Automatic seismic phase picking based on unsupervised machine learning classification and content information analysis

E. V. Cano*, J. Akram*, and D. B. Peter*

*King Abdullah University of Science and Technology (KAUST),

Thuwal, Saudi Arabia

Running head: Automatic seismic phase picking

ABSTRACT

Accurate identification and picking of P- and S-wave arrivals is important in earthquake and exploration seismology. Often, existing algorithms lack in automation, multi-phase classification and picking, as well as performance accuracy. A new fully automated four-step workflow for efficient classification and picking of P- and S-wave arrival times on microseismic datasets is presented. First, time intervals with possible arrivals on the waveform recordings are identified using the fuzzy c-means clustering algorithm. Second, these intervals are classified as corresponding to P, S, or unidentified waves using the polarization attributes of the waveforms contained within. Third, the P-, S-, and unidentified-waves arrival times are picked using the Akaike information criterion picker on the corresponding intervals. Fourth, unidentified waves are classified as P or S based on the arrivals moveouts. The application of the workflow on synthetic and real microseismic datasets shows that it yields accurate arrival picks for both high and low signal-to-noise ratio waveforms.
INTRODUCTION

Hypocenter locations are in most cases estimated either using arrival times (e.g., by linearized inversion, grid-search methods; see Buland, 1976; Pavlis, 1986; Moser et al., 1992; Oye and Roth, 2003) or waveform-based approaches (e.g., by time-reverse migration; see Artman et al., 2010; Nakata and Beroza, 2016). In the former approach, accurate arrival picking of P- and S-waves is critical to the accurate estimation of hypocenter locations. These picked arrival times are used in the polarization analysis for receiver orientations and back-azimuths, in the velocity model calibration and ultimately in the direct estimation of hypocenter locations. Therefore, any errors in the arrival-time picks can cause significant uncertainty in the estimated hypocentral parameters.

The arrival picking of P- and S-waves on microseismic datasets is, nonetheless, a challenging endeavor due to poor signal-to-noise ratio (SNR) of the waveforms and large data volumes (days to weeks of continuous recordings). Previously, numerous automatic arrival picking methods have been proposed (see, e.g., Akram and Eaton (2016) who compare different algorithms including the short- and long-term average ratio, STA/LTA, Akaike information criterion (AIC), phase arrival identification-kurtosis (PAI-K), and cross-correlation pickers). Recently, many supervised and unsupervised machine learning methods have also gained considerable popularity. Gentili and Michelini (2006) pick P- and S-phases using shallow neural networks with four manually defined input features, including variance, absolute values of skewness and kurtosis, and a combination of skewness and kurtosis. Similarly, Maity et al. (2014) use a neural network with two hidden layers and four input features including the variance of the sum of absolute values, and other attributes based on the wavelet coefficients and envelope functions. More recently, many applications of deep learn-
ing algorithms for arrival picking have been developed (e.g. Ross et al., 2018; Wang et al., 2019; Zhu and Beroza, 2019).

Neural network approaches typically belong to supervised machine learning methods (e.g., artificial neural networks) and have a high success rate, in cases where a good training dataset is available. Since the ground truth is known \textit{a priori} for the training set, direct quantification of the accuracy of the learning algorithm is possible (Ross et al., 2018). Nonetheless, the availability of an adequate training set often serves as a potential bottleneck, affecting the learning process. It is still ongoing research how to construct an optimal training dataset for a given problem. Supervised deep learning applications thus generally contain ad hoc choices and assumptions on the neural network architecture as well as a sparse coverage of the parameter model space represented by a training dataset. Additionally, training sets need manual labeling which can take a considerable amount of user time.

Conversely, unsupervised machine learning methods can be used in arrival picking without requiring any training dataset as they rely on the data itself. For instance, Zhu et al. (2016) and Chen (2020) pick first arrivals on raw, noisy microseismic data using the fuzzy \( c \)-means clustering (FCM) algorithm. In the arrival picking context, the FCM method computes a time series, called membership function, where abrupt increases indicate wave arrivals. Although FCM is efficient in detecting arrivals, the accurate identification of the instance where the membership function increases is challenging, causing picking inaccuracies. Additionally, some datasets (e.g., downhole microseismic monitoring) require picking of both \( P \)- and \( S \)-waves which the above method cannot determine reliably in its current state. For \( P \)- and \( S \)-arrival picking, additional modifications to the current workflow, as described in Zhu et al. (2016) and Chen (2020), are, therefore, necessary.
Here, we present a new fully-automated workflow capable of picking $P$ and $S$ arrivals not only on events where both phases are present but also on events where only one of them exists (single-phase events). First, we use the FCM as described in Chen (2020) to identify multiple signal intervals (if present) in the analysis window. Second, we classify these intervals either as $P$, $S$, or unidentified wave using polarization analysis on the waveforms in the signal intervals. Third, we pick the arrival times of $P$, $S$, and unidentified waves using the AIC picker on the waveforms in the corresponding intervals. Fourth, we fit the $P$ and $S$ moveouts with a quadratic function and use them to classify unidentified picks as $P$ or $S$. Finally, to evaluate the workflow performance, accuracy, and computational cost, we test it on synthetic and real microseismic data, and for the synthetic data, conduct a hypocenter location analysis.

**FUZZY C-MEANS CLUSTERING**

The fuzzy $c$-means (FCM) clustering partitions a set of $N$ points $X = x_1, \ldots, x_k, \ldots, x_N$, in a $F$-dimensional Euclidean space into $C$ clusters by minimizing the objective function (Dunn, 1973; Zadeh, 1977; Bezdek, 1981; Bezdek et al., 1984; Zhu et al., 2016; Cano et al., 2019; Chen, 2020):

$$J(U, V) = \sum_{k=1}^{N} \sum_{i=1}^{C} (u_{ik})^m \|x_k - v_i\|^2, \quad (1)$$

where $U$ is the partition matrix which elements $u_{ik} \in [0, 1]$ indicate the degree of membership of the point $x_k$ to the cluster $i$, $V = v_1, \ldots, v_i, \ldots, v_C$ is a set of $C$ points $v_i$ that represent the centroid of cluster $i$, $x_k$ is the $k$-th point of $X$, $m \in (1, \infty)$ is the controller of cluster fuzziness, and $\| \cdot \|$ is any norm.

One approach to minimize $J$ is to update the set of centroids $V$ and the partition matrix
U via iterations of:

\[
v_i = \frac{\sum_{k=1}^{N} u_{ik}^m x_k}{\sum_{k=1}^{N} u_{ik}^m}, \quad 1 \leq i \leq C, \quad (2)
\]

\[
u_{ik} = \frac{1}{\sum_{j=1}^{C} \left( \frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{\frac{2}{m-1}}} \quad 1 \leq k \leq N; 1 \leq i \leq C, \quad (3)
\]

where equation 3 has the constraint \( \sum_{i=1}^{C} u_{ik} = 1 \) for all \( k \).

The similarity metric of the points and the shape of the clusters depend on the choice of norm \( \| \cdot \| \). Here, we use the L2 norm, which induces a similarity metric based on Euclidean distance and clusters of hyperspherical shape.

**AKAIKE INFORMATION CRITERION (AIC) PICKER**

The Akaike Information Criterion (AIC) is a model selection technique developed by Akaike (1973), which can also be used for picking the onset of seismic phases on a single component trace. It assumes that a seismic trace can be divided into locally stationary segments where each is modeled as an autoregressive process. The onset time of a wave arrival separates two different segments and is associated with the minimum of the AIC values (Oye and Roth, 2003; Sleeman and Eck, 1999; Akram and Eaton, 2016).

Typically, the calculation of AIC function requires the estimation of autoregressive model coefficients but Maeda (1985) uses the following relation to calculate the AIC function directly from the input trace (waveform):

\[
AIC(k) = k \log(\text{var}\{x(1,k)\}) + (N - k - 1) \log(\text{var}\{x(k+1,N)\}), \quad (4)
\]

where \( x \) is a trace of \( N \) samples, \( k \) ranges from 1 to \( N \), and \( \text{var}\{x\} \) is the variance function.

In this study, we use equation 4 to pick the onset of P- and S-wave arrivals. Since the AIC
picks on the global minimum, it is important that we first identify $P$ and $S$ intervals and then apply the picker to the corresponding intervals (Akram and Eaton, 2016).

**AUTO-PICKING WORKFLOW**

The automatic arrival-time picking workflow (explained in Figure 1) comprises four main stages: (1) signal identification, (2) wave classification, (3) arrival-time picking, and (4) unidentified picks classification. For an adequate performance, our workflow requires a provisionally detected event where $P$- and $S$-wave arrivals occur only once. This is a regular strategy that usually simplifies arrival picking (Akram and Eaton, 2016). Additionally, because we determine $S$-wave picks as the average of picks on the $SV$ and $SH$ components, our method is limited to isotropic media to avoid inaccuracies due to shear-wave splitting.

[Figure 1 about here.]

For the binary clustering problem of signal identification, we define the samples $k = 1, ..., N$, of a trace $d(k)$ as a set of points $X = x_1, ..., x_k, ..., x_N$, in a $F$-dimensional Euclidean space. The elements of $x_k$ represent the value of some feature of $d(k)$ (e.g. mean) at the sample $k$. $F$ represents the number of features, and $N$ is the total number of samples in the analysis window. We denote the cluster number by $i$, where $i = 1$ is the noise cluster and $i = 2$ is the signal cluster, and the centroid of cluster $i$ by $v_i$. Among many existing features, we use the following three in this study:

- Mean of the absolute value of the amplitude:

$$M(k) = \frac{1}{N} \sum_{k-w}^{k+w} |d(k)|,$$  \hspace{1cm} (5)
– Peak power spectral density:

\[
P(k) = \max(|D(k, \omega)|^2),
\]  

(6)

– Short- and long-term average ratio:

\[
Q(k) = \frac{STA}{LTA} = \frac{1}{SW} \sum_{j=k}^{k+SW} |d(j)|, \\
\frac{1}{LW} \sum_{j=k-LW}^{k} |d(j)|
\]  

(7)

where in equation 5, the constant \( w \) is half the length of a window around the sample \( k \), in equation 6, \( D(k, \omega) \) is the modulus of the discrete short-time Fourier transform of \( d(k) \), and in equation 7, \( SW \) and \( LW \) are the lengths of the short- and long-term windows. For a correct signal identification, we suggest to estimate the dominant period of the arrival of interest (i.e., using time-frequency analysis), \( T_{dom} \), and set \( w \approx 0.5 \times T_{dom} \), \( SW \approx 1.5 \times T_{dom} \), and \( LW \approx 5 \times SW \). After computing the features of \( d(k) \), we define the points:

\[
x_k = [M(k), P(k), Q(k)]^\top,
\]  

(8)

and apply the fuzzy clustering (equations 1-3) to obtain \( u_2(k) = u_{i=2,k} \), the membership degree of the sample \( k \) to the signal cluster. For a 3C record, we carry this process on the components \( c = 1, 2, 3 \), and assuming that the signal arrives simultaneously on the three components, we stack their signal-cluster membership degrees, \( u_{2,c}(k) \), to highlight the wave arrivals:

\[
u_s(k) = \frac{1}{3} \sum_{c=1}^{3} u_{2,c}(k),
\]  

(9)

where \( u_s(k) \) is the stacked signal-cluster membership degree at the sample \( k \). Finally, we apply a threshold \( \beta \) to identify the signal intervals. Any continuous interval where \( u_s \) is greater than \( \beta \) for at least 1.5 times \( T_{dom} \) duration is considered to be containing a possible
wave arrival, so the other remaining intervals are deemed noise and discarded from further analysis. Here, we set $\beta$ between 1.0-2.0 times the mean value of $u_s$, depending on the SNR of the data.

Because one or both $P$ and $S$ waves can exist in the analysis window, we need some criteria to identify the intervals corresponding to the desired arrivals. We do so by computing the rectilinearity of the waveforms contained in each interval. First, we form a $(N_i \times 3)$ matrix $D$, where $N_i$ is the number of samples in the analyzed interval. In our workflow, $N_i$ is determined automatically. The limits of each signal interval, and thus $N_i$, are defined by the intersection points of the threshold $\beta$ with the peak of $u_s$ corresponding to the interval of interest. Each column of $D$ contains one of the three waveforms in the interval. After $D$ is created, we compute the rectilinearity as follow:

$$R = 1 - \frac{\sigma_3^2}{\sigma_1^2},$$

(10)

where $R$ is the rectilinearity and $\sigma_1^2, \sigma_3^2$ are the first and third eigenvalues of $D$ (Jurkevics, 1988). Next, we determine the first signal interval with an acceptable high rectilinearity value as the first arrival. If there are no intervals at a later time than the first arrival, we label it as unidentified and classify it as $P$ or $S$ wave at the end of the workflow. This is because, in this situation and at this stage, it is complicated to determine whether the first arrival is a $P$ or $S$ wave, as both waves can exhibit large rectilinearity values. Otherwise, if an interval exist later than the first arrival, we assume the first arrival to be a $P$ wave. Then, we rotate the waveforms to ray-centered coordinates using the polarization information from the selected $P$ interval and find the interval with the maximum $S$ energy on $s1$ and $s2$ components.

Following the wave-interval classification, we pick the onset of $P$, $S$, and unidentified
waves in the corresponding intervals. Although a threshold-based arrival-picking methodology from \( u_s \) (as given in Chen, 2020) can be adopted, the results it yields can be highly unstable for noisy datasets. We, therefore, apply the AIC algorithm to the \( P \) interval on the \( p \) components to pick the \( P \)-wave arrival time. For the \( S \) arrival time, we average the AIC picked times from the \( S \) interval on \( s1 \) and \( s2 \) components. In the case of unidentified waves, we obtain the arrival time using the AIC method on the component with highest SNR on the corresponding interval. It is worth mentioning that the AIC is not the only algorithm that we can use. Other algorithms, such as PAI-K (Saragiotis et al., 2002) and cross-correlation-based (VanDecar and Crosson, 1990; Song et al., 2010) pickers, will work equally fine.

Figure illustrates the application of auto-picking workflow on 3C waveforms from a single receiver level. Both \( P \) and \( S \) phases on the input waveforms have strong amplitudes on each of the three components (Figure a). For each of these components, mean, power spectral density and STA/LTA features are computed. Figure b shows the features for one of the components. In this case, all features and the signal-cluster memberships (Figure c) show a clearly distinguished response for the intervals containing \( P \) and \( S \) arrivals. These arrivals are easily identified by applying a thresholding criterion (Figure d). After the polarization analysis on the identified intervals, a \( P \)-wave interval is selected based on the order of occurrence and a high rectilinearity value. The data are then rotated into ray-centered coordinates \((p, s1, s2)\) to maximize the amplitudes of \( P \) and \( S \) waves on the corresponding components (Figure e). Both \( P \)- and \( S \)-wave arrival times are accurately picked using the AIC picker on the \( p \)-components, and \( s1 \) and \( s2 \)-components, respectively.

[Figure 2 about here.]
Once all the \( P \) and \( S \) arrivals on the event are picked, we classify any unidentified picks as \( P \) or \( S \), as illustrated in Figure 3. First, we temporary label unidentified picks as \( S \) picks (Figure b). Next, we estimate the \( S \)-wave moveout by fitting the \( S \) picks with a quadratic function using the Random Sample Consensus (RANSAC) method (Fischler and Bolles, 1987). For an explanation of how RANSAC is used to estimate moveout curves, we refer the reader to Zhu et al. (2017). Finally, once the \( S \) moveout is estimated (Figure c), we compute the time difference between the unidentified picks and the fitted \( S \) moveout curve. Any unidentified pick between \( \pm T_{dom} \) from the \( S \)-moveout curve is classified as an \( S \) pick, and any remaining unidentified picks are classified as \( P \) (Figure d). We also correct any \( P \) picks on \( S \) waves using the same moveout criteria.

For events where all picks are unidentified (Figure a), the previous strategy cannot be used, as fitting the \( P \) or \( S \) moveout is not possible. In this situation, we generate a database of \( P \) and \( S \) moveouts, obtained from events where \( P \) and \( S \) picks were available. We then fit the moveout of the unidentified arrivals and compare it with the moveouts database (Figure b). We do so by shifting all the moveouts to a common time and computing the following coefficients:

\[
P_{res} = \frac{1}{N_p} \sum_{i=1}^{N_p} \|u_{mov} - p_{mov_i}\|_2, \tag{11}
\]

\[
S_{res} = \frac{1}{N_s} \sum_{i=1}^{N_s} \|u_{mov} - s_{mov_i}\|_2, \tag{12}
\]

where \( u_{mov}, p_{mov}, \) and \( s_{mov} \) are vectors containing the unidentified, \( P \), and \( S \) moveout curves, and \( N_p \) and \( N_s \) are the number of \( P \) and \( S \) moveouts in the database. The uniden-
tified moveout of interest, \( u\text{mov} \), and therefore the associated picks, are classified as \( P \) or \( S \) depending on whether \( P_{\text{res}} \) or \( S_{\text{res}} \) is the smallest value. This approach is useful to identify the wave type of single-phase events as illustrated in Figure c.

[Figure 4 about here.]

RESULTS

To evaluate the performance of the proposed arrival-time picking workflow, we apply it to synthetic and real microseismic data and compare the results with reference picks. For the synthetic data, we also carry a hypocenter location analysis.

To compare our workflow with existing methods, we use the STA/LTA and Chen (2020) algorithms on the synthetic and real datasets. Given that for a 3C record, our workflow potentially obtains one \( P \) and \( S \) pick and the STA/LTA and Chen (2020) methods obtain three first arrival picks (one per receiver component), we consider the picks of the STA/LTA and the Chen (2020) methods with the minimum difference to the reference picks as \( P \) picks. For the synthetic dataset, we use theoretical arrival times (computed using ray-tracing) as reference picks, while for the real dataset, we use manual picks.

To measure the picking accuracy, we compute the residual between the reference and automatic picks. We illustrate our results in scatter plots of arrival-pick residual against arrival SNR. To compute the arrivals SNR, we define a window of noise from the start of the record to one dominant period before the first arrival. Then, we set a window centered on the arrival of interest and compute the ratio between the root-mean-square (RMS) amplitude of each window. In the scatter plots, residuals of \( P \) and \( S \) picks are pictured in blue and red colors, respectively. Residuals of initially unidentified picks are indicated in green, and
the residuals of picks skipped by our workflow but determined by the STA/LTA and Chen (2020) methods are indicated in brown. We also plot black lines at -10 ms and 10 ms to highlight relatively accurate picks. Additionally, we show histograms of the residuals and compute their mean, $\mu$, and standard deviation, $\sigma$. We compute $\mu$ and $\sigma$ using residuals in the -50 ms to 50 ms interval to decrease the influence of large picking errors and obtain meaningful indicators of each algorithms' performance.

**Synthetic data**

*Arrival-time picking*

The synthetic data consist of 100 events recorded by a vertical downhole array of 20 3C receivers located in an elastic homogeneous-layered medium (Figure ). The events are defined by randomly distributed double-couple and tensile sources with moment magnitudes ranging from -3 to 1. A Berlage wavelet (Aldridge, 1990) with a dominant frequency of 30 Hz is used as the source time function, and the wave propagation is simulated using the SPECFEM3D Cartesian package (Komatitsch and Tromp, 1999). We create three synthetic datasets (Figure ) by adding white Gaussian noise (AWGN) with SNR of 20, -8, and -13 dB, and filtering the waveforms between 0.1 Hz to 100 Hz with a zero-phase fourth-order Butterworth bandpass filter. As reference picks, we use 2000 $P$- and 2000 $S$-arrival times computed with ray-tracing (Figure ).

[Figure 5 about here.]

[Figure 6 about here.]

On synthetic dataset one (AWGN SNR of 20 dB), $\mu$ and $\sigma$ of $P$ residuals are relatively
low among the methods, especially for the proposed and STA/LTA pickers (Figure a, d, g). This dataset presents low levels of noise; thus, the three methods have an acceptable outcome, as illustrated in event 33 in Figure . Nevertheless, the proposed workflow exhibits the lowest $\mu$ and $\sigma$ values (-0.66 ms and 2.99 ms), indicating less biased and more reliable picking. The majority of $P$ residuals range between -10 ms to 10 ms, with some outliers from relatively low SNR arrivals (< 20 dB), related to barely detectable $P$ waves (Figure a, d, g). For the proposed workflow, three $P$ picks were initially labeled as unidentified (Table 1). Their residuals are close to 0 ms, indicating accurate picking and classification of unidentified arrivals. Also, total of 221 $P$ arrivals were skipped by our method. This tends to occur when the $P$ arrival is buried by noise, has a poor SNR (receivers 8-11 on Figure a), or was picked on an $S$ wave and corrected by our moveout criteria. Of the skipped $P$ arrivals, the STA/LTA and Chen (2020) methods managed to pick 153 and 42 relatively accurate (residuals between -10 ms to 10 ms). The number of omitted $P$ arrivals picked by the STA/LTA and Chen (2020) methods with a low residual is related to the number of arrivals from which our workflow is losing information.

Regarding the $S$ residuals on synthetic dataset one, most of them range between -10 ms to 15 ms (Figure a). The relatively large $S$ residuals associated with high SNR arrivals (> 40 dB) on Figure a occur due to contamination of the $S$-wave with precursory phases, as observed in Figures (receiver 7) and d. This increases the $\sigma$ value, explaining why $\sigma$ is 2.09 ms larger for $S$ residuals than for $P$ residuals. Moreover, a total of 221 $S$ picks were determined from unidentified picks (Table 1). These picks correspond to the waveforms where $P$ picks were skipped. The associated residuals are in the same range as the rest of the $S$ picks, suggesting a correct classification of unidentified arrivals.
The automatic picking results for synthetic dataset two (AWGN SNR of -8 dB) are less favorable than in dataset one. As observed in event 33 in Figure (receivers 1 to 14), the noise on this dataset masks $P$ arrivals that were detectable before (see Figure). Also, the presence of high amplitude noise foregoing the first arrivals increases. The previous noise-related effects result in an increase of picks on noise preceding first arrivals (large positive residuals), and picking of $P$ arrivals on $S$ waves (large negative residuals). This is especially true for the STA/LTA and Chen (2020) methods, as indicated by the increase of $\sigma$ compared to dataset one (Figure e and h) and illustrated on Figure e and h. In contrast, our workflow is less affected by the noise increment, as suggested by the lower $\sigma$ compared to the STA/LTA and Chen (2020) methods (Figure b) and the relatively-dense cloud of points centered at 0 ms in Figure b. The proposed workflow yields better results than the other two methods because it skips $P$ arrivals completely covered by noise, as shown in Figure a (receivers 1 to 6). Of the 2000 $P$ arrivals, 1014 were skipped (Table 1). Although this is more than half of the total $P$ arrivals in the dataset, the majority of them have poor SNR ($< 5$ dB) or were not recorded. From the skipped $P$ picks, the STA/LTA and Chen (2020) methods picked 295 and 149 between -10 ms and 10 ms, respectively.

As observed from Figure and , the noise on the synthetic dataset two has a small influence on the $S$ arrivals. The $S$ residuals have a similar distribution as in the less-noisy dataset one (Figure b), and the increase in the number of large residuals is small (Figure b). Compared to the $P$ residuals, the $S$ residuals have a lower $\sigma$ value, suggesting that picking is more reliable in $S$ waves that in $P$ waves. As pictured in Figure b, the number of $S$ picks determined from unidentified picks increased considerably compared to dataset one. This
is expected, as unidentified picks occur on waveforms where the $P$ wave was not detected. A total of 1014 $S$ picks were determined from unidentified picks, most of which have similar residuals than the rest of $S$ picks, implying correct classification of unidentified arrivals.

[Figure 8 about here.]

On synthetic dataset three (AWGN SNR of -13 dB), the noise covers more $P$ arrivals than in dataset two, and the amplitude of early noise grows. As consequence, the number of picks on noise with similar amplitude to the arrivals and picking of $P$ arrivals on $S$ waves increases (receivers 1 to 3 on Figure b and c). The distributions of $P$ residuals are relatively similar compared to dataset two (Figure c, f, and i). The Chen (2020) method has the lowest $\mu$ value, which indicates less biased picking than the STA/LTA and proposed methods. However, our workflow exhibits the lowest $\sigma$ suggesting it is the most reliable picker. The number of large $P$ residuals increases on all methods, especially for low SNR arrivals ($< 5$ dB; Figure c, f, and i). For the proposed workflow, some large $P$ residuals occur on relatively high SNR arrivals ($> 7$ dB). These residuals are related with picking on early noise. The proposed workflow determined 16 $P$ picks from unidentified picks. Some of these reclassified picks have large residuals, however, this is because the unidentified picks occur on early noise and not due to incorrect reclassification. Additionally, our workflow skipped 1239 $P$ arrivals (Table 1). As illustrated by the clusters of brown points in Figure f and i, most of the skipped arrivals have poor SNR ($< 5$ dB) and were not determined accurately by the STA/LTA and Chen (2020) methods, suggesting that the skipped $P$ arrivals were buried by noise.

Similar to dataset two, $S$ arrivals are less affected by noise than $P$ arrivals (Figures and ). The $\mu$ and $\sigma$ values are considerably lower for $S$ residuals than for $P$ residuals.
There is not a high increase of $\mu$ and $\sigma$ compared to previous datasets (Figure c). Also, the number of large $S$ residuals remains almost the same, with exception of an outlier around 140 ms related to incorrect detection of $P$ and $S$ waves (Figure e). As hinted by the similar residuals between the 1237 $S$ picks determined from unidentified arrivals and the remaining $S$ picks (Figure c), the unidentified picks classification was successful.

[Hypocenter location]

To determine if our workflow results are useful for accurate event location, we locate the 100 events of the synthetic data using the true velocity model and the picks and events back-azimuths estimated by our workflow. Then, we compute the location error using the true hypocenters as reference. For each event in the following analysis, we define the location RMS error (RMSE) as the RMS of residuals on the north, east and depth coordinates. We also define the arrival picking RMSE as the RMS of $P$ and $S$ picks residuals. Additionally, we compute the difference between the true event back-azimuth and the back-azimuth
estimated by our workflow on each 3C record. We define the back-azimuth RMSE as the RMS of back-azimuth residuals.

On Figure , the dots show the location errors on the north, east and depth coordinates of the located hypocenters. We can observe that on the three datasets, the largest location uncertainty occurs on the north coordinate, (transverse direction), followed by the depth and by the east coordinate (radial direction). For synthetic dataset one, the majority of events (84) have a location RMSE below 30 m, a low value considering the array geometry. On dataset two, five single-phase (S wave) events were not located. A total of 52 events have a location RMSE below 45 m, which is a realistic value considering the level of noise on this dataset. The location errors increase on dataset three. The location RMSE is under 58 m for 50 of the 95 located events.

Figure shows scatter plots of location RMSE against arrival-picking and back-azimuth RMSE. The dots color indicate the number of receivers used during the event localization (receivers with available P and S picks). As expected, the location RMSE is positively correlated with arrival picking and back-azimuth RMSE in all datasets. That is, location is poor for events with inaccurate arrival picks and estimated back-azimuths. Also, we can observe that the location RMSE tends to be lower for events where more than ten receivers were used. Nonetheless, this is not general, as some events with low arrival-picking and back-azimuth RMSE were located relatively accurately using less than five receivers.

[Figure 14 about here.]

[Figure 15 about here.]
Real data

The real microseismic dataset was acquired during a hydraulic fracturing operation by an array of 20 3C receivers placed in a vertical monitoring well. The recorded waveforms were sampled at 0.5 ms interval. Using time-frequency analysis, we estimate that the dominant frequency of the arrivals is 100 Hz. To denoise the data, we filter the waveforms using a zero-phase fourth-order Butterworth bandpass filter. Due to receiver limitations, we set the lower cutoff frequency to 10 Hz. We set upper cutoff frequencies between 200 to 300 Hz. For this study, we use only a dataset containing 40 previously detected events. To compare the automatic picks, we carry manual picking on waveforms where \(P\) and \(S\) arrivals were recorded. Of the potential 800 \(P\) arrivals and 800 \(S\) arrivals, we retrieve 694 \(P\) and 714 \(S\) picks (106 \(P\) and 86 \(S\) missing reference picks).

Figure shows the \(P\) residuals distribution obtained by the automatic picking algorithms. For all methods, most of residuals are in the -10 ms to 10 ms range, indicating relatively precise picking, as shown in Figures and . The distribution of the STA/LTA method exhibits the largest spread and a positive bias, which is a consequence of picks on pre-signal noise as illustrated in Figure b (receivers 2 to 4). The the proposed workflow exhibits the lowest \(\mu\) and \(\sigma\) values among the methods. As in the synthetic data, the arrival-picking omission on waveforms with \(P\) arrivals buried in noise (receivers 16 and 19 in Figure a) reduces the number of incorrect picks by our workflow. In this dataset, 95 \(P\) arrivals were skipped by our method (Table 2), from which 61 correspond to missing reference picks. Of the remaining 34 arrivals, the STA/LTA and Chen (2020) methods picked 10 and 22 with a residual between -10 ms to 10 ms, and the rest correspond to outliers (Figure ). Also, a total of 5 \(P\) picks were correctly determined from unidentified picks.
Compared to $P$ residuals, $S$ residuals have a slightly higher $\mu$ and $\sigma$ (Figure ). Great part of residuals occur between -10 to 10 ms, indicating relatively accurate picking as shown in Figures a, a, and a. Additionally, high ($> 30$ dB) and low ($< 20$ dB) SNR outliers are present. The high SNR outliers on Figure are related to complex waveforms where high amplitude phases exist after the $S$ wave, resulting on $P$ picking on the $S$ wave, and $S$ picking on a late phase (Figure b). Regarding classification of unidentified arrivals, a total of 85 $S$ picks were obtained from unidentified picks, all with similar residuals as the rest of $S$ picks. This indicates a successful classification of unidentified arrivals.

DISCUSSION

Overall, the proposed workflow obtains lower mean and standard deviation values of $P$ residuals than the STA/LTA and Chen (2020) methods on the synthetic and real datasets.
This suggests that the $P$-arrival picking carried by our workflow is less biased and more stable than that of the other two methods. The main observed problems of the STA/LTA and Chen (2020) methods are picking on pre-signal noise and $P$ picking on $S$ waves. High amplitude noise tends to generate high STA/LTA values, and depending on the used trace features, this also occurs for FCM. This results in picks on pre-signal noise, as these methods set the arrival time on the earliest jump of STA/LTA and membership values. On the other hand, in cases where $P$ waves are masked by noise, the earliest increase in STA/LTA and membership values is usually related to $S$ waves, resulting in $P$ picks on $S$ waves. Because our workflow selects $P$ arrivals based on the rectilinearity of waveforms contained in signal intervals determined from FCM membership values, high amplitude noise is avoided most of the times. Also, the capacity of our workflow to detect waveforms containing one "unidentified" arrival and to classify it as a $P$ or $S$ aids in the picking omission of not recorded phases and allows picking and phase identification on single-phase events.

Based on the mean and standard deviation values of $S$ residuals on the synthetic dataset one and the real dataset, $P$ picking is slightly more accurate and reliable than $S$ picking. The $S$ residuals present slightly higher mean and standard deviation values than those of $P$ residuals on these two datasets. However, when the noise level increases and buries $P$ waves, $S$ picking by our workflow is more robust. As illustrated in Figures -, $S$ waves are less affected by noise increments, facilitating arrival picking.

The proposed workflow is not exempt from drawbacks. Because the first arrival is set as the earliest signal interval with high rectilinearity, noise with high rectilinearity values may be set as the first arrival. We also need to consider that the rectilinearity estimation of $P$ waves can be affected by noise. Another factor that decreases the workflow performance is waveform complexity. Contamination of $S$ waves by preceding phases can reduce the
picking accuracy, and the presence of phases other than direct $P$ and $S$ waves may result in incorrect arrival windowing. Additionally, on all datasets, our workflow skipped weak $P$ arrivals that were picked by the STA/LTA and Chen (2020) methods with acceptable precision. The previous occurred especially on the synthetic data, where the majority of these weak arrivals were recorded only on one component. Because we average the three components’ membership values to determine signal intervals, low SNR arrivals recorded on one component may be missed by our workflow. Despite that the STA/LTA and Chen (2020) picked these arrivals with decent accuracy, it is important to remember that these two methods obtain three picks per 3C record, and here we only consider the pick with minimum residual. In practice, one pick per receiver must be determined, which may reduce the picking accuracy of these methods when low SNR arrivals are recorded only in one component.

Despite the previous drawbacks, the picking carried by our workflow is accurate enough to obtain acceptable hypocenter locations. For events with high SNR waveforms ($> 20$ dB) as the ones in the synthetic dataset one, the arrival picks yielded by our method result in relatively accurate locations for 84 of 100 events (location RMSE $< 30$ m). As the noise level increases, the computed arrival times and event back-azimuths are less accurate, increasing the hypocenter location error. In synthetic dataset two, 52 events were located with an RMSE below 45 m, and in dataset three, the location RMSE of 50 events was below 58 m. These are acceptable values considering the low SNR of the waveforms on these datasets (-5 dB to 15 dB for dataset two and -5 dB to 10 dB for dataset three).

Regarding the speed and computational costs of the presented workflow, we ran the algorithm on MATLAB using a single-core of the Intel Core i7-9750H CPU processor at 2.6 GHz clock speed. For one 3C record of 0.8 s duration at 0.5 ms sampling rate, the
algorithm picks both P and S arrivals in $\sim 0.1$ s. For existing large datasets, the algorithm could be further parallelized to run in an embarrassingly parallel scheme, as each 3C trace is analyzed independently.

CONCLUSIONS

We present a new AIC assisted fuzzy c-means (FCM) clustering based auto-picking workflow for efficient identification and picking of P- and S-wave arrival times. The workflow is capable of skipping picking of phases (direct P and S waves) not recorded on the waveforms, reducing the number of inaccurate picks. Our workflow also allows picking and phase identification of single-phase events by estimating and comparing the $P$ and $S$ moveouts of analyzed events. This workflow is fully automatic, meaning that almost all of the parameters (e.g., window duration for trace features computation and time difference threshold to classify unidentified picks based on arrival moveouts) are associated with a user-specified estimate of the signal’s dominant period. The computational costs of this workflow are very low compared to supervised machine learning approaches.

As other arrival-time pickers, this workflow has limitations. First, this workflow only works on previously detected events, as the main workflow component involves a fuzzy c-means clustering to partition signals from noise. For noise-only traces, FCM generates two clusters containing noise with different behaviour, making signal detection erratic. Second, the workflow accuracy may decrease in the presence of phases different than direct $P$ and $S$ waves. Third, high rectilinearity noise may result in incorrect identification of $P$ waves. Fourth, our workflow may omit picking on arrivals with minuscule amplitude recorded only on one component. By analyzing hypocenter locations with different noise levels, we determine that the picking accuracy is compromised for waveforms with SNR $< 20$ dB.
Despite these drawbacks, tests on synthetic and real data show that our method is more robust than existing methods. Furthermore, there is room to improve the proposed workflow by using more sophisticated trace features or different picking methods on the detected $P$ and $S$ intervals.

ACKNOWLEDGMENTS

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST). The authors would like to thank Leo Eisner, the editor, and the anonymous reviewers for their helpful comments that improved this work. The real data is granted by an anonymous provider. Our open-source workflow and data used in this work can be found at http://doi.org/10.5281/zenodo.4553858.
REFERENCES


Fischler, M. A., and R. C. Bolles, 1987, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, in Readings in


Ross, Z. E., M. A. Meier, and E. Hauksson, 2018, P Wave Arrival Picking and First-Motion...


LIST OF FIGURES

1 Arrival-picking workflow. .................................................. 31

2 Arrival-time picking using the proposed workflow. a) Input 3C waveforms. b) Vertical component features. c) Fuzzy memberships for the three components. d) Stacked membership. e) Rotated data and arrival picks. The highlighted blue and red curves show the $P$ and $S$ intervals obtained from the fuzzy $c$-means clustering. The blue and red vertical lines represent the $P$ and $S$ picks. f) AIC values for the rotated data. The time associated with the minimum AIC value is the arrival time of a phase arrival. .................................................. 32

3 Classification of unidentified picks for events with available P and S picks. a) Unidentified picks. The $P$, $S$, and unidentified picks are indicated by the blue, red, and green lines, respectively. b) Picks used for $S$ moveout fitting (magenta lines). c) Estimated $S$ moveout (red dotted line). Any pick laying on the dark area will be classified as $S$ pick. d) Classified picks. The unidentified picks on panel c) are now classified as $S$ picks. .................................................. 33

4 Classification of unidentified picks for events without $P$ and $S$ picks. a) Unidentified picks. b) Moveout comparison. The $P$, $S$, and unidentified moveouts are in blue, red, and green, respectively. c) Classified picks. The initially unidentified picks are classified as $S$ picks. .................................................. 34

5 Synthetic experiment setup. a) Source-receiver geometry. b) Medium model. The left panel shows the $P$ and $S$ velocity models and the right panel the density model. .................................................. 35

6 Example of a synthetic event with different noise levels. The illustrated $P$ and $S$ picks are reference picks computed using ray-tracing. a) Synthetic dataset one (SNR=20 dB). b) Synthetic dataset two (SNR=-8 dB). c) Synthetic dataset three (SNR=-13 dB). .................................................. 36

7 Event 33 of synthetic dataset one (average $P$-wave SNR = 14.9 dB; average $S$-wave SNR = 43.1 dB). a) Proposed workflow picks. b) STA/LTA picks. c) Chen (2020) picks. The blue and red lines indicate the $P$ and $S$ picks. Skipped arrivals are indicated by asterisks. .................................................. 37

8 Event 33 of synthetic dataset two (average $P$-wave SNR = -0.3 dB; average $S$-wave SNR = 17.4 dB). a) Proposed workflow picks. b) STA/LTA picks. c) Chen (2020) picks. The blue and red lines indicate the $P$ and $S$ picks. The * and < symbols indicate skipped arrivals and early picks outside the figure time range, respectively. .................................................. 38

9 Event 33 of synthetic dataset three (average $P$-wave SNR = 0.2 dB; average $S$-wave SNR = 12.1 dB). a) Proposed workflow picks. b) STA/LTA picks. c) Chen (2020) picks. Symbols as indicated in Figure. .................................................. 39
10 Histograms of P-pick residuals obtained by the proposed workflow and the STA/LTA and Chen (2020) methods on the synthetic datasets. The black line indicates the residuals probability distribution in the -50 ms to 50 ms interval. The mean and standard deviation values are indicated by $\mu$ and $\sigma$, respectively. ............................................. 40

11 P-pick residuals obtained by the proposed workflow and the STA/LTA and Chen (2020) methods on the synthetic datasets. The residuals of P picks initially labeled as unidentified are in green and the residuals of picks skipped by the proposed workflow in brown. ............................................. 41

12 Histograms of S-pick residuals obtained by the proposed workflow on the synthetic dataset one (a), two (b), and three (c). Symbols as indicated in Figure ............................................. 42

13 Panels (a)-(c): S-pick residuals obtained by the proposed workflow on the synthetic datasets. The residuals of S picks initially labeled as unidentified are in green. d) Inaccurate S pick due to precursory phase contamination. e) Incorrect P and S picking. The P interval is erroneously set as noise with high rectilinearity, resulting in an incorrect determination of the S wave. In panels d) and e), the blue and red colors indicate P and S waves, respectively. The reference picks are indicated by the dashed lines and the automatic picks by the continuous lines. The highlighted curves indicate the windows where picking was carried using the AIC picker. ............................................. 43

14 Hypocenter location errors on the north, east, and depth coordinates. Each dot represents the hypocenter localization error of one event. ............................................. 44

15 Hypocenter location RMSE against arrival picking RMSE (panels a to c) and against back-azimuth RMSE (panels d to f). ............................................. 45

16 Event 8 of the real dataset (average P-wave SNR = 40.8 dB; average S-wave SNR = 41 dB). a) Proposed workflow picks. b) STA/LTA picks. c) Chen (2020) picks. Symbols as indicated in Figure ............................................. 46

17 Event 13 of the real dataset (average P-wave SNR = 23.2 dB; average S-wave SNR = 30.2 dB). a) Proposed workflow picks. b) STA/LTA picks. c) Chen (2020) picks. Symbols as indicated in Figure ............................................. 47

18 Event 28 of the real dataset (average P-wave SNR = 13.8 dB; average S-wave SNR = 19 dB). a) Proposed workflow picks. b) STA/LTA picks. c) Chen (2020) picks. Symbols as indicated in Figure ............................................. 48

19 Histograms of P-pick residuals obtained by the proposed workflow (a), the STA/LTA method (b), and the Chen (2020) method (c) on the real dataset. Symbols as indicated in Figure ............................................. 49

20 P-pick residuals obtained by the proposed workflow (a), the STA/LTA method (b), and the Chen (2020) method (c) on the real dataset. Color code as indicated in Figure ............................................. 50

21 Histograms of S-pick residuals obtained by the proposed workflow on the real dataset. Symbols as indicated in Figure ............................................. 51
a) S-pick residuals obtained by the proposed workflow on the real dataset.
b) Incorrect picking on complex waveform. Colors as indicated in Figure . . 52
LIST OF TABLES

1  Number of picked, initially unidentified, and skipped arrivals by the proposed workflow on the synthetic datasets. ........................................... 54
2  Number of picked, initially unidentified, and skipped arrivals by the proposed workflow on the real dataset. ........................................... 55
Figure 1. Arrival-picking workflow.

178x194mm (300 x 300 DPI)
Figure 2. Arrival-time picking using the proposed workflow. a) Input 3C waveforms. b) Vertical component features. c) Fuzzy memberships for the three components. d) Stacked membership. e) Rotated data and arrival picks. The highlighted blue and red curves show the $P$ and $S$ intervals obtained from the fuzzy $c$-means clustering. The blue and red vertical lines represent the $P$ and $S$ picks. f) AIC values for the rotated data. The time associated with the minimum AIC value is the arrival time of a phase arrival.
Figure 3. Classification of unidentified picks for events with available \textit{P} and \textit{S} picks. a) Unidentified picks. The \textit{P}, \textit{S}, and unidentified picks are indicated by the blue, red, and green lines, respectively. b) Picks used for \textit{S} moveout fitting (magenta lines). c) Estimated \textit{S} moveout (red dotted line). Any pick laying on the dark area will be classified as \textit{S} pick. d) Classified picks. The unidentified picks on panel c) are now classified as \textit{S} picks.
Figure 4. Classification of unidentified picks for events without \textit{P} and \textit{S} picks. a) Unidentified picks. b) Moveout comparison. The \textit{P}, \textit{S}, and unidentified moveouts are in blue, red, and green, respectively. c) Classified picks. The initially unidentified picks are classified as \textit{S} picks.

863x666mm (600 x 600 DPI)
Figure 5. Synthetic experiment setup. a) Source-receiver geometry. b) Medium model. The left panel shows the \textit{P} and \textit{S} velocity models and the right panel the density model.
Figure 6. Example of a synthetic event with different noise levels. The illustrated \textit{P} and \textit{S} picks are reference picks computed using ray-tracing. a) Synthetic dataset one (SNR=20 dB). b) Synthetic dataset two (SNR=-8 dB). c) Synthetic dataset three (SNR=-13 dB).
Figure 7. Event 33 of synthetic dataset one (average \textit{P}-wave SNR = 14.9 dB; average \textit{S}-wave SNR = 43.1 dB). a) Proposed workflow picks. b) STA/LTA picks. c) \cite{Chen_2020} picks. The blue and red lines indicate the \textit{P} and \textit{S} picks. Skipped arrivals are indicated by asterisks.
Figure 8. Event 33 of synthetic dataset two (average $P$-wave SNR = -0.3 dB; average $S$-wave SNR = 17.4 dB). a) Proposed workflow picks. b) STA/LTA picks. c) \cite{Chen_2020} picks. The blue and red lines indicate the $P$ and $S$ picks. The $*$ and $<$ symbols indicate skipped arrivals and early picks outside the figure time range, respectively.

768x583mm (600 x 600 DPI)
Figure 9. Event 33 of synthetic dataset three (average \textit{P}-wave SNR = 0.2 dB; average \textit{S}-wave SNR = 12.1 dB). a) Proposed workflow picks. b) STA/LTA picks. c) \cite{Chen_2020} picks. Symbols as indicated in Figure 8f{Fig8}. 

769x582mm (600 x 600 DPI)
Figure 10. Histograms of \textit{P}-pick residuals obtained by the proposed workflow and the STA/LTA and \cite{Chen_2020} methods on the synthetic datasets. The black line indicates the residuals probability distribution in the -50 ms to 50 ms interval. The mean and standard deviation values are indicated by $\mu$ and $\sigma$, respectively.

1143x923mm (600 x 600 DPI)
Figure 11. *P*-pick residuals obtained by the proposed workflow and the STA/LTA and \cite{Chen_2020} methods on the synthetic datasets. The residuals of *P* picks initially labeled as unidentified are in green and the residuals of picks skipped by the proposed workflow in brown.
Figure 12. Histograms of \textit{S}-pick residuals obtained by the proposed workflow on the synthetic dataset one (a), two (b), and three (c). Symbols as indicated in Figure 10.

\[\mu = 2.99\text{ ms} \quad \sigma = 5.08\text{ ms}\]
\[\mu = -3.02\text{ ms} \quad \sigma = 5.51\text{ ms}\]
\[\mu = -3.25\text{ ms} \quad \sigma = 7.22\text{ ms}\]
Figure 13. Panels (a)-(c): S-pick residuals obtained by the proposed workflow on the synthetic datasets. The residuals of S picks initially labeled as unidentified are in green. d) Inaccurate S pick due to precursory phase contamination. e) Incorrect P and S picking. The P interval is erroneously set as noise with high rectilinearity, resulting in an incorrect determination of the S wave. In panels d) and e), the blue and red colors indicate P and S waves, respectively. The reference picks are indicated by the dashed lines and the automatic picks by the continuous lines. The highlighted curves indicate the windows where picking was carried using the AIC picker.

1116x602mm (600 x 600 DPI)
Figure 14. Hypocenter location errors on the north, east, and depth coordinates. Each dot represents the hypocenter localization error of one event.

1181x978mm (600 x 600 DPI)
Figure 15. Hypocenter location RMSE against arrival picking RMSE (panels a to c) and against back-azimuth RMSE (panels d to f).

1158x652mm (600 x 600 DPI)
Figure 16. Event 8 of the real dataset (average \textit{P}-wave SNR = 40.8 dB; average \textit{S}-wave SNR = 41 dB). a) Proposed workflow picks. b) STA/LTA picks. c) \cite{Chen_2020} picks. Symbols as indicated in Figure ef{Fig8}.

851x673mm (600 x 600 DPI)
Figure 17. Event 13 of the real dataset (average $P$-wave SNR = 23.2 dB; average $S$-wave SNR = 30.2 dB). a) Proposed workflow picks. b) STA/LTA picks. c) Chen (2020) picks. Symbols as indicated in Figure ef{Fig8}.
Figure 18. Event 28 of the real dataset (average \textit{P}-wave SNR = 13.8 dB; average \textit{S}-wave SNR = 19 dB). a) Proposed workflow picks. b) STA/LTA picks. c) \cite{Chen_2020} picks. Symbols as indicated in Figure ef(Fig8).
Figure 19. Histograms of $P$-pick residuals obtained by the proposed workflow (a), the STA/LTA method (b), and the \cite{Chen_2020} method (c) on the real dataset. Symbols as indicated in Figure ef(Fig10).

1118x332mm (600 x 600 DPI)
Figure 20. \textit{P}-pick residuals obtained by the proposed workflow (a), the STA/LTA method (b), and the \cite{Chen_2020} method (c) on the real dataset. Color code as indicated in Figure ef{Fig11}. 

1125x353mm (600 x 600 DPI)
Figure 21. Histograms of \textit{S}-pick residuals obtained by the proposed workflow on the real dataset. Symbols as indicated in Figure 10.

\(\mu = -1.24 \text{ ms}\)
\(\sigma = 6.66 \text{ ms}\)
Figure 22. a) S-pick residuals obtained by the proposed workflow on the real dataset. b) Incorrect picking on complex waveform. Colors as indicated in Figure ef{Fig13}.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>$P$ arrivals</th>
<th>$S$ arrivals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Picked</td>
<td>Initially unidentified</td>
</tr>
<tr>
<td>1</td>
<td>1779</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>986</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>761</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>P arrivals</td>
<td>S arrivals</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>Picked</td>
<td>Initially</td>
</tr>
<tr>
<td></td>
<td>unidentified</td>
<td>unidentified</td>
</tr>
<tr>
<td>720</td>
<td>5</td>
<td>95</td>
</tr>
</tbody>
</table>
DATA AND MATERIALS AVAILABILITY

Data associated with this research are available and can be accessed via the following URL: Note: A digital object identifier (DOI) linking to the data in a general or discipline-specific data repository is strongly preferred.