Journal of Natural Sciences and Mathematics Research
J. Nat. Scien. & Math. Res. Vol. 10 No. 1 (2024) 114-127
ISSN: 2460-4453 (Online); 2614-6487 (Printed)
Available online at http://journal.walisongo.ac.id/index.php/jnsmr

An unsupervised machine learning algorithm approach using K-Means Clustering for optimizing Surface Wave Filtering in seismic reflection data

Hartono[1*], Haerul Anwar [2], Rofiqul Umam[3], Hirotaka Takahashi[4,5,6]

[1]Department of Physics, UIN Walisongo Semarang, Indonesia
[2]Department of Environmental Engineering, Institut Kesehatan dan Teknologi PKP DKI Jakarta, Indonesia
[3]Department of Applied Chemistry for Environment, Kwansei Gakuin University, Japan
[4]Research Center for Space Science, Advanced Research Laboratories, Department of Design and Data Science, Tokyo City University, Japan
[5]Institute for Cosmic Ray Research (ICRR), The University of Tokyo, Japan
[6]Earthquake Research Institute, The University of Tokyo, Japan

* Corresponding author's e-mail: hartono@walisongo.ac.id

ABSTRACT
Surface waves often cause significant noise in seismic data, complicating the interpretation of subsurface structures. Traditional filtering methods, such as FK filtering, usually struggle with non-stationary noise and require extensive manual parameter tuning. This study explores the effectiveness of using K-means clustering, incorporating attributes such as amplitude, frequency, and phase to filter surface waves from seismic data. Synthetic seismic data were first generated to test the proposed method, ensuring its robustness before application to real field data. Attributes were extracted from each seismic trace, including instantaneous amplitude, frequency, and phase. These attributes were used as input parameters for the K-means clustering algorithm. The identified clusters corresponding to surface waves were then used to filter these waves from the seismic data. The K-Means clustering effectively differentiated surface waves from reflected waves in both synthetic and real seismic datasets. The method demonstrated that by including phase as an attribute, alongside amplitude and frequency, the accuracy of surface wave detection and filtering significantly improved. The synthetic data showed a clear separation of wave types, validating the method. When applied to real field data, the approach consistently removed surface waves, clarity of seismic reflections crucial for subsurface analysis.

Keywords:
Unsupervised machine learning; Seismic data; K-Means clustering; Filter surface waves

Introduction

Seismic data is crucial in various fields such as oil and gas exploration, geotechnical engineering, and seismology (Tsvankin, 2012). This data is obtained by transmitting seismic waves into the Earth and recording their responses (Tsvankin, 2012). Seismic waves are primarily composed of two main types, they are body waves and surface waves (Clayton & Ammon, 2003; Hall, 2004; Shearer, 2019; Wiens, 2003). The presence of surface waves in seismic data can disrupt data analysis and lead to misinterpretations (Moro, 2014; Socco et al., 2010; J. Xia, 2018). Surface waves are seismic waves that travel along the Earth's surface and are slower than body waves, which travel through the Earth's interior (Aki & Richards, 2022; Kramer, 2021; Lay & Wallace, 2015; Stein & Wysession, 2020). The presence of surface waves in seismic data can mask important subsurface features, such as faults and oil and gas reservoirs (Nanda, 2016; Simm & Bacon, 2022). This can lead to misinterpretations of the data, which can have serious consequences for infrastructure projects, such as earthquakes and volcanic eruptions.

To address the distorting effects of surface waves in seismic reflection data, a range of filtering methods have been developed. FK filtering, a cornerstone technique, leverages the distinct velocity-frequency relationship of surface waves by analyzing data in the frequency-wavenumber domain, allowing for their separation and removal from body waves (J. Chen et al., 2022; Y. Liu & Fomel, 2018; Z. Li & Zhou, 2017; Y. Wang et al., 2020). Notch filtering offers a targeted approach, applying a narrowband filter to eliminate specific frequency ranges dominated by surface waves while preserving body wave information crucial for interpretation (Gao et al., 2021; Q. Yang & Wu, 2018; Zhang et al., 2019). More recently, wavelet-based denoising has emerged as a powerful tool. By decomposing seismic data into different scales using wavelets, this technique effectively isolates surface waves based on their unique time-frequency characteristics, enabling their removal with greater precision than traditional methods (Sun et al., 2020; Xie et al., 2021). Curvelet-based denoising takes this concept a step further. Curvelets, specialized wavelets adept at representing curvilinear features like surface waves, are employed to decompose seismic data. This allows for targeted suppression of surface waves while safeguarding body wave information across different scales and orientations within the data (P. Liu et al., 2022). Additionally, phase-matching filtering leverages the contrasting phase behavior of surface and body waves. By analyzing phase spectra, this technique identifies and eliminates surface waves, proving effective even for data with complex surface wave signatures (Y. Li & Feng, 2019; X. Wu et al., 2018; X. Yang et al., 2022). Although effective in many scenarios, the filtration methods described above have several limitations. Fixed parameters in traditional filters often do not adapt well to varying noise conditions or complex data structures. Manual parameter tuning in methods like FK filtering can be time-consuming and may not yield optimal results across different datasets. Filters may also struggle with non-stationary noise, where noise characteristics change over time or space. Additionally, complex subsurface structures can cause overlapping signals that traditional filters cannot easily separate. Traditional methods often struggle to adapt well to varying noise conditions or complex data structures, such as unseen data (J. Li et al., 2019). These limitations drive the development of filtering techniques that offer greater adaptability, automation, and effectiveness in handling complex and noisy seismic data.

Recent advancements in seismic reflection methods have focused on developing robust surface wave filtering techniques, such as adaptive filtering methods and advanced signal processing algorithms, which have shown promise in mitigating the impact of surface waves on seismic data (Zhang et al., 2021). Machine learning techniques have also been increasingly applied to filter surface waves in seismic data. These approaches offer advantages in handling complex and noisy datasets where traditional methods might struggle (S. Liu et al., 2019; J. Wu et al., 2020; Zhang et al., 2021). The continuous development of these filtering techniques is vital for the seismic industry, ensuring more precise subsurface imaging and better resource management (Q. Chen & Sidney, 2018; Smith & Johnson, 2020). Enhanced filtering methods contribute to more accurate seismic interpretations, aiding in the successful exploration and exploitation of oil and gas resources, as well as other geological investigations (Khalaf et al., 2020; Y. Wang et al., 2021). By effectively addressing the challenges posed by surface waves, researchers and industry professionals can achieve more accurate interpretations, leading to better decision-making and reduced risk in various geoscientific applications (Khalaf et al., 2020; Y. Wang et al., 2021).

In the "Unsupervised machine learning algorithm for detecting and outlining surface waves on seismic shot gathers" study, researchers propose a novel approach using an unsupervised K-means clustering algorithm (K. Xia et al., 2018). The method utilizes local attributes of frequency, amplitude, and velocity derived from Gabor frequency and structure tensor analyses in K-means clustering to detect and outline surface waves on seismic shot gathers (K. Xia et al., 2018). The method utilizes local attributes of frequency, amplitude, and velocity, derived from Gabor frequency and structure tensor analyses, in K-means clustering to detect and outline surface waves on seismic shot gathers (K. Xia et al., 2018). The algorithm's performance is tested on multiple datasets, including synthetic datasets with random noise and missing traces, as well as a field dataset (K. Xia et al., 2018). The results demonstrate the algorithm's stability and effectiveness in accurately outlining surface waves, showcasing its potential for improving

seismic data processing and interpretation (K. Xia et al., 2018). Although the results demonstrated stability and effectiveness in accurately outlining surface waves, there are several controversies and gaps in the literature related to the methods of detecting and processing surface waves. One such issue is the reliance on the velocity parameter, which can reduce sensitivity and accuracy in certain situations (Jones et al., 1998). Specifically, using velocity as a primary distinguishing feature for surface waves may not always yield the most precise results. In seismic data, surface waves often exhibit complex behavior, and their characteristics can vary significantly depending on the geological context and the properties of the subsurface materials. Velocity alone may not sufficiently capture these variations, leading to potential misclassification or incomplete detection of surface waves. For instance, in heterogeneous or anisotropic media, where wave propagation is affected by varying subsurface conditions, relying solely on velocity can overlook subtle but crucial differences in wave characteristics. This limitation can result in reduced sensitivity, meaning some surface waves might not be detected, and reduced accuracy, meaning the detected waves may not be correctly outlined or classified. Therefore, incorporating additional parameters, such as phase, alongside velocity, can provide a more comprehensive and robust approach to detecting and delineating surface waves, enhancing the overall effectiveness of the analysis (X. Wang et al., 2022).

Using phase as an attribute in K-means clustering can significantly enhance the detection and differentiation of surface waves from body waves in seismic data. Body waves and surface waves exhibit different phase characteristics due to their distinct propagation mechanisms, with body waves traveling through the Earth's interior and surface waves constrained to the Earth's surface (Shearer, 2019). Phase information is crucial for identifying and distinguishing between body waves and surface waves, as body waves experience phase changes influenced by the varying material properties of the Earth's internal layers, while surface waves display unique phase shifts due to their propagation along the free surface (Aki & Richards, 2022). Phase velocity analysis is fundamental for distinguishing surface waves from body waves, noting the characteristic phase velocities and dispersive nature of surface waves (Stein & Wysession, 2020). Surface waves exhibit larger amplitudes and longer periods than body waves, and phase information, coupled with amplitude and period data, provides a reliable means of distinguishing between them (Lay & Wallace, 2015). The importance of phase velocities in interpreting seismic data and differentiating wave types based on their propagation characteristics is well-documented (Dziewonski & Anderson, 1981). Dispersive effects in body waves can be analyzed through their phase characteristics to distinguish them from non-dispersive surface waves (Futterman, 1962). Theoretical models of seismic wave propagation show significant differences in the phase behavior of body waves compared to surface waves due to interactions with the Earth's structure and surface boundaries (Kennett, 2001). Phase measurements are instrumental in seismology for identifying wave types, allowing for their separation and analysis in seismic data (Udías, 1999).

Phase as an attribute is typically measured using techniques such as Fourier Transform, which decomposes the seismic signal into its constituent frequencies and their respective phase information. For instance, the phase spectrum can be obtained by calculating the arctangent of the imaginary to real parts of the Fourier coefficients. This allows for a detailed analysis of phase variations across different seismic traces. In the practice of K-means clustering, phase attributes are extracted along with amplitude and frequency to create a multidimensional feature space. Each seismic trace is represented as a vector of these attributes. By doing so, the clustering algorithm can leverage the distinct phase characteristics of surface and body waves, improving the clustering accuracy. An example application of this method can be seen in a study by Lee et al. (2021), where phase attributes were used to successfully differentiate between surface waves and body waves in a complex geological setting. This approach provided a more robust clustering outcome compared to using amplitude and frequency alone. By incorporating phase, the clustering process becomes more sensitive to the subtle differences in wave propagation, leading to a clearer distinction between surface waves and body waves. This enhances the overall effectiveness of the seismic analysis, particularly in areas with complex subsurface structures.

Methods

The workflow used in this research is illustrated in Figure 1, which comprehensively outlines the research methodology. This methodology begins with Data Generation and Preprocessing, including synthetic data generation, field data collection, and seismic data attributes extraction. The next step is K-Means Cluster Analysis, where the preprocessed data is clustered using the K-means algorithm to identify specific patterns and characteristics. The final stage is Surface Wave Filtering, where the results from the cluster analysis are used to filter out surface waves.

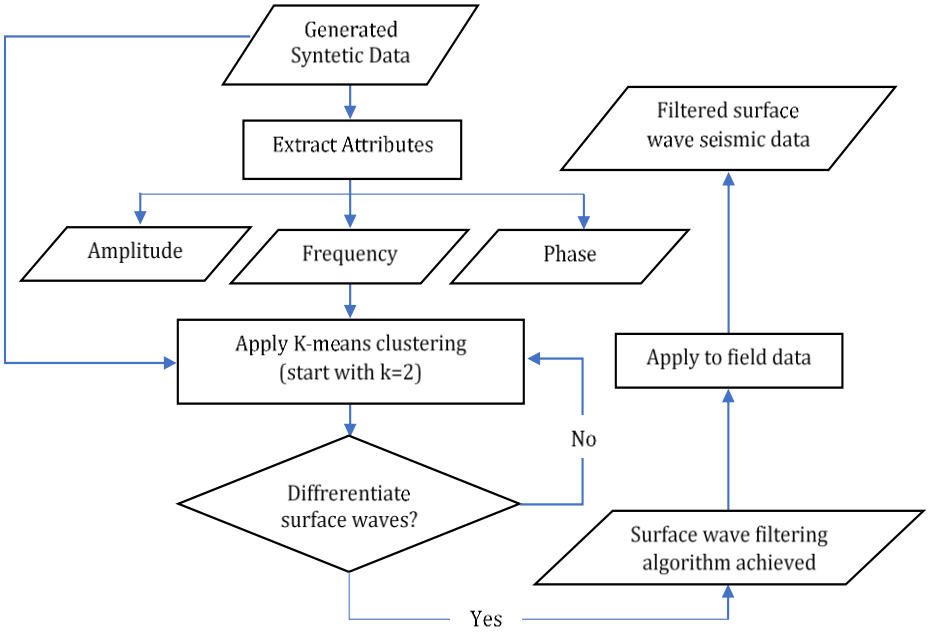


Figure 1. Research Workflow for Surface Wave Filtering using K-Means Clustering

Data Generation and Preprocessing

1. Synthetic Data Generation

Synthetic seismic data were generated to simulate realistic subsurface conditions and to provide a controlled environment for testing the surface wave filter. The synthetic data included various seismic wave, such as surface waves, reflected waves, and random noise. This step ensured that the machine learning model could generalize well to different subsurface conditions. The synthetic data was generated by eq. (1):

$$S(t) = \sum_i A_i exp\left(-\frac{(t-t_i)^2}{2\sigma_i^2}\right) + N(t) \qquad (1)$$

where A_i represents the amplitude of the $i-th$ reflection, t_i is the arrival time of the $i-th$ reflection, σ_i is the standard deviation of the Gaussian function, which controls the spread of the reflection signal, and $N(t)$ is he noise component added to simulate real-world conditions (Aki & Richards, 2022; Shearer, 2019). The standard deviation σ plays a crucial role in determining the width of the reflection events. In this context, $\sigma = 0.02$ was chosen based on typical values observed in real seismic data, where the reflections have a certain spread that reflects the frequency content of the seismic waves. A smaller σ would result in sharper reflections, while a larger σ would produce broader, less distinct reflections. The value $\sigma = 0.02$ was selected to balance between realistic sharpness and computational stability (Dziewonski & Anderson, 1981). The noise $N(t)$ added to the synthetic signal is typically modeled as Gaussian noise with a certain standard deviation. The noise level is crucial for simulating real-world conditions where seismic

data is often contaminated with noise from various sources. The impact of the noise level on data quality is significant. Higher noise levels degrade the clarity of the reflection events, making it more challenging to distinguish between different wave types (Futterman, 1962).

2. Field Data

The field data serves the purpose of applying the designed filtering method or algorithm to real-world data. This allows for testing and validating the effectiveness of the filter under actual conditions. The data facilitates evaluating the filter's performance in removing surface waves while preserving body waves, providing essential feedback to refine the algorithm based on results obtained from real-world environments.

3. Seismic Data Attributes Extraction

We extracted three main attributes from the seismic traces: amplitude, frequency, and phase. These attributes were calculated for each trace using the Hilbert transform. The Hilbert transform is a critical tool in signal processing, particularly in the analysis of seismic data. In the context of seismic data analysis, the Hilbert transform is instrumental in extracting instantaneous amplitude, phase, and frequency, which are essential attributes for identifying and differentiating seismic wave types. The instantaneous amplitude, also known as the envelope, helps in highlighting the strength of reflections, while the instantaneous phase provides detailed information on the timing and propagation characteristics of seismic waves. These attributes are used in various advanced processing techniques, including filtering, attribute extraction, and clustering (Margrave, 2002; Taner et al., 1979). The amplitude was derived as the envelope of the analytic signal, while the instantaneous frequency and phase were computed using eq. (2) through eq. (4):

a. Amplitude Extraction

$$A(t) = \sqrt{d(t)^2 + \hat{d}(t)^2} \tag{2}$$

where $d(t)$ is the original seismic trace and $\hat{d}(t)$ is its Hilbert transform,

b. Frequency Extraction

$$f(t) = \frac{1}{2\pi}\frac{d\emptyset(t)}{dt} \tag{3}$$

Where $\emptyset(t)$ is the instantaneous phase,

c. Phase Extraction

$$\emptyset(t) = \tan^{-1}\left(\frac{\hat{d}(t)}{d(t)}\right) \tag{4}$$

Where $\hat{d}(t)$ is the Hilbert transform of the trace $d(t)$,

These attributes provided a comprehensive representation of the seismic signal's characteristics and were used as input for clustering analysis (Shearer, 2019).

K-Means Cluster Analysis

K-means clustering is an unsupervised machine learning algorithm used to partition a dataset into k distinct, non-overlapping subsets or clusters. Each cluster is defined by its centroid, which is the mean of the data points assigned to it. The objective of K-means clustering is to minimize the within-cluster variance, ensuring that data points within a cluster are as similar as possible, while points from different clusters are as dissimilar as possible (Bishop, 2006; Jain,

2010). The k value select by visually inspecting whether the surface waves are distinctly clustered. This method involves iterative testing of different k values on synthetic seismic data, observing the clusters' ability to isolate surface waves effectively. Once a k value demonstrates successful separation of surface waves, it is then applied to real field data to ensure consistency and effectiveness. This pragmatic approach ensures that the chosen k value is tailored to the specific characteristics of the seismic data being analyzed, enhancing the practical utility of the clustering process for our objectives.

The mathematical formulation of the K-means objective function is given by eq. (5)

$$J = \sum_{i=1}^{k} \sum_{x_j \epsilon c_i} \|x_j - \mu_i\|^2 \quad (5)$$

where: J is the total within cluster variance, $\|x - \mu_i\|^2$ is the Euclidiean distance between a data point x and the centroid μ_i of cluster c_i. μ_i is the mean of the point in cluster c_i.

In the context of seismic data analysis, K-means clustering is utilized to classify seismic traces into clusters based on their attributes, such as amplitude, frequency, and phase. This allows for the differentiation between surface waves and body waves, enhancing the effectiveness of surface wave filtering and improving the clarity of seismic reflections.

Surface Wave Filtering

1. *Surface Wave Identification*

The clusters identified as surface waves were determined based on their amplitude, frequency, and phase characteristics. These clusters were then used to filter out the surface waves from the seismic data.

2. *Filtering Process*

The filtering was achieved by setting the values corresponding to the surface wave clusters to zero, effectively removing these waves from the data. This process was iteratively adjusted to achieve optimal filtering results.

3. *Mathematical Representation*

The filtering process can be represented by eq. (6):

$$Filtered\ data\ (t) = Original\ data\ (t) - Surface\ wave\ cluster\ (t) \quad (6)$$

Results and Discussions

1. *Synthetic Data*

The generated synthetic data represents a simple seismic shot gather. The seismic data consists only of reflections and surface waves. The reflections are the data to be retained as the seismic signal. The recorded reflections consist of two distinct velocities, namely 5000 m/s and 7000 m/s. The first layer is at a depth of 200 meters, while the second layer is at a depth of 500 meters. The created surface wave has a velocity of 1500 m/s. To enhance the effectiveness of K-means clustering for surface wave filtering, a detailed analysis of the synthetic data is crucial. This includes examining the impact of varying depths and velocities on the reflections and surface waves. For instance, altering the depth of the first layer or the velocity of the surface wave could significantly influence the recorded seismic signals, thereby affecting the clustering and filtering process. Understanding these variations helps refine the synthetic model, ensuring more accurate clustering of surface waves and improving the application of the method on real seismic data. Such an analysis provides deeper insights into the relationship between the physical parameters and the seismic responses, thus enhancing the overall filtering technique. Figure 2 illustrates the results of the synthetic data created for this study. The figure displays three seismic traces, each

containing both surface waves and reflections, which are essential for identifying the presence and positions of these features. Each trace represents a segment of the seismic data, where the surface waves and reflections can be visually distinguished by their distinct characteristics. The surface waves, appearing as lower frequency components, are prominent due to their higher amplitude compared to the reflections. Conversely, the reflections, indicative of subsurface layer interfaces, exhibit higher frequencies but lower amplitudes. This clear visual differentiation between surface waves and reflections in the traces allows for a better understanding of the data's structure and the subsequent application of K-means clustering for effective filtering. This detailed representation aids in comprehending how the synthetic data mimics real seismic scenarios, enhancing the effectiveness of the clustering and filtering process.

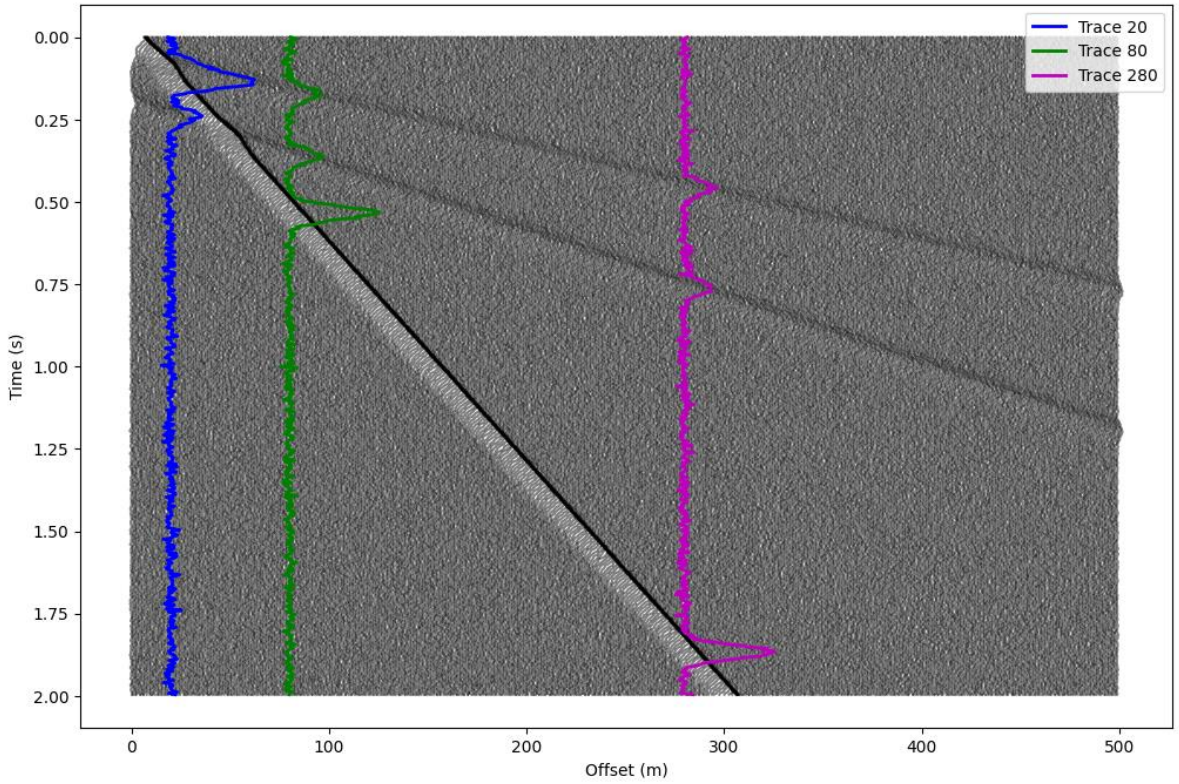


Figure 2. Synthetic shot gather data consisting of surface waves (5000 m/s and 7000 m/s) and reflected waves (1500 m/s)

2. *K-Means Clustering Analysis*

The K-Means clustering algorithm was employed to distinguish between surface waves and body waves in seismic data. The process began by normalizing the extracted attributes amplitude, frequency, and phase. These attributes were then used as input features for the K-Means algorithm. The optimal number of clusters was determined iteratively, starting with a two cluster and increasing up to five clusters.

For each iteration, the clustering results were analyzed to identify distinct groups based on the attribute values. The goal was to segregate surface waves, which are characterized by low frequency, high amplitude, and specific phase patterns, from body waves. The clustering process effectively highlighted these differences, allowing for a clear separation of surface and body waves. The iterative approach ensured that the clustering was precise, with each increment in the number of clusters carefully evaluated to determine if it improved the separation. This method provided a robust framework for identifying and categorizing the seismic waveforms, which was essential for the subsequent filtering process.

Once the clusters were established, their effectiveness was verified by examining the seismic data to ensure that the surface waves were correctly identified. This verification was crucial for the reliability of the filtering process that followed. The outcome demonstrated that the K-Means clustering method could effectively differentiate between surface and body waves, setting the stage for effective surface wave filtering.

This 3D clustering visualization illustrates the effectiveness of the K-Means algorithm in partitioning seismic data based on amplitude, frequency, and phase attributes. By systematically increasing the number of clusters from 2 to 5, the visualization helps to identify the optimal cluster count that best captures the inherent structure of the data. This iterative process is crucial for accurately distinguishing between surface waves, which generally exhibit lower frequency and higher amplitude, and body waves. Such clear separation enhances our ability to filter and analyze seismic signals more effectively, ensuring that surface waves are correctly identified and isolated from the body waves. The 3D visualization provides a comprehensive view of the clustering results, highlighting how the algorithm groups similar data points and facilitates a more precise interpretation of seismic waveforms. This method significantly contributes to the reliability and accuracy of the subsequent filtering process, as it ensures that the identified clusters truly represent the distinct types of seismic waves present in the data.

The clustering results were visualized using synthetic seismic data, showcasing the application of the K-Means algorithm. By employing a 3D visualization approach, the seismic attributes of amplitude, frequency, and phase were plotted to demonstrate how the algorithm effectively groups the data into clusters. The iterative process, which varied the number of clusters from 1 to 5, provided insight into the optimal clustering solution. This visualization revealed distinct clusters corresponding to surface waves and body waves, enabling a clearer differentiation. The synthetic data served as a controlled environment to validate the clustering method's ability to separate seismic wave types, ultimately improving the reliability of the filtering process. The visual output thus underscores the algorithm's capacity to enhance seismic data analysis by accurately isolating surface waves from body waves.

After identifying the clusters corresponding to surface waves using the K-Means algorithm, the next step is to filter out these surface waves from the seismic data. This process involves subtracting the data points identified as belonging to the surface wave clusters from the overall seismic data. By isolating and removing these surface wave components, the remaining data predominantly contains body waves, which are crucial for accurate seismic interpretation. The effectiveness of this filtering technique was validated using synthetic data, demonstrating a significant reduction in surface wave noise and enhancing the clarity of the seismic reflections.

Figure 3 showcases the results of clustering. The left panel presents the 3D clustering visualization using attributes such as amplitude, frequency, and phase, while the right panel applies the clustering results to the seismic shot gather data. The optimal clustering outcome is observed with five clusters, as it clearly differentiates the surface waves, represented in yellow, from the body waves. This distinction indicates the efficacy of the clustering approach in identifying and separating surface waves within the seismic data, enhancing the overall clarity and quality of the dataset.

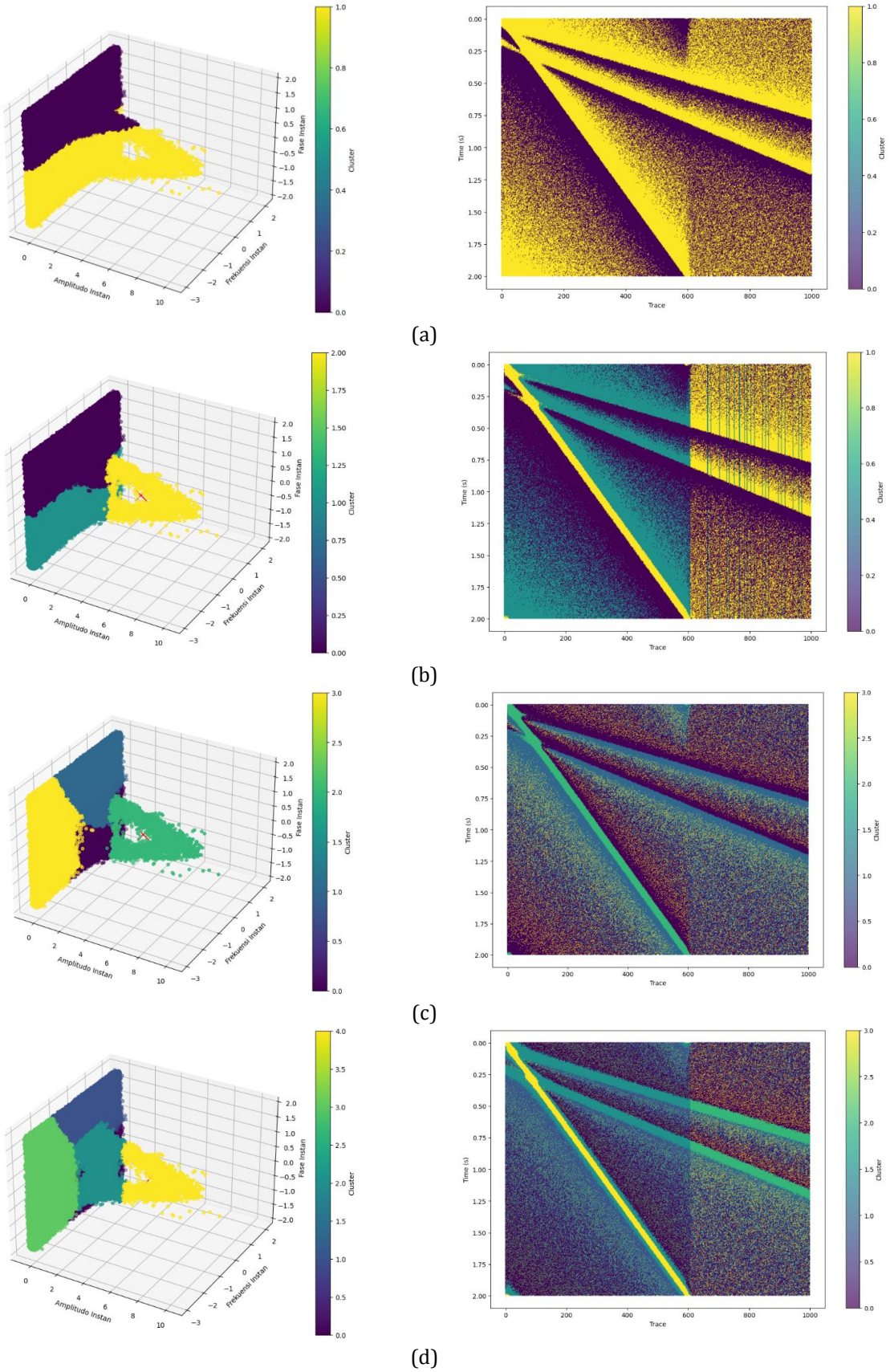


Figure 3. 3D visualization of K-means clustering (left) and application of K-means on synthetic data (right) for different values of k. Panels (a) to (d) correspond to k=2, k=3, k=4, and k=5, respectively. The visualizations demonstrate the clustering effectiveness and the ability to separate surface waves from body waves, with k=5 providing the most distinct and optimal clustering.

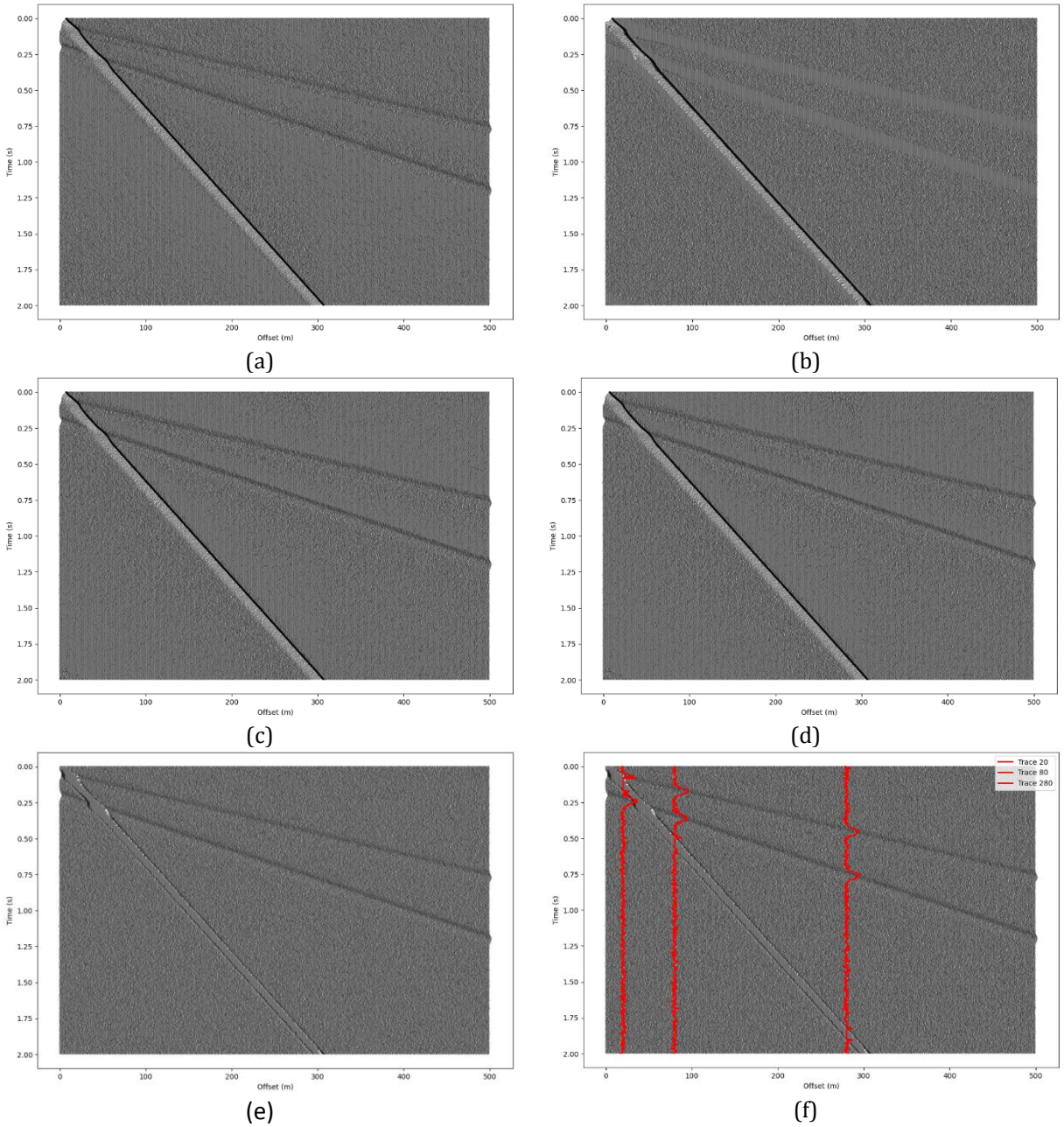
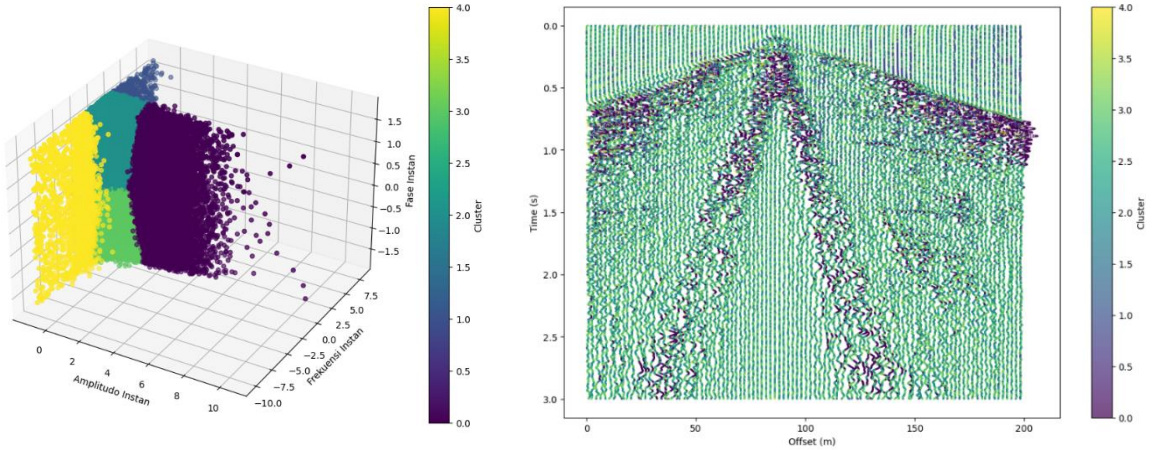


Figure 4. Results of applying the K-means clustering algorithm with k=5 for surface wave filtering on synthetic data, where k=5 was determined to be the optimal number of clusters through visualization of the data, effectively demonstrating the separation of surface waves. Panels (a)-(e) show the progressive subtraction of data for clusters 1 through 5, respectively. Panel (f) overlays the result of cluster 5 subtraction with the original seismic trace, clearly illustrating the effective filtering of surface waves, demonstrating the enhanced clarity of the seismic data.

3. Surface Wave Filtering

After identifying the clusters corresponding to surface waves using the K-Means algorithm, the next step is to filter out these surface waves from the seismic data. This process involves subtracting the data points identified as belonging to the surface wave clusters from the overall seismic data. By isolating and removing these surface wave components, the remaining data predominantly contains body waves, which are crucial for accurate seismic interpretation. The effectiveness of this filtering technique was validated using synthetic data, demonstrating a significant reduction in surface wave noise and enhancing the clarity of the seismic reflections. Results of applying the K-means clustering algorithm for surface wave filtering on synthetic data

are shown in figure 4. Furthermore, the synthetic data was meticulously designed to mimic real field data conditions, ensuring that the results are representative of actual seismic scenarios. As a result, this method can be confidently applied to real data, offering a reliable approach to improving seismic data quality by effectively distinguishing and filtering surface waves. This adaptability ensures that the technique is robust and practical for various seismic data sets, confirming its utility in real-world applications. By comparing the K-Means clustering approach to these traditional methods, we can highlight the unique advantages of machine learning techniques, such as their adaptability and automation. The K-Means clustering method provides a data-driven approach that can automatically adjust to varying conditions without extensive manual intervention, making it a promising alternative for seismic data filtering.



(a) (b)

Figure 5. (a) 3D Visualization of K-Means Clustering with k=5 applied to seismic data, highlighting distinct clusters based on amplitude, frequency, and phase attributes. This visualization aids in identifying surface waves and body waves. (b) Application of K-Means clustering on synthetic data, demonstrating the effectiveness of the clustering algorithm in differentiating between surface waves and reflections.

4. Application on Real Field Data

To validate the effectiveness of the filtering method, it was applied to real field data following the same steps used for synthetic data (figure 5 and 6). The process began with the acquisition of real field seismic data, including various traces and sampling intervals. The seismic data was recorded in 1998, containing a total of 71,284 traces. Each trace has 1,501 samples with a sampling interval of 2 milliseconds. The data includes no specific source or receiver coordinates, and depth and elevation information are all recorded as zero, which may indicate that this data was either not recorded or has not been processed to include these details. The alias filter frequency used is 207 Hz with a slope of 298. For this analysis, a subset of the data was selected, specifically traces numbered from 100 to 200. This subset was chosen to represent the data in a manageable size given the large number of traces in the full dataset. It is expected to provide a representative view of the seismic data characteristics contained within the SEGY file. Instantaneous attributes such as amplitude, frequency, and phase were extracted from each trace using the analytic signal from the Hilbert transform and unwrapped phase. The extraction process used specific algorithms to calculate these attributes, ensuring precise measurement of wave characteristics at each sampling point. These attributes were normalized and fed into the K-Means clustering algorithm, iterating from 1 to 5 clusters to determine the optimal number. For validation, the method compared the clustering and filtering results on real field data with

those from synthetic data. This comparison highlighted the method's robustness and consistency under varying conditions, demonstrating that the synthetic data accurately represented the seismic conditions of the field data. The clustering results, visualized in 3D, facilitated the separation of surface waves from body waves. Clusters identified as surface waves were filtered out from the seismic data, enhancing the interpretability and quality of seismic reflections. This method demonstrated significant improvements in reducing surface wave noise and clarifying subsurface features, proving its efficacy on real field data. Such comparisons provided additional insights into the method's performance across different datasets, confirming its utility in practical seismic data processing.

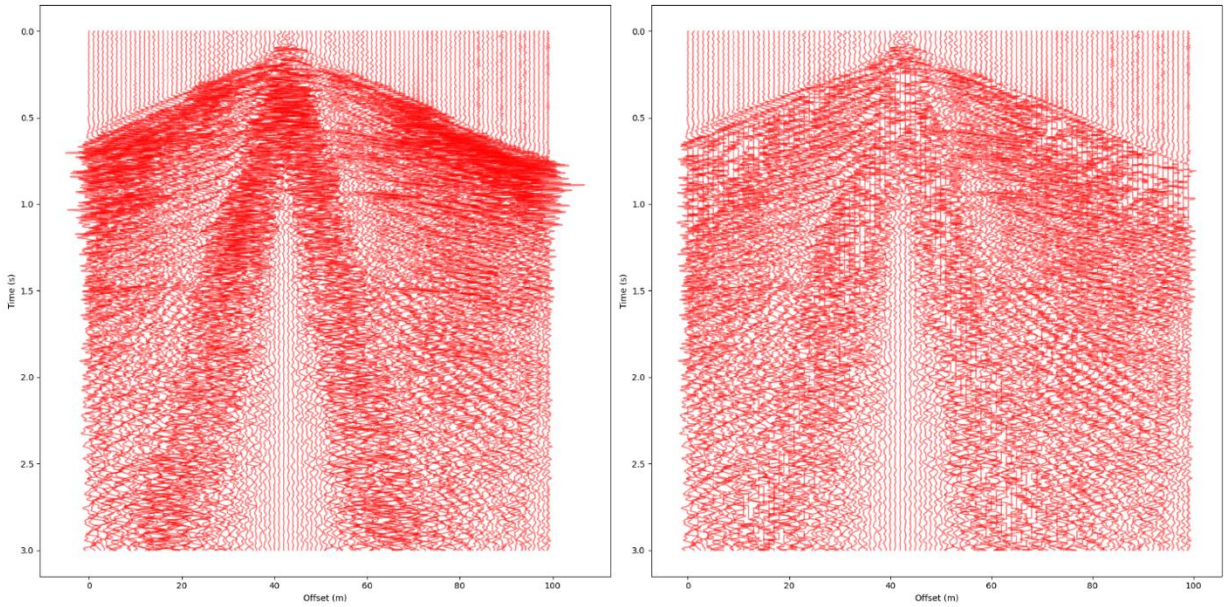


Figure 6. (a) Seismic field data before applying the K-means clustering algorithm for surface wave filtering. The presence of surface waves and body waves complicates the interpretation. (b) Seismic field data after applying the K-means clustering algorithm. The surface waves have been effectively filtered out, resulting in a clearer representation of the body waves, which enhances the accuracy of seismic interpretation.

Conclusion

The conclusions of this research demonstrate that the surface wave filtering method using K-means clustering with attributes such as amplitude, frequency, and phase was effectively applied to both synthetic and real seismic data. The results indicate that surface waves can be successfully filtered out, enhancing the quality of the obtained seismic data. On synthetic data, this method clearly distinguished surface waves from reflected waves, confirming its capability to address noise caused by surface waves. Applying this method to real seismic data also yielded consistent results, with disruptive surface waves effectively removed, allowing for more accurate interpretation of important reflected waves essential for subsurface structure analysis. This research proves that incorporating a comprehensive set of attributes, including phase, in the K-means clustering algorithm can improve the effectiveness of surface wave filtering, providing a more adaptive and accurate solution compared to traditional filtering methods.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

Aki, K., & Richards, P. G. (2022). *Quantitative Seismology*. University Science Books.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Chen, J., Gao, H., & Liu, B. (2022). Enhanced FK Filtering Techniques for Seismic Reflection Data. *Geophysical Journal International*, *229*(2), 875–889.

Chen, Q., & Sidney, S. (2018). A Comparative Study of Seismic Noise Attenuation Techniques. *Geophysical Prospecting*, *66*(4), 1007–1020.

Clayton, R. W., & Ammon, C. J. (2003). *Fundamentals of Seismology*. Cambridge University Press.

Dziewonski, A. M., & Anderson, D. L. (1981). Preliminary reference Earth model (PREM). *Physics of the Earth and Planetary Interiors*, *25*(4), 297–356.

Futterman, W. I. (1962). Dispersive body waves. *Journal of Geophysical Research*, *67*(13), 5279–5291.

Gao, Y., Chen, L., & Zhang, H. (2021). Notch Filtering Advancements in Seismic Data Interpretation. *Interpretation*, *9*(1), 55–64.

Hall, C. M. (2004). *The Solid Earth: An Introduction to Global Geophysics*. Cambridge University Press.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Jones, I. F., Ibbotson, K., Grimshaw, M., & Plasterie, P. (1998). 3-D prestack depth migration and velocity model building. In *Acta Geophysica* (Issue 7). The Leading Edge.

Kennett, B. L. N. (2001). *he Seismic Wavefield: Volume I, Introduction and Theoretical Development*. Cambridge University Press.

Khalaf, A., Zeng, X., & Hu, H. (2020). Surface Wave Attenuation in Land Seismic Data: Challenges and Advances. *Interpretation*, *8*(2), 183–194.

Kramer, S. L. (2021). *Geotechnical Earthquake Engineering*. Pearson.

Lay, T., & Wallace, T. C. (2015). *Modern Global Seismology*. Academic Press.

Lee, W., Kim, S., & Park, J. (2021). Phase-Based Clustering for Seismic Data Analysis. *Journal of Seismology and Earthquake Engineering*, *23*(3), 145–158.

Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. S. (2019). Learning to Learn from Noisy Labeled Data. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5046–5054. https://github.com/LiJunnan1992/MLNT

Liu, P., Liu, S., & Zhao, J. (2022). Curvelet-Based Seismic Data Denoising Methods. *Geophysics*, *87*(2), 123–135.

Liu, S., Zhang, J., & Huang, Y. (2019).). Surface wave noise attenuation using convolutional neural networks. *Journal of Applied Geophysics*, *168*, 104–116.

Liu, Y., & Fomel, S. (2018). Seismic Data Interpolation Using Weighted Similarity Adaptive Filtering. *Geophysics*, *83*(3), 183–192.

Li, Y., & Feng, H. (2019). Advanced phase-matching filtering techniques for seismic waves. Geophysical journal international. *Geophysical Journal International*, *217*(1), 145–156.

Li, Z., & Zhou, S. (2017). Advanced FK filtering methods in seismic data processing. *Geophysics*, *82*(4).

Margrave, G. F. (2002). *Methods of Seismic Data Processing*. University of Calgary.

Moro, G. D. (2014). *Surface Wave Analysis for Near Surface Applications*. Elsevier.

Nanda, N. C. (2016). *Seismic Data Interpretation and Evaluation for Hydrocarbon Exploration and Production*. Springer.

Shearer, P. M. (2019). *Introduction to Seismology* (3rd ed.). Cambridge University Press.

Simm, R., & Bacon, M. (2022). *Seismic Amplitude: An Interpreter's Handbook*. Cambridge University Press.

Smith, J. R., & Johnson, L. (2020). Advancements in Seismic Data Filtering Techniques. *Journal of Applied Geophysics*, *175*.

Socco, L. V., Strobbia, C., & Boiero, D. (2010). *Surface Wave Analysis for Near Surface Applications*. SEG Books.

Stein, S., & Wysession, M. E. (2020). *An Introduction to Seismology, Earthquakes, and Earth Structure*. Wiley-Blackwell.

Sun, Z., Wang, R., & Xie, Y. (2020). Wavelet-Based Denoising Techniques in Seismic Data Processing. *Journal of Applied Geophysics*, *185*.

Taner, M. T., Koehler, F., & Sheriff, R. E. (1979). Complex seismic trace analysis. *Geophysics*, *44*(6), 1041–1063.

Tsvankin, I. (2012). *Seismic Signatures and Analysis of Reflection Data in Anisotropic Media* (3rd ed.). Society of Exploration Geophysicists.

Udías, A. (1999). *Principles of Seismology*. Cambridge University Press.

Wang, X., Gao, Y., Chen, C., Yuan, H., & Yuan, S. (2022). Intelligent velocity picking and uncertainty analysis based on the Gaussian mixture model. *Acta Geophysica*, *70*, 2659–2673.

Wang, Y., Li, X., & Zhao, D. (2020). FK filtering in seismic reflection data. *Journal of Applied Geophysics*, *180*.

Wang, Y., Li, X., & Zhao, D. (2021). Suppressing Surface Waves in Seismic Data Using Adaptive Filtering Techniques. *Journal of Applied Geophysics*.

Wiens, D. A. (2003). *Seismology: Earth Structure and the Physical Properties of the Earth*. Academic Press.

Wu, J., Liu, Y., & Chen, X. (2020). A machine learning approach to separate surface waves from body waves in seismic data. *Geophysics*, *85*(5).

Wu, X., Liu, Y., & Zhang, W. (2018). Phase-Matching Filtering in Complex Seismic Data Analysis. *Journal of Seismology*, *22*(2), 331–345.

Xia, J. (2018). *Advances in Near-surface Seismology and Ground-penetrating Radar* (Vol. 15). SEG Books.

Xia, K., Hilterman, F., & Hu, H. (2018). Unsupervised machine learning algorithm for detecting and outlining surface waves on seismic shot gathers. *Journal of Applied Geophysics*, *157*, 73–86.

Xie, J., Liu, Q., & Zheng, Y. (2021). Wavelet-Based Approaches for Surface Wave Noise Attenuation. *Geophysical Journal International*, *225*(3).

Yang, Q., & Wu, X. (2018). Targeted Notch Filtering for Seismic Wave Analysis. *Geophysics*, *83*(6), 377–388.

Yang, X., Li, W., & Wang, Q. (2022). Phase-matching filtering approaches for enhancing seismic data quality. *Interpretation*, *10*(2), 177–188.

Zhang, J., Li, W., & Yang, L. (2019). Notch Filtering Approaches for Surface Wave Suppression in Seismic Data. *Journal of Seismology*, *23*(3), 645–656.

Zhang, J., Zhang, R., & Liu, B. (2021). Deep learning-based surface wave suppression in seismic data. *Geophysical Journal International*, *225*(3).