Geophysical Conference*, 1083–1086.
- Zhang H. and Thurber C. H. 2003. Double-difference tomography: The method and its application to the Hayward Fault, California. *Bulletin of the Seismological Society of America* **93**, (5), 1875–1889.
- Zhebel O. and Eisner L. 2015. Simultaneous microseismic event localization and source mechanism determination. *Geophysics* **80**, (1), KS1–KS9.
- Zhu L., Liu E. and McClellan J. H. 2015. Full waveform microseismic inversion using differential evolution algorithm. *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 591–595.
- Zhu L., Liu E. and McClellan J. H. 2016. An Automatic Arrival Time Picking Method Based on RANSAC Curve Fitting. *78th EAGE Conference and Exhibition*, Th SBT3 03.
- Waldhauser F. and Ellsworth W. L. 2000. A Double-Difference Earthquake Location Algorithm: Method and Application to the Northern Hayward Fault, California. *Bulletin of Seismological Society of America*, **90**, (6), 1353–1368.

A Parametric model of moveout curves

In many microseismic applications, accurate velocity models may not be available. However, a layered medium is commonly assumed for a shale rock region, in which case an estimate of the event location can be inferred from the arrival-time moveout curve across the monitoring geophone array. Using arrival times not only has a clear physical meaning but also it turns out to be computationally efficient. The primary requirement for this method to work is that there exists a parametric model $T(x)$ that approximates the arrival time T versus horizontal offset x . By estimating the finite number of model parameters, an event location can be uniquely determined. Over the years, such parametric models have been gradually updated and generalized for various types of media.

A.1 Homogeneous medium

For a homogeneous medium, the geometry of ray tracing is shown in Figure 17a. For an event originating at time T_0 and location (x_0, h) , a sensor at x will receive the signal at time $T(x)$, so the relation between source-to-receiver travel time $T(x) - T_0$ and the horizontal offset $(x - x_0)$ is

$$v^2[T(x) - T_0]^2 = \underbrace{v^2[T(x_0) - T_0]^2}_{=h^2} + (x - x_0)^2. \quad (13)$$

where we note that the zero-offset travel time is $T(x_0) - T_0 = h/v$. The travel-time equation (13) can be rewritten in a form that is recognizable as the standard form of a hyperbola

$$\frac{[T(x) - T_0]^2}{[T(x_0) - T_0]^2} - \frac{(x - x_0)^2}{v^2[T(x_0) - T_0]^2} = 1. \quad (14)$$

Thus, an event originating at time T_0 and location $(x_0, v[T(x_0) - T_0])$ can be uniquely determined by estimating the parameters T_0 , x_0 , $T(x_0)$, and v in equation (13), or (14). The estimation involves fitting a hyperbola to the picked arrival times $T(x_j)$ in a linear surface array.

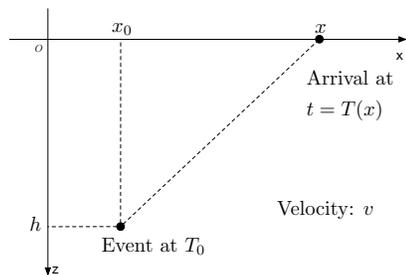


Figure 15:

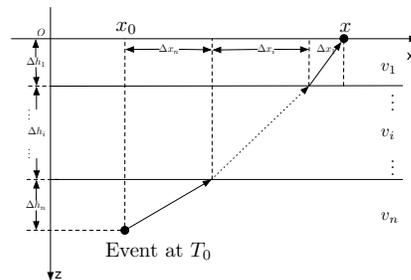


Figure 16:

Figure 17: Geometry of ray paths and travel time for (a) homogeneous medium with velocity v , and (b) a layered media

A.2 Layered isotropic media

A layered isotropic media, shown schematically in Figure 17b, is a little more complex than a homogeneous medium. The travel time ΔT_i and horizontal offset Δx_i of the i -th layer can be modeled as follows:

$$\begin{aligned}\Delta x_i &= h_i \tan \theta_i = \frac{ph_i v_i}{\sqrt{1 - (pv_i)^2}}, \\ \Delta T_i &= \frac{\Delta h_i}{v_i \cos \theta_i} = \frac{h_i}{v_i \sqrt{1 - (pv_i)^2}},\end{aligned}\tag{15}$$

where $p = \sin \theta_i / v_i$ is the ray parameter in Snell's law which is constant over all layers. Within each layer, the travel time has a hyperbolic relationship with offset, i.e., $\Delta T_i(x_i)$ defines a hyperbola, so the overall travel time $\Delta T(x) = T(x) - T_0$ computed as the sum is not exactly a hyperbola

$$\begin{aligned}\Delta x &= \sum_{i=1}^n \Delta x_i = \sum_{i=1}^n \frac{ph_i v_i}{\sqrt{1 - (pv_i)^2}}, \\ \Delta T &= \sum_{i=1}^n \Delta T_i = \sum_{i=1}^n \frac{\Delta h_i}{v_i \sqrt{1 - (pv_i)^2}}.\end{aligned}\tag{16}$$

However, (Dix, 1955) proved that a layered isotropic media behaves approximately like the homogeneous model when the offset x is close to zero. In other words, a hyperbolic moveout curve is observed near $\Delta x = 0$ for an isotropic layered model with an equivalent velocity of

$$v_{\text{RMS}}^2 = \frac{\sum_{i=1}^n v_i^2 \Delta T_i}{\sum_{i=1}^n \Delta T_i}.\tag{17}$$

In microseismic monitoring the receiving array is usually positioned over the top of the monitored events, so the offset x should be close to zero and the approximation (17) can be used. Moreover, Dix (1955) gave a correction for a tilted layered model as well—the (approximate) relationship between T and x is still described by a hyperbolic curve

$$\Delta T(x)^2 = \Delta T(0)^2 + \frac{\Delta x^2}{(v_{\text{RMS}} / \cos \theta)^2}.\tag{18}$$

where θ is the tilt angle.

A.3 Parametric model for TI media

It is the parametric model rather than a hyperbolic curve that is essential to the data fitting method we will propose. In cases of transverse isotropic (TI)

media, Dellinger *et al.* (1993) gave an elliptic approximation of the arrival-time moveout curve

$$\Delta T(x)^2 = \frac{T(0)^4 + (F_W + 1)T(0)^2 V_{\text{NMO}}^{-2} \Delta x^2 + F_W^2 V_{\text{NMO}}^4 \Delta x^4}{T(0)^2 + F_W^2 V_{\text{NMO}}^{-2} \Delta x^2}, \quad (19)$$

where x is the offset, $T(0)$ is the vertical travel time, V_{NMO} is the near-offset NMO velocity, and F_W is a dimensionless anisotropy parameter. Although equation (19) is a rational form with a fourth-order numerator and seems to be far from a hyperbolic curve, the basic idea is still valid: fitting a parametric model for $T(x)$ to localize an event. In this paper, we use the simpler hyperbolic model as a demonstration. The method to be proposed can be extended easily to other types of parametric models that adapt to different types of media.