

**FIRST ARRIVALS TRAVELTIME PICKING USING
UNSUPERVISED MACHINE LEARNING (DBSCAN)
AND SUPER-VIRTUAL REFRACTION
INTERFEROMETRY TECHNIQUE**

BY

MUHAMMAD AWAIS

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

MS GEOPHYSICS

NOVEMBER 2022

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

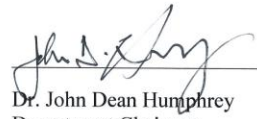
DHAHRAN - 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **MUHAMMAD AWAIS** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN GEOPHYSICS**.



Dr. Sherif Mahmoud Hanafy
(Advisor)



Dr. John Dean Humphrey
Department Chairman



Prof. Abdullatif Al-Shuhail
(Member)



Prof. Suliman Saleh Al-Homidan
Dean of Graduate Studies



Dr. Umair Bin Waheed
(Member)

10/12/2022
Date



© Muhammad Awais

2022

Dedication

To my father Mr. Muhammad Waris and my mother, wife, son, siblings, and sister
in-laws for their support, efforts, love, and prayers.

ACKNOWLEDGMENTS

I begin with the name of Allah Almighty, the most gracious and merciful. I am grateful to Allah (SWT) who provided me with this opportunity and gave me the courage and focused mind to study at Geoscience Department and complete my master's degree successfully. I am thankful to the Saudi Arabia Higher Education Commission and King Fahd University of Petroleum and Minerals for providing the academic platform and financial support during my master's studies.

I like to express my deepest gratitude and appreciation to my advisor Dr. Sherif Mahmoud Hanafy for his constant guidance, motivation, support, cooperation, and supervision during my graduate research work. His understanding, expertise, and patience helped me a lot during my studies. I feel proud of doing research tasks under his supervision. I also express my deep gratitude and thankfulness to Prof. Abdullatif Al-Shuhail and Dr. Umair Bin Waheed as my thesis committee members for their experienced support, contribution, encouragement, highly valuable suggestions, and guidance during my research work.

I would like to pay thanks to individuals for their support: Mr. Saifullah Shafique, and Mr. Nawaf Al-Omari who helped me with MATLAB programming learning. I can't forget the consistent guidance of Mr. Muhmmad Naeem who not only assisted in my research work but also provided me with constant motivation, support, and help. His good spirit helped me overcome setbacks and stay focused on my studies. I also would like to pay thanks to all faculty, staff, and students of the Geosciences Department at King Fahad University of Petroleum & Minerals for their support through my pleasant learning experience.

I am very grateful to my family: parents, wife, brothers, and sisters for their cooperation, encouragement, support, patience, understanding, and prayers during my academic journey. Especially my brothers Mr. Muhammad Waqas, Mr. Muhammad Usman Waris, Mr. Muhammad Arslan, and Mr. Muhammad Saif for their guidance, support, cooperation, motivation, and affection. I am also very grateful for the support and cooperation of my uncle (Mr. Muhammad Hammed), aunt, and cousins.

TABLE OF CONTENTS

AKNOWLEDGMENTS	V
TABLE OF CONTENTS.....	VII
LIST OF TABLES.....	X
LIST OF FIGURES.....	XI
LIST OF ABBREVIATIONS.....	XV
ABSTRACT	XVI
ملخص الرسالة	XVIII
CHAPTER 1 INTRODUCTION.....	1
1.1 General	1
1.2 Types of first break picking methods.....	2
1.2.1 Manual method	2
1.2.2 Automatic methods	3
1.3 Automatic first break picking approach for low SNR data	4
1.4 Objectives	6
1.5 Justification of the study	7
1.6 Benefits of the study.....	7
1.7 Structure of the thesis.....	8
CHAPTER 2 LITERATURE REVIEW	10
2.1 Previous studies related to first break traveltime picking	10
2.2 Previous studies related to DBSCAN applications.....	13
2.3 Previous studies related to Super-virtual Refraction Interferometry (SVI) applications	15

CHAPTER 3 BACKGROUND THEORY	16
3.1 Machine learning	16
3.2 DBSCAN technique	18
3.3 Seismic Interferometry (SI)	21
3.4 Super-virtual Refraction Interferometry (SVI)	24
CHAPTER 4 WORKING PRINCIPLES OF DBSCAN AND SVI	27
4.1 DBSCAN technique	27
4.2 Super-virtual Refraction Interferometry (SVI)	28
CHAPTER 5 WORKFLOW AND METHODOLOGY	30
5.1 DBSCAN algorithm	30
5.1.1 Selecting a sliding window	32
5.1.2 Applying Linear Moveout Correction (LMO) and stacking	32
5.2 SVI algorithm	35
5.3 Proposed methodology.....	37
CHAPTER 6 DATASET INFORMATION AND PREPROCESSING	38
6.1 Synthetic data	38
6.2 Ithra field data	40
6.3 Yilmaz field data	41
6.4 Preprocessing.....	43
6.5 Parameters determination for synthetic and field data	45
6.5.1 Parameters determination for synthetic data	45
6.5.2 Parameters determination for field data	46
6.6 BandPass filter	47
CHAPTER 7 RESULTS AND DISCUSSION.....	49
7.1 Results	49
7.1.1 Synthetic results with no noise	49

7.1.2	Synthetic results with high-Level random noise and no restricted scanning window	53
7.1.3	Synthetic results with high-Level random noise and restricted scanning window	56
7.1.4	Synthetic results	56
7.2	Ithra field data results.....	60
7.3	Yilmaz field data results	64
7.4	Discussion	66
CHAPTER 8 FUTURE PROSPECTS AND CONCLUSION		71
REFERENCES.....		73
VITAE		79

LIST OF TABLES

Table 1:	Parameters selected for different datasets. These are the same for the	
	initial/final first break.	47

LIST OF FIGURES

Figure 1.1: Schematic representation of the proposed methodology.....	6
Figure 3.1: DBSCAN clustering result with epsilon = 1 and minimum points = 10, Blue color circle points shows cluster 1, magenta color circle shows bluster 2, whereas red circles are either outlier/noise points, the Figure generated by using MATLAB, version 2020a.	21
Figure 3.2: Reflection seismograms (zero and shifted zero offsets) over a sandstone lens with complex near-surface/overburden (a), representing static removal using Seismic Interferometry (b), (Schuster, 2009).	23
Figure 3.3: Schematic diagram for generating super-virtual refraction traces using SVI (Hanafy and Al-Hagan, 2012).....	26
Figure 4.1: Schematic diagram for generating super-virtual refraction traces using SVI (Hanafy and Al-Hagan, 2012).....	29
Figure 5.1: DBSCAN clustering in the presence of data and noise. (a) shows the selected window with 5 trace and 100 sample numbers, (b) is the stacked LMO image, (c) shows the stacked LMO image after applying the threshold, and (d) represents DBSCAN cluster formation.	34
Figure 5.2: DBSCAN shows no or many clusters in the presence of noise only. (a) shows the selected window with 5 trace and 100 sample numbers, (b) is the stacked LMO image, (c) shows the stacked LMO image after applying the threshold, and there is no DBSCAN cluster formation.	34
Figure 5.3: Flowchart indicating the different steps involved in SVI.....	36

Figure 5.4: Research methodology accurate first break picking automatic using DBSCAN and SVL.....	37
Figure 6.1: Synthetic velocity model showing a two-layer velocity model with first and second layer's velocities of 500 m/s and 1500 m/s, respectively.	38
Figure 6.2: Shot gather 1 from the synthetic data.	39
Figure 6.3: Shot gather 60 from the synthetic data.	39
Figure 6.4: Shot gather 121 from the synthetic data.	40
Figure 6.5: Ithra field's data shot gather 1, end spread.	41
Figure 6.6: Ithra field's data shot gather 100, split spread.	41
Figure 6.7: Yilmaz shot gather 6 recorded in the Far East.	42
Figure 6.8: Yilmaz shot gather 25 recorded in Alberta.....	42
Figure 6.9: Synthetic shot gather 60 showing small amplitude first arrivals, the area is indicated by black two head arrows.	43
Figure 6.10: Synthetic shot gather 60 after 20 times increased amplitude as compared to the original data.	44
Figure 6.11: Two black arrows show two different slopes of the refraction layers.....	47
Figure 6.12: Shot gather 25 (a) before band pass filter (b) after bandpass filter.	48
Figure 7.1: DBSCAN pickings (blue circles) compared with manual (magenta cross marker) for synthetic shot gather 60 without noise. Black arrows show mismatched locations of DBSCAN picking with manual picking.	50
Figure 7.2: a) Non-linear slope trend in the window, b) Stacked LMO image shows scattered energy c) Stacked LMO image after applying the threshold, d) In the end, it gives deviated first break picking value.....	51

Figure 7.3: a) Increasing linear slope trend in the window, b) Stacked LMO image c) Stacked LMO image after applying the threshold, d) In the end, it gives the wrong first break picking value.	52
Figure 7.4: DBSCAN picking without restricted scanning window and extrapolation with 50% random noise.	54
Figure 7.5: DBSCAN picking with restricted scanning window and extrapolation with 50% random noise for synthetic shot gather 1.	55
Figure 7.6: a) and f) show shot gather 29 and 69 with added random noises, magenta and blue lines showing manual and DBSCAN initial guess, respectively, b) and g) show the muting around the expected location of the first arrivals using initial guess from DBSCAN, c) and h) showing the enhanced SNR after SVI, d) and i) show the manual picking (magenta line) match with the DBSCAN picking after SVI (blue line), e) and j) presenting the difference between manual and DBSCAN first arrivals traveltimes before applying SVI (red line) and after applying SVI (blue line), respectively. The dashed green lines display the threshold values. The red arrow shows the shot location.	58
Figure 7.7: Histograms comparing traveltimes differences between manual and final DBSCAN's automatic first break's picking after SVI for all shot gathers. Green dashed lines show the threshold values of ± 8.33 msec.	59
Figure 7.8: a) and g) show Ithra field data shot gathers 25 and 100 with manual pickings (magenta line), respectively, b) and h) showing the DBSCAN initial guess (blue line), c) and i) show the muting around the expected	

location of the first arrivals using the initial guess, d) and j) showing the enhanced SNR after SVI, e) and k) show the re-DBSCAN pickings after SVI (blue line), f) and l) representing the matching of manual picking (magenta line) with DBSCAN after SVI (blue line). The black arrow shows enhanced picking than the manual. The red arrow shows the shot location. 62

Figure 7.9: a) and b) show the differences between manual and DBSCAN first arrivals traveltimes pickings for Ithra shot gather 25 and 100 before applying SVI (red line) and after applying SVI (blue line), respectively. ... 63

Figure 7.10: Histograms comparing traveltimes differences between manual and final DBSCAN's automatic first break's picking after SVI for all shot gathers. Green dashed lines show the threshold values of ± 8.33 msec. 64

Figure 7.11: Matching of initial guess DBSCAN picking (blue line) with Manual (magenta line), (a) for shot gather 6, and (b) for 25. For shot gather 25, the last 4 traces are muted as they were just noises..... 65

Figure 7.12: a) and b) showing the differences between manual and initial guess of DBSCAN first arrivals traveltimes pickings for shot gather 25 and 100, respectively..... 65

LIST OF ABBREVIATIONS

DBSCAN	:	Density-Based Spatial Clustering Application with Noise
SVI	:	Super-Virtual refraction Interferometry
SNR	:	Signal-to-Noise Ratio
minPts	:	Minimum number of Points

ABSTRACT

Full Name : [Muhammad Awais]
Thesis Title : [First Arrivals Traveltime Picking Using Unsupervised Machine Learning (DBSCAN) And Super-Virtual Refraction Interferometry Technique]
Major Field : [Geophysics]
Date of Degree : [November 2022]

An accurate and trustworthy first break picking plays an important role in static correction calculations, velocity analysis, and deconvolution. First break traveltimes picking accuracy ensures correct and reliable seismic data processing results. Many methods are available to perform the first break pickings automatically. However, most of the methods do not give robust results when the signal-to-noise ratio (SNR) of the data is low. In this study, I used an unsupervised machine learning technique called density-based spatial clustering application with noise (DBSCAN) along with super-virtual refraction interferometry (SVI) to automatically pick the first arrivals' traveltimes. In the case of noisy data, using only DBSCAN for first break picking shows poor accuracy. To overcome this drawback, SVI is used to enhance the SNR of the raw data, and then DBSCAN is applied to the enhanced data for accurate picking. I first use DBSCAN to define an initial muting window for the SVI, then, SVI is used to enhance the SNR of the first arrivals, and finally, I re-apply DBSCAN to pick the final first arrivals' traveltimes. The proposed approach is tested on synthetic and field data, where the manual and the cross-correlation pickings are compared to the DBSCAN's automatic pickings. For the synthetic data, the automatic first arrival's traveltimes picked by DBSCAN before using

SVI for all shots was poor, whereas, after applying SVI, it improved drastically, and almost 98.4% of the DBSCAN's automatic first arrivals traveltimes pickings compared with the manual fall within the allowable error of ± 8.33 msec. For all shot gathers of Ithra field data, results show that DBSCAN's automatic first break's traveltime picking before using SVI was poor again, whereas, after applying SVI, 86.4% of the pickings compared with raw manual pickings were lying within the allowable error of ± 8.33 msec. I also test a trace-by-trace version of the proposed approach on publicly available shot records (Yilmaz shot gathers 6 and 25) to benchmark my technique for automatic first break traveltimes picking's accuracy with other established methods. Results show that the performance of my methodology is better than the other methods by picking 91% and 99% traveltimes within an allowable error of ± 20 msec from the manuals for shots 6 and 25, respectively.

ملخص الرسالة

الاسم الكامل: محمد اويس

عنوان الرسالة: الوافدون الأولون انتقاء وقت السفر باستخدام التعلم الآلي غير الخاضع للإشراف (DBSCAN) وتقنية الانكسار الافتراضي الفائق

التخصص: الجيوفيزياء

تاريخ الدرجة العلمية: (نوفمبر) 2022

يلعب انتقاء الاستراحة الأول الدقيق والجدير بالثقة دورًا مهمًا في حسابات التصحيح الثابت ، وتحليل السرعة ، والتفكيك. تضمن دقة اختيار أوقات السفر في الاستراحة الأولى نتائج معالجة البيانات الزلزالية الصحيحة والموثوقة. تتوفر العديد من الطرق لإجراء عمليات التقاط الفاصل الأول تلقائيًا. ومع ذلك ، فإن معظم الطرق لا تعطي نتائج قوية عندما تكون نسبة الإشارة إلى الضوضاء (SNR) للبيانات منخفضة. في هذه الدراسة ، استخدمت أسلوبًا غير خاضع للإشراف للتعلم الآلي يسمى تطبيق المجموعات المكانية القائمة على الكثافة مع الضوضاء (DBSCAN) جنبًا إلى جنب مع قياس الانكسار الافتراضي الفائق (SVI) لاختيار أوقات سفر الوافدين الأوائل تلقائيًا. في حالة البيانات المشوشة ، فإن استخدام DBSCAN فقط للتقاط الفاصل الأول يظهر دقة ضعيفة. للتغلب على هذا العيب ، يتم استخدام SVI لتحسين نسبة الإشارة إلى الضوضاء (SNR) للبيانات الأولية ، ثم يتم تطبيق DBSCAN على البيانات المحسنة لانتقاء دقيق. أستخدم DBSCAN أولاً لتحديد نافذة كنم الصوت الأولية لـ SVI ، ثم يتم استخدام SVI لتحسين نسبة الإشارة إلى الضوضاء (SNR) للقادمين الأوائل ، وأخيرًا ، أعيد تطبيق DBSCAN لاختيار أوقات سفر الوافدين النهائيين. يتم اختبار النهج المقترح على البيانات التركيبية والميدانية ، حيث تتم مقارنة اختيارات الدليل والترابط المتبادل مع عمليات الانتقاء التلقائية لـ DBSCAN. بالنسبة إلى البيانات التركيبية ، كانت أوقات السفر التلقائية للوصول الأول التي اختارتها DBSCAN قبل استخدام SVI لجميع اللقطات ضعيفة ، بينما بعد تطبيق SVI ، تحسنت بشكل كبير ، وحوالي 98.4 ٪ من عمليات الوصول التلقائي لأوقات السفر في DBSCAN مقارنةً بالدليل تقع ضمن الخطأ المسموح به ± 8.33 ميلي ثانية. بالنسبة لجميع مجموعات اللقطات لبيانات حقل Ithra ، تُظهر النتائج أن التقاط وقت السفر التلقائي للقطر الأول من DBSCAN قبل استخدام SVI كان ضعيفًا مرة أخرى ، بينما ، بعد تطبيق SVI ، كانت 86.4 ٪ من اللقطات مقارنة بالاختيارات اليدوية الأولية تقع ضمن الخطأ

المسموح به وهو ± 8.33 مللي ثانية. أختبر أيضًا نسخة التتبع عن طريق التتبع للنهج المقترح على سجلات اللقطات المتاحة للجمهور (تجمع 6 Yilmaz Shot و 25) لقياس تقنيتي في الاستراحة الأوتوماتيكية لدقة أوقات السفر مع الطرق الأخرى المعمول بها. تظهر النتائج أن أداء منهجيتي أفضل من الطرق الأخرى من خلال اختيار 91% و 99% من أوقات السفر ضمن خطأ مسموح به قدره ± 20 مللي ثانية من كتيبات اللقطات 6 و 25 ، على التوالي.

CHAPTER 1

INTRODUCTION

1.1 GENERAL

The first arrival or first break can be defined as the onset arrival on a shot gather that is comprised of the refracted energy associated with the bottom of the weathering layer (Yilmaz, 2001). First break picking is usually a common step in seismic reflection/refraction data processing (Al-Anezi et al., 2012; Al-Shuhail and Al-Shaibani, 2013; Almadani et al., 2015). First arrival traveltime is the first stage in transforming the seismic data into a geological model. Among various steps to follow in conventional seismic data processing, one of the important steps is the static correction. The correctness of the first breaks traveltime picking is essential to improve the effectiveness of reflection/refraction-based methods to calculate the static correction for processing seismic data. There are different challenges that could be faced while doing the calculations for the static correction. Examples include areas with complex near-surface geology, data with a poor SNR, or weak source wavelet penetration, (Yilmaz, 2001; Al-Mashhor et al., 2019). Ground condition and coupling at source and receiver locations can also influence (Cox, 1999). So, first arrivals picking is a crucial step. Besides the static correction, this step affects the subsequent steps such as velocity analysis (Alkhalifah and Tsvankin, 1995), and deconvolution (Bell and Sejnowski, 1995). Further, there are many renowned seismic data processing programs available such as Promax

(Fajaryanti et al., 2018), Focus and Grisys (F.-Y. WANG et al., 2015). To process the seismic data, these systems involve a very important step which is to pick the first arrival's traveltimes. Moreover, the amount of seismic data that has been collected in the field has also increased (Huahui et al., 2018). Therefore, automatic first arrival picking methods are required urgently as the volume of seismic data is continuously increasing (Akram and Eaton, 2016).

1.2 TYPES OF FIRST BREAK PICKING METHODS

1.2.1 MANUAL METHOD

Mainly, first arrival traveltime picking techniques are divided into manual and automatic (Cox, 1999). Manual first break detection is done by discernible examination of the amplitudes by an adept processor. However, there are many difficulties in manual picking, particularly for data polluted with noise. These difficulties are related to many effects, for instance, the experience of a human operator, the precision and ability of the operator's eyes, the scale of the imaging and data amplitude, the accuracy of the picking cursor axes, and the signal-to-noise ratio. Moreover, this method requires a lot of time once it deals with large datasets or are unfeasible when there comes a requirement for real-time seismic data processing (Senkaya and Karsli, 2014). Further, if the dataset is noisy, the manual inspection becomes harder. Consequently, the picking process may take between 20-30% of the whole processing time, particularly if the dataset is big (Sabbione and Velis, 2010).

1.2.2 AUTOMATIC METHODS

Nevertheless, many different algorithms and techniques using machine learning are developed to perform first arrivals traveltime picking automatically. Many scholars have proposed various approaches for automatic first arrivals detection. Gelchinsky and Shtivelman (1983) employed the joint use of correlation technique with statistical technique to restrict the results of first break picking. The energy ratio method was first presented by Coppens (1985). A fractal-based algorithm for identifying first arrivals on seismic traces was given by Boschetti et al. (1996). Mousa and Al-Shuhail (2012) presented a first break enhancement technique by benefiting from the τ -p transform on energy-ratio to pick the first break in seismic shot gathers. Senkaya and Karsli (2014) made use of the cross-correlation method to prepare the data for first-break recognition. Chi-Durán et al. (2017) have developed new algorithms benefiting from the Fourier transform and the fractal dimension technique to determine the arrival traveltimes automatically. Despite certain advantages, these methods suffer from a number of challenges on field datasets. The primary problem is that these methods do not give robust results in cases of low energy levels or low SNR (Pan et al., 2019).

Many advanced automatic first break traveltime picking algorithms have been developed based on convolutional neural networks (CNNs) (Yuan et al., 2018; Ross et al., 2018; Hollander et al., 2018). Nevertheless, methods based on CNNs present many difficulties (Saad and Chen, 2021) besides being computationally expensive, time-consuming, intensive data labeling, and slow convergence (Benediktsson et al., 1990; Han et al., 2021).

1.3 AUTOMATIC FIRST BREAK PICKING APPROACH FOR LOW SNR DATA

In this research, I proposed the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) along with the super-virtual Refraction Interferometry technique to pick the traveltimes of the first arrivals. DBSCAN is an unsupervised machine learning method that uses clustering of the data to pick the traveltimes of the first breaks. A variety of clustering techniques are effectively applied for multiple purposes such as outlier recognition, medical imaging, pattern recognition, and gene sequencing, (Gan et al., 2007), automatic interpretation of geophysical data sections (Sabor et al., 2021), automatic velocity picking (bin Waheed et al., 2019), seismic zones identification (Karri et al., 2018), anomaly detection (Çelik et al., 2011; Wibisono et al., 2021), etc.

Using DBSCAN to pick first arrival traveltimes was proposed by K. Lu and Feng (2018) and tested by Abdullah (2020) during his thesis work. They achieved accurate results for noiseless and very low-noise datasets. K. Lu and Feng (2018) used DBSCAN and convolution neural network (CNN) to find the rough location of the first breaks, which is, then, used to mute the data for the SVI application, and then pick the first breaks using other standard automatic or manual picking methods.

So, to improve the results of DBSCAN picking, one way to increase the performance is to increase the SNR. Therefore, before applying DBSCAN for the final picking, I plan to use super-virtual refraction interferometry (SVI) to increase the SNR of the far offset arrivals. SVI technique has been successfully applied to different seismic data generated by land and marine surveys to enhance the far offset arrivals by different researchers

(Hanafy et al., 2011; Hanafy and Al-Hagan, 2012; Bharadwaj et al., 2012; Bharadwaj et al., 2013; Vaezi and DeMeersman, 2014; Li et al., 2018). The general idea of SVI is to construct and stack super-virtual refractions to improve the SNR of coherent refractions. By enhancing the SNR, the DBSCAN job becomes much easier to identify accurate first arrivals traveltime. In my research work, I follow K. Lu and Feng's (2018) approach, however, I first apply only DBSCAN to get an initial guess of the first break picking area, then use this guess to define the muting window followed by re-applying DBSCAN after the SVI step to automatically pick the first arrival. This results in accurate results even for data with low SNR.

The schematic idea is to first add random noise into the synthetic data whereas the field data has noise already. Apply the DBSCAN to get an initial guess for the picking. This guess will give an idea to mute around the actual first breaks. This muted data will be fed to SVI to enhance the SNR. This enhanced SNR data will be given to the DBSCAN again to make the accurate picking. The schematic idea diagram is shown below in Figure 1.1.

After the successful application of DBSCAN with SVI on the synthetic data, field data has been collected and this process is repeated to get the final output with accurate traveltimes picking. To benchmark the results with other automatic traveltime picking methods, the results are compared with the cross-correlation technique developed by Senkaya and Karsli (2014), further, the methodology has been tested on two publicly available datasets.

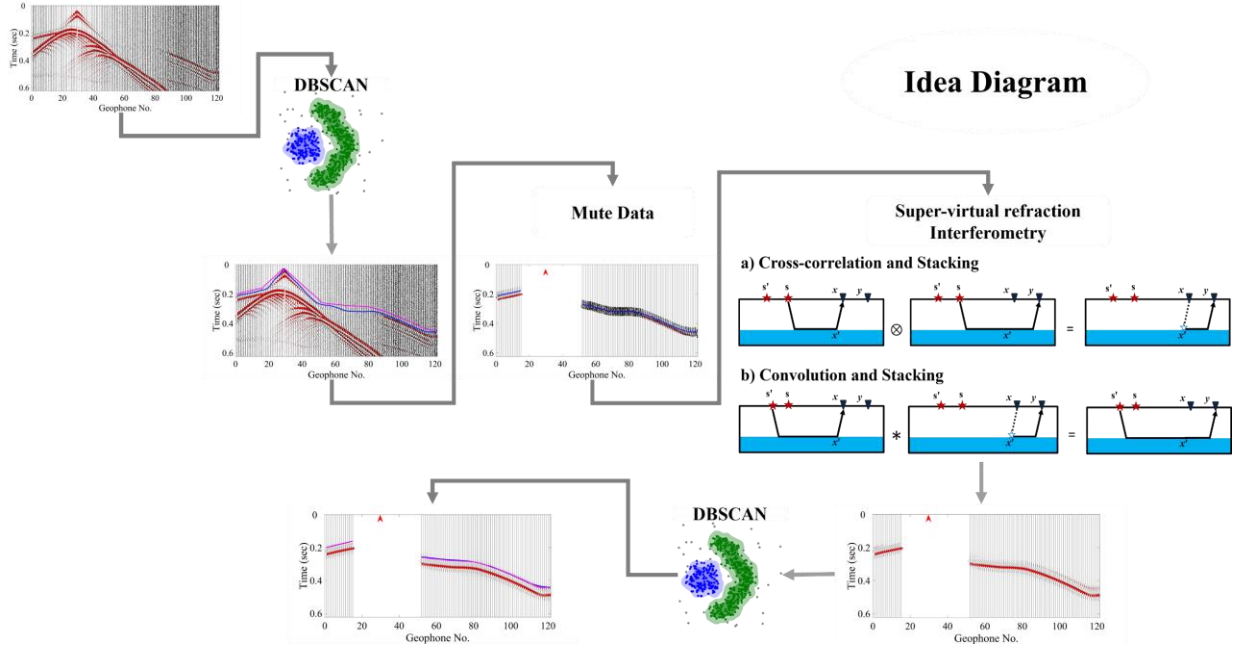


Figure 1.1: Schematic representation of the proposed methodology.

1.4 OBJECTIVES

1. Enhance the signal-to-noise ratio of the first arrivals using super-virtual refraction interferometry (SVI)
2. Accurate first arrival travel time picking using unsupervised machine learning technique (DBSCAN)
3. Compare the results before and after applying SVI and picked by DBSCAN both for synthetic and real datasets

1.5 JUSTIFICATION OF THE STUDY

There are two justifications for this study:

1. DBSCAN has been employed by K. Lu and Feng (2018) to locate the muting window for the SVI application. It has also been employed to pick the first break traveltime in noise-free or low-noise data, however, it has not been used to pick the accurate first break traveltime in the presence of high-level random noise.
2. Many automatic methods available in the literature e.g., Yuan et al. (2018) and Yalcinoglu and Stotter (2018) are using supervised machine learning techniques to pick the first break traveltimes. Whereas, in this study, an unsupervised machine learning technique has been employed that is much more flexible and doesn't require any labeling of the data like the supervised machine learning techniques.

1.6 BENEFITS OF THE STUDY

As we know that automatic first break picking is very important for seismic data processing e.g., static correction, velocity analysis, and deconvolution. Further, many systems available in the market require first break picking for seismic data processing. Many methods available in the literature struggle to pick the accurate first break in the presence of low SNR data. In this study, DBSCAN along with SVI has been used to pick the first break accurately in the presence of low SNR data. So, the benefit of this study would be to pick efficiently the accurate first break traveltimes automatically when the data has poor SNR. The success of this method is very promising as the DBSCAN is an

unsupervised machine learning technique that does not require any data labeling like supervised machine learning techniques.

1.7 STRUCTURE OF THE THESIS

This thesis is divided into eight chapters as follows:

Chapter 1 starts with an introduction to first break picking and its importance in seismic data processing. Then it introduces different first arrival traveltime picking methods and their limitations. Next, the approach is briefly explained with an idea diagram. Lastly, study objectives, justification, and benefits are described.

Chapter 2 mentions the literature review of first arrival traveltime picking methods from past to recent studies. It covers semiautomatic and automatic methods being used for first break picking and outlines their room for improvement. Further, it describes the applications of the DBSCAN clustering algorithm, Seismic Interferometry (SI), and SVI techniques.

Chapter 3 contains the background theory of machine learning, DBSCAN, Seismic Interferometry, and SVI techniques.

Chapter 4 describes the working principle of both techniques (DBSCAN and SVI) used throughout this research work.

Chapter 5 contains three sections. The first section starts with the DSBCAN method working steps and explains how to select different parameters e.g., sliding windows and clustering parameters to start the methodology. The second section shows the working steps of the SVI technique. The last part displays the complete methodology employed in this study to pick the first break traveltime.

Chapter 6 starts with synthetic, Ithra, and Yilmaz field datasets information. Then it tells the preprocessing steps that need to be followed. For example, it describes the guidelines for the selection of different parameters needed to be determined for the DBSCAN application and the band pass filter application for the Ithra field data before the SVI application.

Chapter 7 concludes the results and discussions of the proposed method and suggests some recommendations for future studies.

Chapter 8 states the prospects of the study and concludes the research work.

CHAPTER 2

LITERATURE REVIEW

2.1 PREVIOUS STUDIES RELATED TO FIRST BREAK TRAVELTIME PICKING

In an early stage, Peraldi and Clement (1972) presented a cross-correlation of neighboring traces to find the delay times between the first break and the peak of the signal. The method was based on the similarity of the neighboring traces. But it did not prove efficient in two cases: when the variation in the waveform of adjacent seismic traces is large and when there are absent traces. By using the difference in time between the first break and the peak of the wave, Hatherly (1982) proposed a method by using the least-squares approach with the first kick idea to identify first arrivals. Nevertheless, the method gives poor results when the resemblance between the traces is poor aside from the reality that the algorithm comprises quite complex calculations. Gelchinsky and Shtivelman (1983) employed the joint use of correlation technique with statistical technique to restrict the results of first break pickings. The method gives good pickings like those comparable with manual first break pickings when the SNR is high. This technique produces large errors when SNR is low. The energy ratio method was first proposed by Coppens (1985). He made the use of energy of the trace within one cycle and of the whole-time window for determining the first break. This approach relatively works better in the presence of noise, therefore, many scholars directed further research depending on it. Baer and Kradolfer (1987) benefit from an envelope function and non-

linear amplifier to detect phase pick automatically. Boschetti et al. (1996) proposed a fractal-based algorithm to find the first break on seismic traces by examining the change in fractal dimension along the trace. The fractal dimension is a measure of the “roughness” of the trace. The divider method was used to calculate this dimension. This method relies on a difference in the dimension which occurs when the trace transforms from noise to noise plus signal (i.e., first break). The nature of this difference varies from trace to trace, but a traceable change is always found to occur. But this method works slower in the presence of significant noise; therefore, it is intended to use for high SNR. Liao et al. (2011) presented another automatic first break picking approach built on time-frequency analysis (spectral decomposition). Mousa et al. (2011) introduced a new method that is built on the digital color-segmentation of energy ratios of refracted seismic data. In general, this technique works better where the first two energy peaks are separated from each other for far-offset traces. This method’s results are comparable to Coppen’s energy ratio method. Mousa and Al-Shuhail (2012) proposed a first arrival improvement method using the τ -p transform on energy-ratio seismic shot gathers. Senkaya and Karsli (2014) used the cross-correlation technique for automatic first arrival detection. This method is comprised of three steps. A source wavelet is firstly determined according to the noise content of the data, secondly, it is cross-correlated trace by trace with the data. Then, the first arrival area on the data is highlighted manually and finally maximal points of cross-correlated data which indicate the first break times are picked. Chi-Durán et al. (2017) have carried out new strategies built on the Fourier transform and the fractal technique to find the arrival times automatically. Despite certain advantages, these methods still show some application challenges. The primary problem is, these

methods give less robust results in cases of low energy levels or low SNR (Pan et al., 2019).

An automatic first break detection technique was proposed by Pan et al. (2019) by using an image connectivity algorithm and multiple time windows. This technique efficiently solves issues related to data with low SNR and low energy level first arrivals. But selecting the connected parameter is a crucial step in the presented approach. In some situations, it is hard to find this parameter by experience, which unavoidably leads to repeated cycles. Thus, it needs to build another algorithm that can adjust the parameters automatically.

Recently, picking first arrival traveltimes focused on using different methods of machine learning such as applying a method based on reinforcement learning (Ma and Luo, 2018), and Yalcinoglu and Stotter (2018) applied support vector machine. Many advanced automatic first break traveltime picking algorithms have been developed based on convolutional neural networks (CNNs) (Yuan et al., 2018; Ross et al., 2018; Hollander et al., 2018). However, these methods based on CNNs have many difficulties, e.g., firstly, in the testing phase, CNN may not identify the rotated/shifted objects which is thought as a disadvantage while dealing with 2-D images (Zeiler and Fergus, 2014; B. Jia and Huang, 2020), secondly, the use of pooling later in the deeper layers reduction in data size that results in information loss (Zeiler and Fergus, 2014; Sabour et al., 2017), thirdly, it can be deceived in case if the object is mislocated that may require data augmentation (Zeiler and Fergus, 2014; Sabour et al., 2017; B. Jia and Huang, 2020). Further, methods based on CNNs are computationally expensive, time-consuming, slow convergent, and require

intensive data labeling (Benediktsson et al., 1990; Han et al., 2021; Saad and Chen, 2021).

Al-Mashhor et al. (2019) developed a method for first break picking of first arrivals based on the maximum difference between adjacent points of the trace envelope (MDPE). The presented approach automatically selects the traveltime of the first arrivals depending on the maximum difference between neighboring elements of the envelope of a complex seismic trace. This method starts by calculating the Hilbert transform of the complex seismic trace to get an envelope, converting its measurements to decibel values, then computing the moving median of this envelop by using a sliding window, followed by determining the differences of the moving median points and then mark the first break where this difference is maximum. The approach was evaluated on synthetic gathers with very low random noise addition, and the results were compared to manual picking. The results showed that the method detected the picks with more than 88% accuracy. Further, this method was assessed on real seismic data with relatively high SNR data and found that the method estimated the picks with an accuracy of greater than 85%. However, MDPE requires more testing with various data sets to understand its limitations.

2.2 PREVIOUS STUDIES RELATED TO DBSCAN APPLICATIONS

DBSCAN is a clustering technique. A variety of clustering techniques are effectively applied for multiple purposes in different fields such as outlier recognition, medical imaging, pattern recognition, gene sequencing, etc. (Gan et al., 2007). It has also been

applied in the automatic interpretation of geophysical data sections (Sabor et al., 2021), automatic velocity picking (bin Waheed et al., 2019), seismic zones identification (Karri et al., 2018), and anomaly detection (Çelik et al., 2011; Wibisono et al., 2021).

Using DBSCAN to pick the first arrival traveltimes was proposed by K. Lu and Feng (2018) and tested by Abdullah (2020), they achieved accurate results for noiseless and low-noise data sets. The algorithm picks the first arrival's traveltime very accurately in case of noiseless or very low noise data but suffers in case of data with a high level of random noise. To cope with noise in the data, K. Lu and Feng (2018) used DBSCAN and CNN to find the rough location of the first breaks, which is, then, used to mute the data for the super-virtual refraction interferometry (SVI) application to enhance the far-offset head wave arrivals, and then pick the first breaks by using automatic/manual picking methods.

In my thesis work, I follow Lu and Feng's (2018) approach, however, to define the muting window for SVI, I use only DBSCAN with a restricted scanning window and extrapolate for the missing picks approach to get an initial guess and then re-apply DBSCAN after the SVI step to, automatically, pick the first arrivals.

Abdullah (2020) tested the DBSCAN algorithm during his thesis work only for noiseless and very low-level random noise data and achieved accurate results. The reason for the poor results in the case of high-level random noise is that the first arrivals get masked by the random noise in the data and once the results, after applying linear moveout correction, are stacked, the signals are canceled out and the algorithm misses the accurate picks of the first arrivals. Secondly, using DBSCAN, he was scanning the whole data

from the first sample to the last sample by selecting a sample and trace window in search of the first break, and this caused DBSCAN to pick many outliers due to the random noises present in the data.

2.3 PREVIOUS STUDIES RELATED TO SUPER-VIRTUAL REFRACTION INTERFEROMETRY (SVI) APPLICATIONS

To increase the SNR of the far-offset arrivals, super-virtual refraction interferometry is a robust and effective SNR enhancer for far-offset head waves (Bharadwaj et al., 2011; Bharadwaj et al., 2012). The general idea of SVI is to construct and stack super-virtual refractions to improve the SNR of coherent refractions. Numerous scholars used the SVI theory successfully to enhance the far offset arrivals' SNR. For example, Hanafy and Al-Hagan (2012) and Al-Shuhail (2015) used SVI to increase the SNR of the far offsets to improve picking accuracy. In a similar vein, Bharadwaj et al. (2013) used SVI to enhance the first break SNR of ocean bottom seismic data, which significantly increased the number of reliable pickable first break traces. Hanafy et al. (2011) employed SVI on land seismic data to achieve a similar output as Bharadwaj et al. (2013). Additional successful applications on different seismic data to enhance the far-offset arrivals include Vaezi and DeMeersman (2014), and Li et al. (2018). By enhancing the SNR, the DBSCAN job becomes much easier to identify first arrivals traveltimes accurately.

CHAPTER 3

BACKGROUND THEORY

3.1 MACHINE LEARNING

Machine learning is a subset of Artificial intelligence that dates to the 1950s, and started as the idea that computers can mimic the human brain (Samuel, 1959). The goal of machine learning is to write an algorithm or code to solve any human brain tasks. Machine learning consists of methods/techniques which learn from the labeled or unlabeled data to enhance decisions making across a multitude of tasks (Qadrouh et al., 2019). It is being used to solve industrial or real-life problems more efficiently. Different examples include image/object/speech/pattern/handwriting/outlier/anomaly identification, photo tagging, web page ranking, anti-spam filtering, medical imaging, gene sequencing, etc., (Gan et al., 2007; Çelik et al., 2011; Schmidhuber, 2015). Moreover, there are numerous applications of machine learning techniques in the fields of petroleum and geophysics such as petrophysical properties e.g., porosity and permeability estimation (Hamada and Elshafei, 2010; Ahmadi and Chen, 2019), reservoir characterization (Helmy et al., 2010), seismic velocity estimation from wireline data (Aleari, 2015), seismic faults identification (Araya-Polo et al., 2017), interpolation of seismic data (Y. Jia and Ma, 2017), wavelet estimation (L.-X. Wang and Mendel, 1992), velocity analysis (Calderón-Macías et al., 1998), automatic velocity picking of seismic data (bin Waheed et al., 2019), inversion and interpretation problems (Kim and Nakata, 2018; P. Lu et al., 2018), etc. Machine learning is typically categorized into two subtypes:

a) Supervised machine learning methods

This is probably the most common type of machine learning being used. In the supervised techniques, we make the computer how to learn. In these algorithms, we know what the input is and what would be the output. We train our algorithm on this data and evaluate it by asking a similar question on which the algorithm was trained and getting the correct answer. To get an accurate output, the algorithm task is to find patterns and develop a mathematical model between the input and output to predict the answer accurately. Supervised machine learning can be divided into two types: classification, and regression. If the problem has a discrete-valued answer (e.g., 0 & 1), it's classification learning while if the output is a continues value, it will be regression learning. Classification learning problem examples include classifying, if an email is spam or not, and if cancer is deadly or benign. We may have two or more classes of answers rather than just two. Regression learning problems examples include predicting a house price, and airline tickets fair.

b) Unsupervised machine learning methods

In these approaches, we have the input data, but we do not know what the answers/outputs are. In other words, unlike the supervised algorithm, these methods do not require any training/labeled dataset. The main task of these algorithms is to find patterns or structures based on similarities or features and form grouping/clustering inside the data. Therefore, unsupervised machine-learning techniques are much more flexible than supervised machine-learning techniques. Unsupervised machine learning is also of two types: clustering and non-clustering (Kaufman and Rousseeuw, 2009). Clustering learning uses partitioning or hierarchical approaches to form clusters in the data. K-

means and DBSCAN algorithms are examples of partitioning and hierarchical approaches, respectively. In clustering learning, the algorithm doesn't change or alter the data to form clusters, but it finds some similarity patterns and groups them (e.g., gene sequencing). Whereas, in non-clustering learning, algorithms changes or transforms the data. A cocktail party is an example of non-clustering learning because it alters the data to identify different voice patterns in the party.

3.2 DBSCAN TECHNIQUE

Partitioning and hierarchical are the two main types of clustering algorithms (Kaufman and Rousseeuw, 2009). The first one divides the data into k clusters, where k is an input parameter that is not easily determined for different applications. K-mean and k-medoid approaches are examples of partitioning clustering techniques. The second one does not require k as an input parameter and initiates a hierarchical break down of the dataset into smaller sets iteratively until every subset contains only one object. The hierarchical fragmentation is described by a dendrogram which can be constructed either by an agglomerative or divisive approach (Ester et al., 1996). DBSCAN is an example of a hierarchical clustering technique. DBSCAN involves two main input parameters to perform clustering of the data. These two parameters are ϵ (epsilon) and the minimum number of objects/points (minPts). The epsilon is a distance parameter and is used to define the neighborhood of a point. All the points which are inside the epsilon distance are called as the neighborhood points. The minPts is stated as the lowest number of the data points in the neighborhood of that point to form a cluster. Using these two parameters, DBSCAN forms clusters within a dataset. A high density of the

points/objects in an area means a cluster otherwise noise or outlier (bin Waheed et al., 2019). DBSCAN has many advantages over other clustering methods e.g., k-mean, or mean-shift clustering. This method does not require the input of the number of clusters to form within a dataset like k-mean, and it identifies outliers as noise unlike the k-mean and the mean-shift clustering algorithm where each point will belong to a cluster no matters if it's a very different point. In addition, arbitrary sized and shaped (e.g., circular, rectangular, and ellipse etc.) clusters can be identified successfully using DBSCAN.

DBSCAN starts by determining the data points into three types to start the clustering procedure.

- i. Core points
- ii. Border points
- iii. Noise/outlier points

To do this, it starts with any random point in the data which is not determined as a cluster or noise point and extracts its neighborhood using the epsilon distance. If the neighborhood contains at least minPts, this point will be marked as a core otherwise noise point. If this point is a core point, then it expands the cluster using the parameters (epsilon and minPts) and adds more and more points into the same cluster until no more points can be added. However, if this point is a noise point, the algorithm chooses another point which is not been visited before and sees if it's a core point or not. The point labeled as the noise might become part of a cluster when visited from other points. In this case, it will be labeled as a border point and will be assigned to a cluster whose core point

is the closest one. This process is repeated until all points are visited and assigned to a cluster or marked as noise/outlier (bin Waheed et al., 2019).

The epsilon and minPts can be estimated by the hit and trial method, or the epsilon value can be assessed using the k-distance method for the input data. This k-distance can be calculated using different pairwise distance measurement approaches such as Euclidean, Square Euclidean, and Manhattan or City Block method, etc. In the k-distance method using the Euclidean approach, for every point in the data, it calculates the distance to the k^{th} nearest points and plots the sorted points vs the calculated distance. This plot shows a knee and the ideal value for the epsilon gets located at the maximum curvature of the knee. After this value of the epsilon, the noisy points start to be included in the cluster. A higher value of epsilon means a smaller number of clusters in the data while a low value means many clusters formation for the same minPts.

For the minimum points, the general guidelines are:

- For a noisy dataset, the value should be large and vice versa.

Using MATLAB, I have run the code, and the result of clustering is shown in Figure 3.1. DBSCAN identified two clusters in the data. All empty circular-shaped points with different colors are the data points. Blue color points show cluster one, magenta color points show cluster two whereas red color points are noise/outliers. The epsilon distance set for the data is one and the minPts accepted in the distance epsilon are 10.

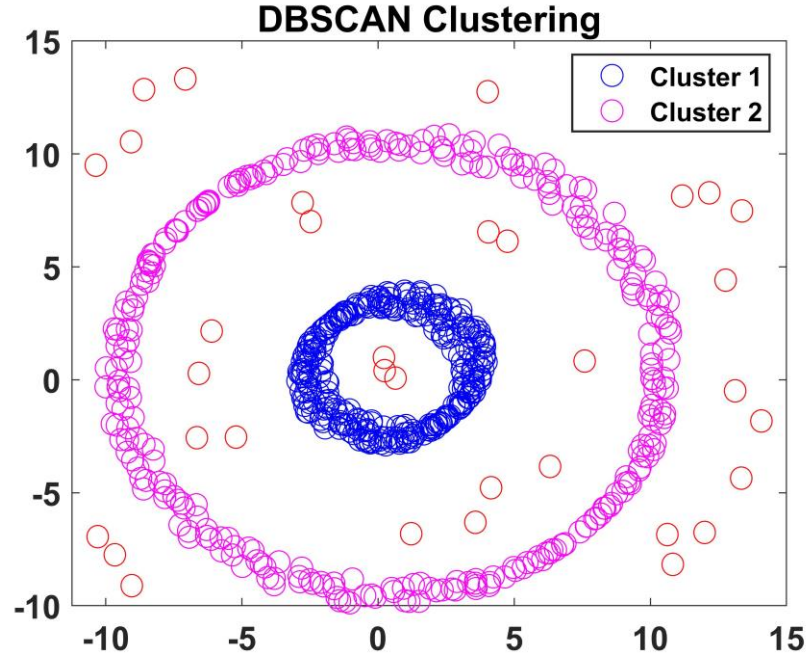


Figure 3.1: DBSCAN clustering result with epsilon = 1 and minimum points = 10, Blue color circle points shows cluster 1, magenta color circle shows bluster 2, whereas red circles are either outlier/noise points, the Figure generated by using MATLAB, version 2020a.

3.3 SEISMIC INTERFEROMETRY (SI)

Seismic Interferometry can be defined as, in general, an interference phenomenon between pairs of seismic signals to generate patterns to estimate the subsurface properties. These patterns, formed by cross-correlation and summation of seismic traces, are useful to get information about the earth's elastic characteristics (Schuster, 2009). In other words, seismic interferometry is to re-datum the data by cross-correlation and summation for different shot gathers to remove the overburdened statics. SI has been employed in different applications e.g., the transformation of vertical seismic profile to single well profile, vertical seismic profile to surface seismic profile, super-virtual refraction interferometry (SVI), etc. A simple illustration to describe the Seismic

Interferometry idea is shown in Figure 3.2 below where the model represents a 2D seismic data profile recorded over a sandstone lens comprised of a complicated overburden. The seismograms are recorded by assuming a 1D wave propagation model and only considering vertical wave recording for simplicity. Each trace is recorded by designing the layout as having one shot and one receiver at each location. The record shows us two events recorded; the earlier one from the top of the sandstone lens shown by dashed lines and the later one from the bottom of the sandstone lens shown by solid lines as shown in Figure 3.2. The objective to determine the shape of the sandstone lens is interrupted by the complex overburden caused by the velocity anomalies (Figure 3.2a). These interruptions are sometimes called residual statics which need to be removed to get a clear shape of the sandstone lens using the seismic data. To do so, we observe that the two-way traveltimes for all the traces through the overburden from the top of the sandstone lens back to the earth's surface is equal to τ_{LML} , which changes for each trace depending on the complicated near-surface statics caused by velocity anomalies (Figure 3.2a). To eliminate this static, we can shift each trace by τ_{LML} , and this shifting is roughly equivalent to cross-correlating two traces. The result after this shifting is shown in Figure 3.2b where we not only removed the effect of the complex near-surface but also, we moved our source locations closer to the target which improves the resolution. There are many benefits of Seismic Interferometry e.g., we don't need to know the source location, velocity model, and excitation time. SI has been employed in different applications e.g., converting vertical seismic profile to single well profile, vertical seismic profile to surface seismic profile, etc., (Schuster, 2009). Moreover, many researchers used SI

successfully to enhance the SNR of the seismic data (H.J. Nicolson, 2011; H. Nicolson et al., 2012; Al-Shuhail et al., 2013; Kang, 2014; Liu et al., 2014).

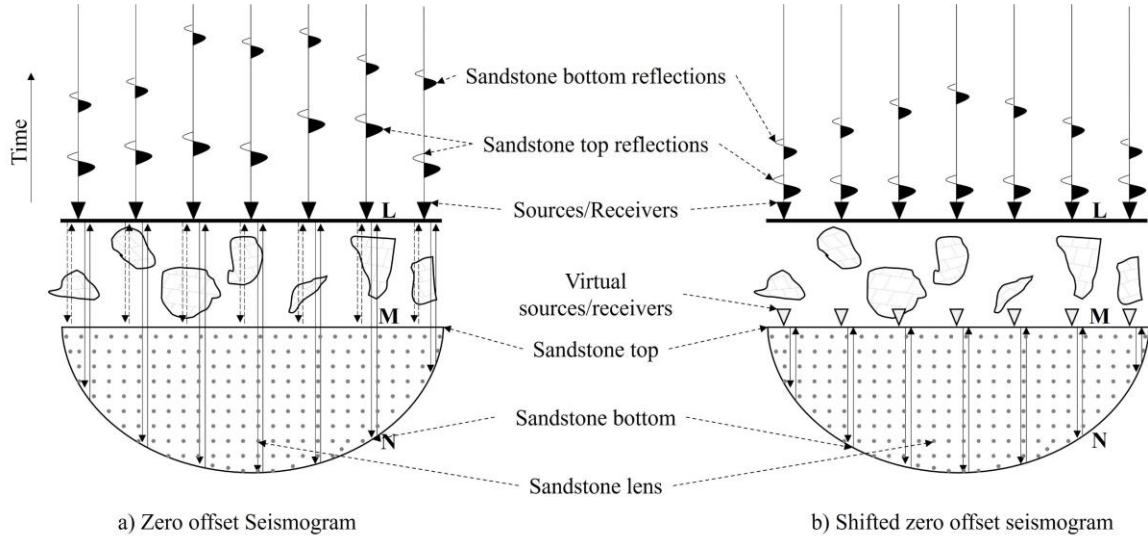


Figure 3.2: Reflection seismograms (zero and shifted zero offsets) over a sandstone lens with complex near-surface/overburden (a), representing static removal using Seismic Interferometry (b), (Schuster, 2009).

There are different types of Seismic Interferometry e.g., reflection, and refraction (Dong et al., 2006; Schuster, 2009; Al-Hagan et al., 2014). Refraction interferometry was introduced by Dong et al. (2006) to tackle the problem associated with refraction imaging in the controlled-source data. The issue with refraction imaging was the low signal-to-noise (SNR) ratio and this ratio decreases with an increase in distance between source and receiver. As result, the first arrival traveltimes picking of the refraction data becomes limited and this consequence adversely affects the application use of refraction imaging. By using refraction interferometry, this problem was partially mitigated, and the SNR of the far-offset arrivals was enhanced by $\sqrt{N_s}$, where N_s are the post-critical source positions that are coincident with receivers associated with the generation of the first break. The general idea of this method is to correlate two traces and stack them for

different shot positions to yield a virtual refraction with enhanced SNR (Figure 3.3a). There were four problems associated with the Dong method of SI (Al-Hagan et al., 2014). After the correlation step, the excitation time remains unknown which is an important input to find the depth of the refractor (see Figure 3.3a after correlation in which the virtual trace has a virtual source at geophone position x' having $-xx'$ excitation time). The second problem was that the offset between the source and receiver of the virtual trace becomes smaller due to smaller traveltimes and ray-paths. Thirdly, the source location remains unknown, and lastly, the wavelet becomes distorted after the cross-correlation step. In remedy to the first problem, Dong et al. (2006) proposed that the virtual source position can be repositioned to the earth's surface by rectifying the stacked virtual refraction traces to an observed traveltime in the original data. Nevertheless, to cope with these issues and to further enhance the SNR of the far offset arrivals, Bharadwaj et al. (2011) introduced the theory of "Super-virtual Refraction Interferometry" shown in Figure 3.3 (a & b). SVI relocated the virtual source by convolving it with the recorded trace which resolved the issue of unknown excitation time and unknown source position. Also, by this, the receiver spread can be extended to its maximum recording extent and the absolute first break arrival time is appropriately accounted for as well. However, the output wavelet remains distorted after SVI which cannot be fully resolved so far.

3.4 SUPER-VIRTUAL REFRACTION INTERFEROMETRY (SVI)

The theory of Super-virtual Refraction Interferometry was introduced by (Bharadwaj et al., 2011; Bharadwaj et al., 2012). The typical concept of SVI is to use the far-field

reciprocity equation of cross-correlation and convolution (Equations a and b are given below). It first generates and stacks virtual refractions using cross-correlation of traces for all post-critical source positions followed by convolving the virtual refractions with the raw data and stacking these convolved refractions which are known as super-virtual refractions. By this, the SNR of the far offset first break improves drastically. Numerous scholars used the SVI theory successfully to enhance the far offset arrivals' SNR e.g., Hanafy and Al-Hagan (2012) and Al-Shuhail (2015) used SVI to increase the SNR of the far offsets to improve the manual and automatic first break picking, Bharadwaj et al. (2013) used ocean bottom seismic data to enhance the first break SNR and this significantly increased the number of reliable pickable first breaks traces, (Hanafy et al., 2011; Hanafy and Al-Hagan, 2012) employed SVI for the land seismic data to achieve a similar output as of (Bharadwaj et al., 2013).

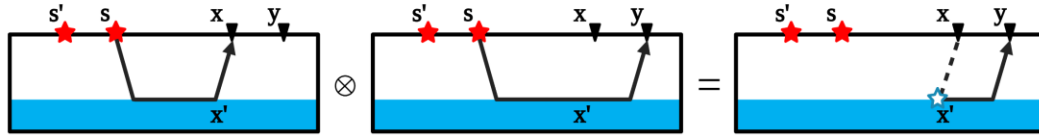
$$\text{Im}[G(\mathbf{x} | \mathbf{y})^{\text{virt.}}] \approx k \int_{\text{post-crit sources}} G^*(\mathbf{x} | \mathbf{s}) G(\mathbf{y} | \mathbf{s}) ds \quad (a)$$

$$G(\mathbf{y} | \mathbf{s}')^{\text{super}} \approx 2ik \int_{\text{receives}} G(\mathbf{x} | \mathbf{y})^{\text{virt.}} G(\mathbf{x} | \mathbf{s}') dx \quad (b)$$

SVI requires one main input to run, and this input is the determination of the muting window around the first break area. Otherwise, there would be unwanted artifacts in the output. These artifacts are caused by the violation of the assumptions of the reciprocity equation. As a solution, to get rid of these artifacts, Dong et al. (2006) gave the idea to mark a proper window around the first break so that only the head-wave refractions are correlated with each other. It will remove the artifacts generated by unwanted cross-correlation among direct arrivals, reflections, and refractions. To achieve the best results, the window is

aimed so that all first arrivals except those coming from the bottom of the same layer that reaches within one to two periods of the estimated first break time are muted.

a) Crosscorrelate and stack over post-critical source positions to generate virtual refractions



b) Convolve and stack over different receivers' locations to generate super-virtual refractions

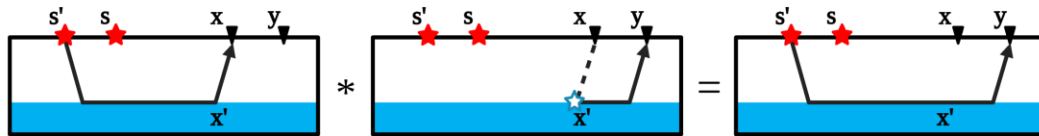


Figure 3.3: Schematic diagram for generating super-virtual refraction traces using SVI (Hanafy and Al-Hagan, 2012).

CHAPTER 4

WORKING PRINCIPLES OF DBSCAN AND SVI

4.1 DBSCAN TECHNIQUE

In DBSCAN, it classifies the objects in an area as a cluster that is tightly packed together. A high concentration area of the objects means a cluster whereas a low concentration area means noise or an outlier (bin Waheed et al., 2019).

DBSCAN has two key inputs to do the clustering: ϵ : The radius of the vicinity nearby the data point and minimum points: The smallest value of data points in a neighborhood of a point. The ϵ and minimum points can be estimated by using the hit-and-trial method. As a general guideline, a small ϵ value with a small number of minimum points for noise-free data and a small ϵ value with a large number of minimum points for noisy data is chosen.

DBSCAN divides the data objects into three categories to perform clustering: Core, Border, and Noise points. Core objects are labeled core objects as if the vicinity of the object contains minimum objects. Border objects: If the vicinity has fewer than minimum objects but it is accessible through other core objects, it is marked as a border object. Noise objects: an object is marked as noise if it is neither a core object nor a border. Using these definitions, the following are the steps of a DBSCAN algorithm (bin Waheed et al., 2019):

1. Start with any arbitrary object which does not belong to a cluster or is labeled as a noise object. Count the number of vicinity objects inside a distance ϵ from that object.

Characterize it as a core object if the vicinity contains objects greater than minimum objects. Otherwise, mark it as noise.

2. When a core object is established, enlarge the group through the addition of all directly accessible objects. If a noise object is included, alter its standing from a noise object to a border object.

3. Steps 1 and 2 are reiterated to all data objects unless all the objects are either marked as a cluster or an outlier.

4.2 SUPER-VIRTUAL REFRACTION INTERFEROMETRY (SVI)

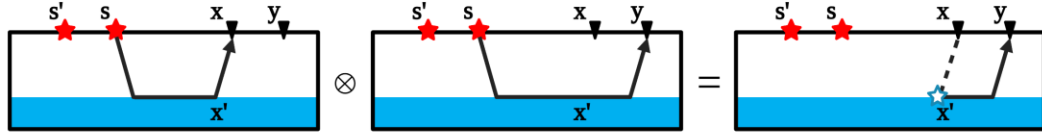
To generate the output for SVI i.e., super-virtual refractions, the equations of reciprocity for correlation and convolution were used as given below in (a) and (b).

$$\text{Im}[G(\mathbf{x} | \mathbf{y})^{\text{virt.}}] \approx k \int_{\text{post-crit sources}} G^*(\mathbf{x} | \mathbf{s}) G(\mathbf{y} | \mathbf{s}) d\mathbf{s}. \quad (a)$$

$$G(\mathbf{y} | \mathbf{s}')^{\text{super}} \approx 2ik \int_{\text{receives}} G(\mathbf{x} | \mathbf{y})^{\text{virt.}} G(\mathbf{x} | \mathbf{s}') d\mathbf{x}. \quad (b)$$

In Figure 4.1(a), $\mathbf{s}$ and $\mathbf{s}'$ are the positions of the sources whereas $\mathbf{x}$ and $\mathbf{y}$ are the receiver's locations. k shows the average wavenumber. For a virtual shot located at $\mathbf{x}'$, and receiver at $\mathbf{y}$, $G(\mathbf{x}|\mathbf{y})$ is the Green's function having a virtual excitation time equivalent to $-\tau_{\mathbf{x}\mathbf{x}'}$. The output of SVI denoted by $G(\mathbf{y}|\mathbf{s}')^{\text{super}}$ is generated by the convolution of field data $G(\mathbf{x}|\mathbf{s}')$ with the virtual data $G(\mathbf{x}|\mathbf{y})^{\text{virt.}}$ as shown in Figure 4.1(b).

a) Crosscorrelate and stack over post-critical source positions to generate virtual refractions



b) Convolve and stack over different receivers' locations to generate super-virtual refractions

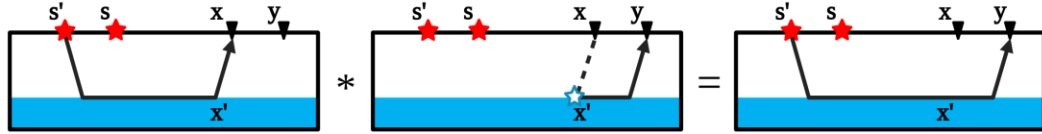


Figure 4.1: Schematic diagram for generating super-virtual refraction traces using SVI (Hanafy and Al-Hagan, 2012).

The steps involved in SVI are illustrated in Figure 4.1 and explained as follows:

a) The virtual trace $\phi_x(x, y, t)$ is created by correlating the field data trace located at x to the trace at y for the source at s . This virtual trace will have traveltime given by $\tau_{x'y} - \tau_{x'x}$. This arrival time will be the same for all post-critical source positions provided that the arrivals are from the same layer, so stacking $\sum_x \phi_x(x, y, t)$ will enhance the SNR of the virtual refraction by $\sqrt{N_v}$. N_v is the number of post-critical source positions that generate this specific refraction.

b) In the second part, the virtual refraction traces generated before are convolved with the field data traces. The results are summed for different geophone locations to get the super-virtual traces with improved SNR by a factor of $\sqrt{N_{sv}}$. N_{sv} , here, is the number of coincident receiver positions that are at post-critical offset for this specific refraction.

CHAPTER 5

WORKFLOW AND METHODOLOGY

There are two techniques (DBSCAN and SVI) used throughout this study. In case of noisy data, DBSCAN will give an initial guess of the picking. This initial guess will guide to the window location to mute around the first breaks. Then SVI will enhance the SNR of this windowed data. In the end, DBSCAN will find the first arrival's traveltime automatically.

5.1 DBSCAN ALGORITHM

The steps used in the proposed DBSCAN algorithm are:

1. Select a sliding window with a predetermined number of traces and time samples (Figure 5.1a).
2. Select a range of velocities and velocity step, and then apply the linear moveout (LMO) for each velocity step.
3. Stack traces after LMO (Figure 5.1b) and apply a threshold to the stacked image to remove the small LMO values (Figure 5.1c).
4. Apply the DBSCAN clustering algorithm to form the cluster (using predefined epsilon and minPts) and determine the optimum velocity and the picking time values from the centroid of the cluster (Figure 5.1d).
5. Extrapolation is used to find picks for traces that DBSCAN could not pick.
6. Smooth the pickings by using the moving average method and apply a constant shift to reduce the error of the picks from the manual and compensate for the SVI delays.

7. Repeat this procedure on all other shot gathers.

Before starting the DBSCAN algorithm, we should apply a constant shift to all the traces so that DBSCAN does not miss the picking for very early arrivals in the window. Mute the amplitude values as much as possible to clean the data. Normally, each first arrival picking is more in time than the previous one, so, it follows a slope. The trace, if it has good SNR, which comes first in the window forms a cluster at the bottom of the window. At the same time, clusters due to the random noise if any, mostly form within the first half of the window. Therefore, in the presence of noise, we should accept the clusters that form within the center to the bottom of the window. This is achieved by applying a cluster's centroid acceptance condition and it improved the initial guess significantly. DBSCAN was able to accurately pick only the noise-free data and failed severely in its presence. The main reason I found was that the algorithm was scanning the whole data from the first to the last sample in the data to find the first break traveltimes. Although, the noise is random, yet, when I apply DBSCAN, it forms clusters in the noisy area. It was making the DBSCAN a huge set back to pick the accurate/near-to-accurate traveltimes. As we know, the first break follows the time offset attribute, and usually, each next first arrival traveltimes is more in time than the previous one, hence, restricting the scanning window after picking the first value could enhance the DBSCAN findings a lot. So, I employed this idea to restrict the scanning window to find the first arrival in the presence of noise. This utilization enhanced the DBSCAN picking efficiency dramatically. Although it could not find the traveltimes perfectly, still it takes us near the accurate picking. I used these picked traveltimes to define a muting window which is a

key input for super-virtual Refraction Interferometry to enhance the SNR of the noisy data.

5.1.1 SELECTING A SLIDING WINDOW

Selection of an appropriate user-defined window with a specified number of traces and samples must be done before any calculation steps (Figure 5.1a & 5.2a). The idea behind this is to specify the number of traces and samples, on the x and y-axis, respectively. It makes sure that the algorithm does not miss any first arrival in the whole dataset. The values of traces and sample windows differ from one data set to another. Nevertheless, as a general guideline, the length of the sample window should be greater enough to have at least one period of the data and a trace window should be chosen enough so that it has enough events which show the slope of the data. As an optimum practice, the trace window should be small (5-10 traces) irrespective of the noise in the data, and the time window should be 1 to 3 periods, preferably one period.

5.1.2 APPLYING LINEAR MOVEOUT CORRECTION (LMO) AND STACKING

There are some seismic events such as first arrivals that have time-offset attributes that can be represented by a straight line. The LMO algorithm is applied to remove the effect of offset on traveltimes and eliminate the dipping of first arrival events to project these events horizontally. LMO uses a constant linear velocity to compute a single time shift applied to each trace in the given dataset. Therefore, it's typically used to “flatten” an event that is linear with offset on a shot gather. Mathematically, LMO is defined as:

$$t_{LMO}=t_0+ x_n/v. \quad (c)$$

where t_{LMO} is the static displacement on the n^{th} trace, t_0 is two-way traveltime corresponding to the zero offset where the first trace is always fixed without applying LMO correction, $x_n=(n-1)\Delta x$ is the offset of the n^{th} trace, v is the constant linear velocity, and Δx is the distance between two geophones/traces.

The first step in the LMO is to select a velocity range within the selected subset of the data to be attempted in equation (c) and then stack the traces for each selected velocity to calculate the output energy (Figure 5.1b and 5.2b). The maximum output energy will give the optimum velocity for the given data subset in the window. To discriminate between signal and noise, I apply DBSCAN on the cluster image (Figure 5.1c, and 5.2c) that is generated using a threshold value to the stacked LMO image (Figure 5.1b and 5.2b). If I have 1 to 3 clusters in the LMO stacked image, then data in the window is categorized as a signal (Figure 5.1c), else the data in the window is considered noise (Figure 5.2c). If the selected window contains signals, then traces will show coherent events and LMO amplitudes can be clustered together (Figure 5.1d). However, a window with only noises shows no coherent events, and, hence, the LMO amplitudes are broadly dispersed, making no cluster or many small-size clusters. Signals are less likely to be misclassified as noise when larger thresholds are used, especially when the data has a low SNR. However, by increasing the likelihood of identifying noise as signals, the trade-off is a decrease in classification accuracy.

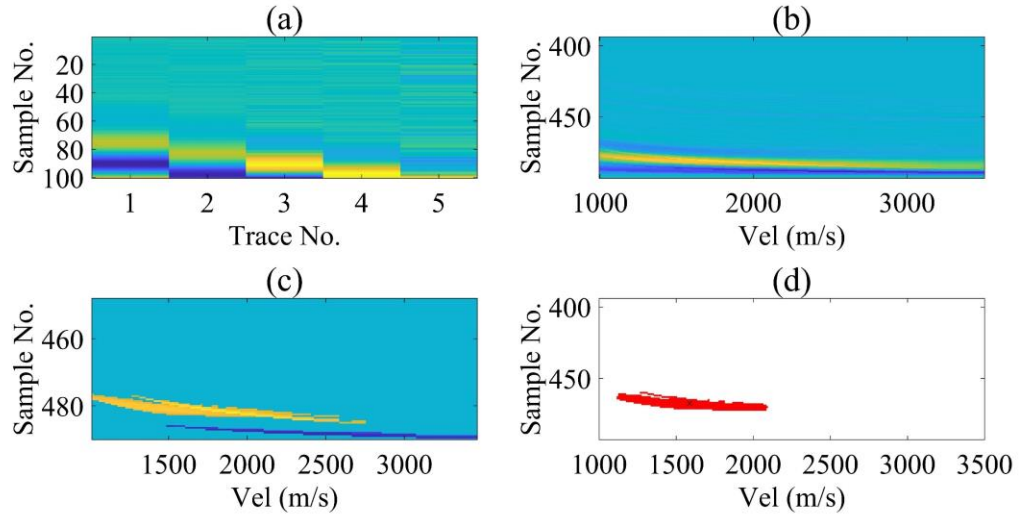


Figure 5.1: DBSCAN clustering in the presence of data and noise. (a) shows the selected window with 5 trace and 100 sample numbers, (b) is the stacked LMO image, (c) shows the stacked LMO image after applying the threshold, and (d) represents DBSCAN cluster formation.

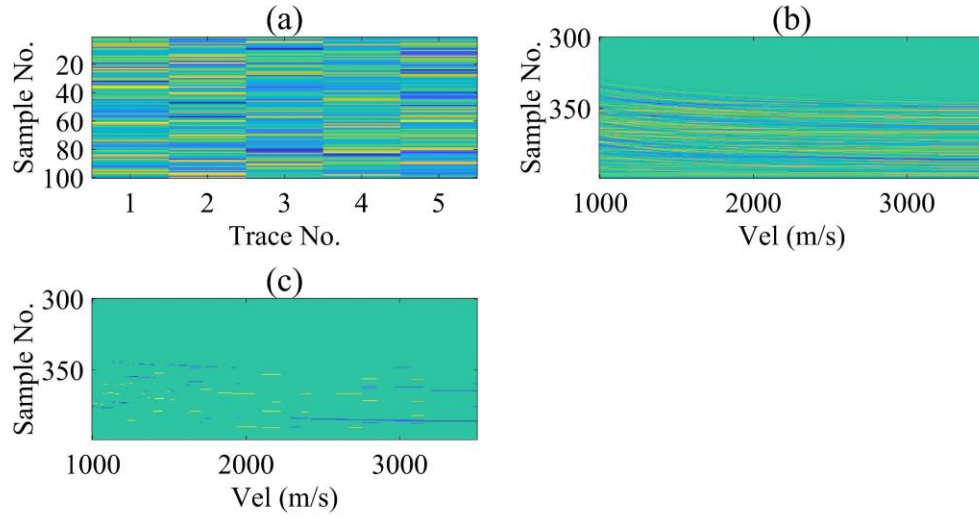


Figure 5.2: DBSCAN shows no or many clusters in the presence of noise only. (a) shows the selected window with 5 trace and 100 sample numbers, (b) is the stacked LMO image, (c) shows the stacked LMO image after applying the threshold, and there is no DBSCAN cluster formation.

5.2 SVI ALGORITHM

After muting the near offsets direct arrivals and around the probable first breaks that are from the bottom of the same layer, SVI gets this muted windowed data as its input and generates the super-virtual traces with improved SNR of the first breaks after the cross-correlation and stacking, and convolution and stacking. A deconvolution filter can be applied after the cross-correlation step on the virtual traces to decrease the ringyness issue in the first breaks. Unluckily, the SNR improvement of SVI is, in some cases, not sufficient for accurate picking of far-offset first breaks. In these cases, SVI needs to be applied iteratively which could improve the SNR of far-offset sufficiently. Therefore, DBSCAN can pick these enhanced SNR arrivals accurately. For iterative SVI, the SVI traces will be used as the new input to the SVI technique. It will result in an output with improved SNR of far-offset traces.

The steps used in the proposed SVI method are shown in Figure 5.3 below:

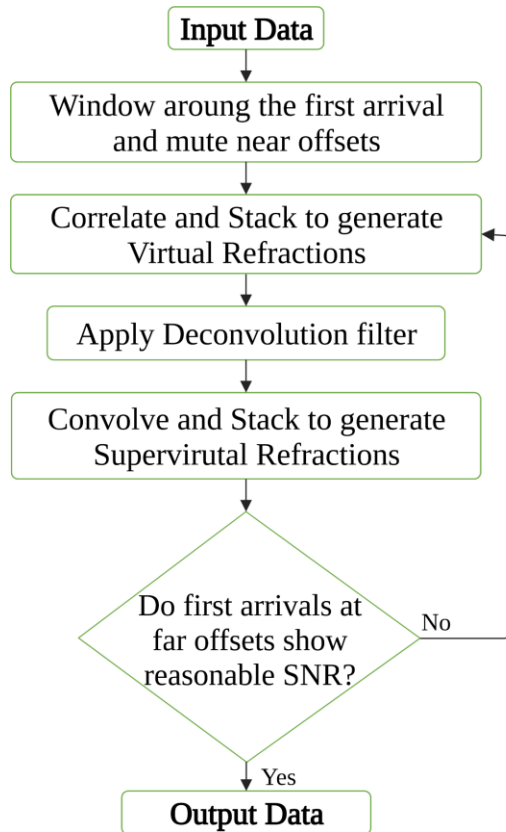


Figure 5.3: Flowchart indicating the different steps involved in SVI.

5.3 PROPOSED METHODOLOGY

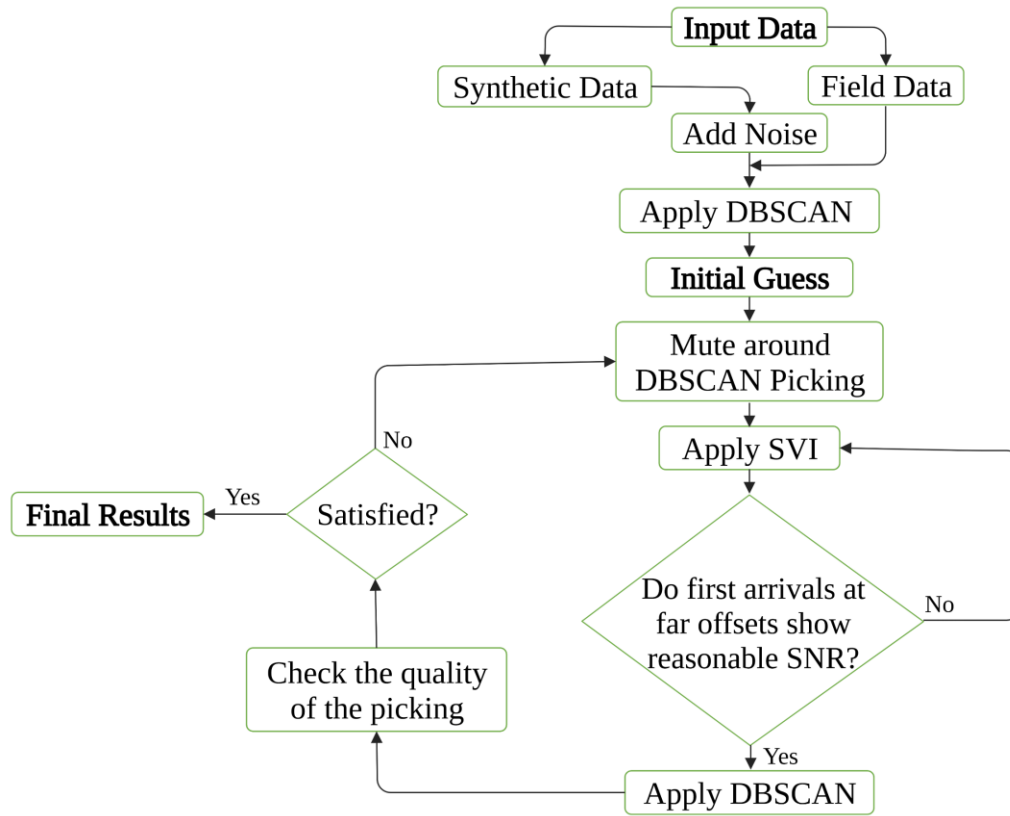


Figure 5.4: Research methodology accurate first break picking automatic using DBSCAN and SVI.

CHAPTER 6

DATASET INFORMATION AND PREPROCESSING

6.1 SYNTHETIC DATA

A two-layer 2D velocity model, with a dimension of 600 m wide and 80 m in depth is used to generate a synthetic dataset using acoustic finite difference forward modeling as shown in Figure 6.1. The layers' velocities are constant, with values of 500 and 1500 m/sec for the first and second layers, respectively. Both layers' density is constant assuming the acoustic medium and layer two has varying thicknesses as shown in Figure 6.1. I designed the layout as having 121 receivers and 121 shots, placing a shot at each receiver location with an interval of 5 m, 0.00025 sec time interval, and total run time equal to 1.250 sec, the peak frequency is 30 Hz using Ricker wavelet. To focus on the refracted data, I am only showing almost half (2500) of the samples recorded.

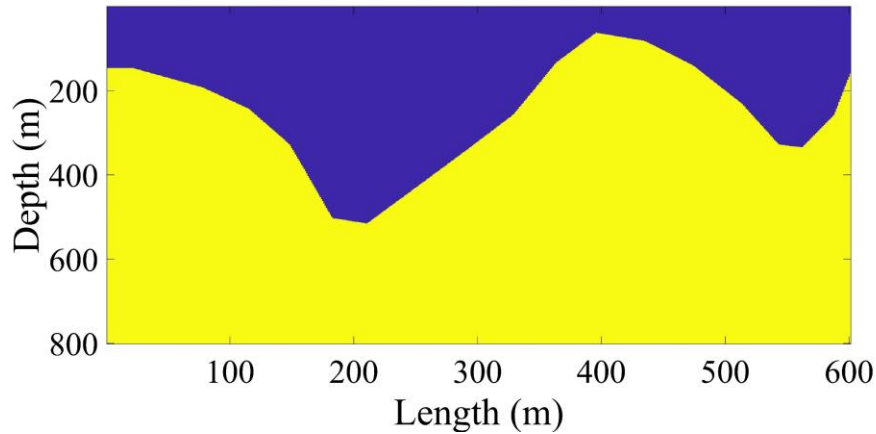


Figure 6.1: Synthetic velocity model showing a two-layer velocity model with first and second layer's velocities of 500 m/s and 1500 m/s, respectively.

Figures 6.2, 6.3, and 6.4 show three examples of shot gathers generated using the velocity model, where two of them shows an end-on spread while the other one has a split spread.

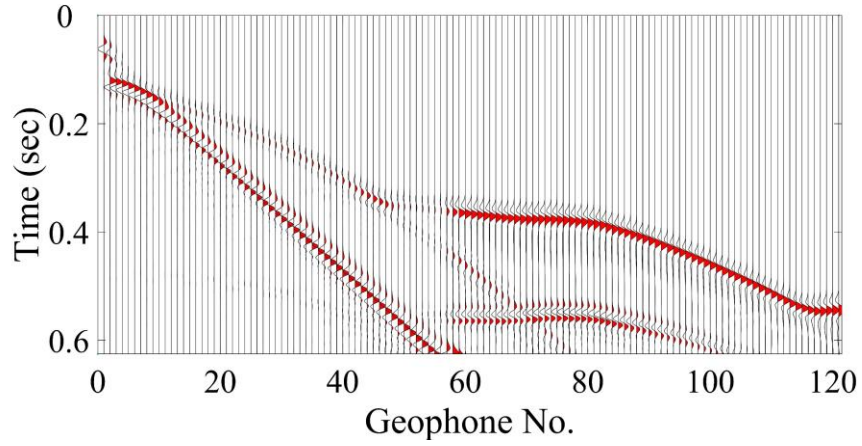


Figure 6.2: Shot gather 1 from the synthetic data.

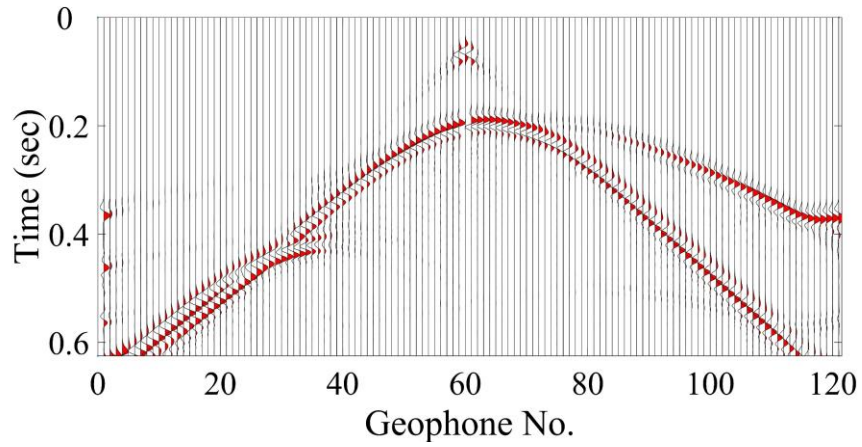


Figure 6.3: Shot gather 60 from the synthetic data.

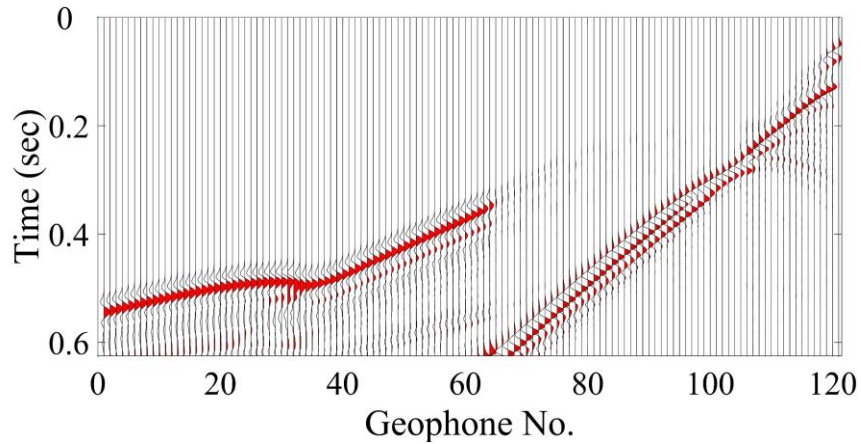


Figure 6.4: Shot gather 121 from the synthetic data.

6.2 ITHRA FIELD DATA

The refraction field data has been collected near Ithra Dammam, eastern Saudi Arabia, using conventional seismic data layout having dimensions of 600 m long profile, placing each shot at each receiver location. Five geodes are used each containing 24 geophones to cover the profile length, so, a total number of 120 geophones are used to record the data with an interval of 5 m, and 120 shot gathers with the same interval as of geophones are collected in the field, sample interval of 0.0005 sec, with a total time to record is two sec. The shot was generated using an accelerated weight drop of 200 lbs. and I used 5 stacks at each shot location. The noise level in the data is severe in many of the shots as you can see in Figures 6.5 and 6.6 for shot gathers 1, and 100. To focus on the refraction data, I am only using less than half (1500) of the samples recorded.

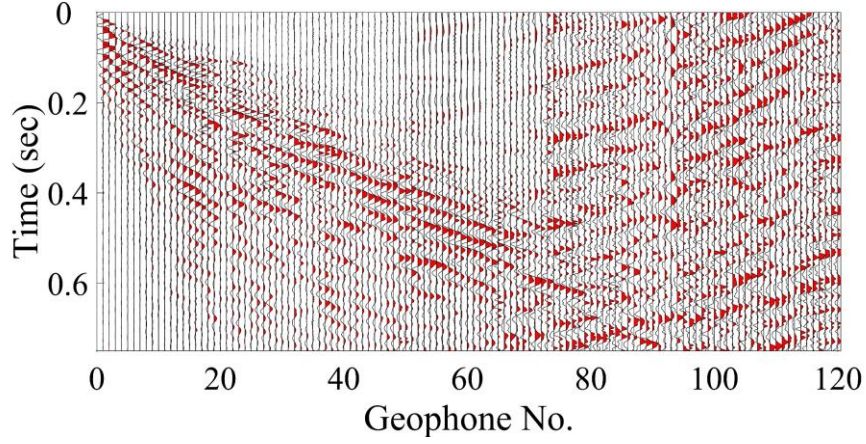


Figure 6.5: Ithra field's data shot gather 1, end spread.

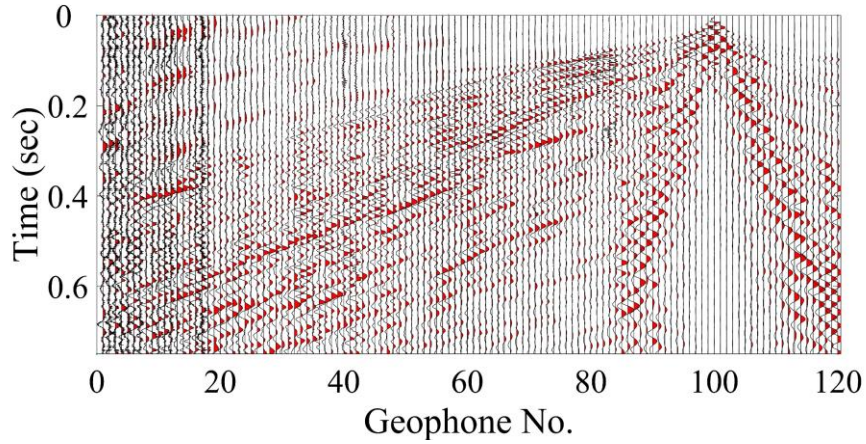


Figure 6.6: Ithra field's data shot gather 100, split spread.

6.3 YILMAZ FIELD DATA

To compare the results of our methodology with other methods, I have chosen two shot gathers (6 and 25) from Yilmaz's published assortment of worldwide field seismic datasets (Yilmaz, 2001). Both shot gathers have been recorded using dynamite as the source. Shot gather 6 has 48 traces, with 4 msec as the time sample interval and receiver spacing of 100 m. While the shot gather 25 has 98 traces, with 2 msec as the time sample interval and receiver spacing of 50 m. Both the shots have good SNR (Figures 6.7 and

6.8) as compared to the synthetic data with added random noise (Figures 11a and 11f) and the Ithra field data (Figures 6.5 and 6.6). So, I am expecting very accurate first break picking as compared with other automatic or semi-automatic methods that have been tested on these two shots against manual picking.

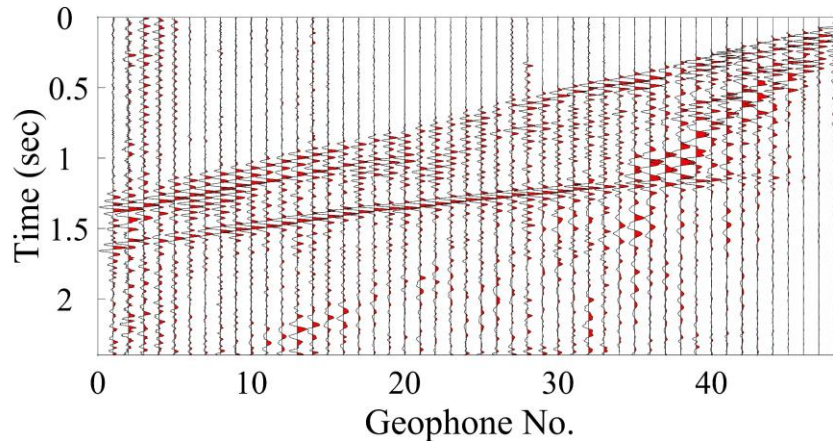


Figure 6.7: Yilmaz shot gather 6 recorded in the Far East.

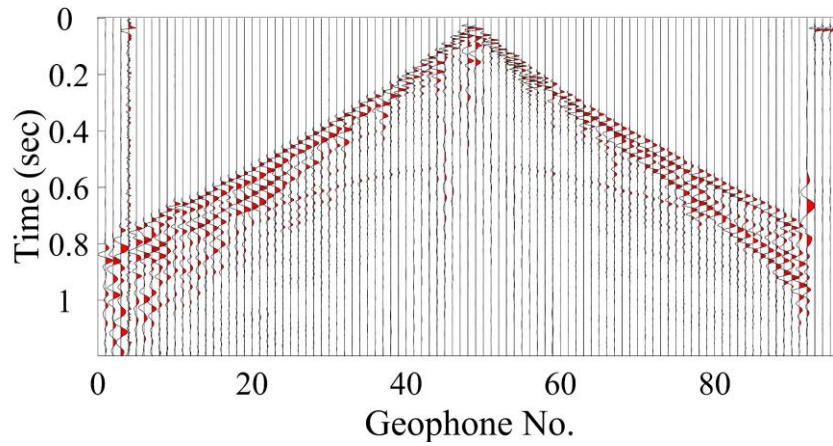


Figure 6.8: Yilmaz shot gather 25 recorded in Alberta.

6.4 PREPROCESSING

Before starting with the DBSCAN method for an initial guess, we must preprocess the data to make sure that our initial guess for synthetic and field data will come as accurately as possible. For synthetic data, I observed that the direct and first break amplitudes were very low in many of the traces as we can see in Figure 6.9 below. The area of the weak amplitude direct and first break on the left side of the shot is shown by the two black color double head arrows (Figure 6.9).

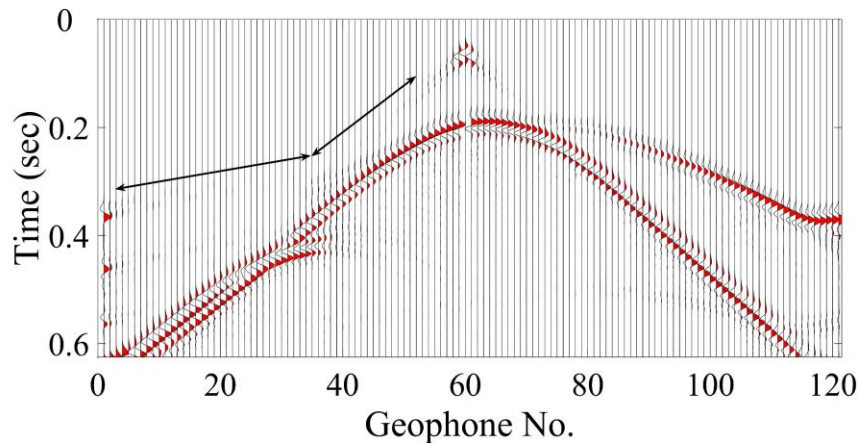


Figure 6.9: Synthetic shot gather 60 showing small amplitude first arrivals, the area is indicated by black two head arrows.

Among others, there could be two ways to improve the amplitudes of weak first arrivals. One can be to apply automatic gain control and the other could be to increase the amplitudes by clipping method. The idea of clipping is that we multiply all the normalized data with a high value and then divide these enhanced amplitudes values by a smaller value followed by normalization again.

By this, the small amplitude values increase by the ratio of high and low values. For example, if we multiply the amplitudes by 50 and then divide by 5, the small amplitude

values will increase by 10 times as compared to the initial amplitude values. These values are selected so that most of the weak amplitude values become visible. For my synthetic data, I applied the clipping method and increased the weak direct and refraction amplitude by 20 times. I did not apply the automatic gain control as the image becomes very ugly after applying it. The result with enhanced amplitudes of direct and first break arrivals are shown by the two black color double-head arrows after applying the clipping method (Figure 6.10). We can clearly observe that those direct and first break arrivals which had very weak amplitude values before (Figure 6.9) get visible after clipping application (Figure 6.10).

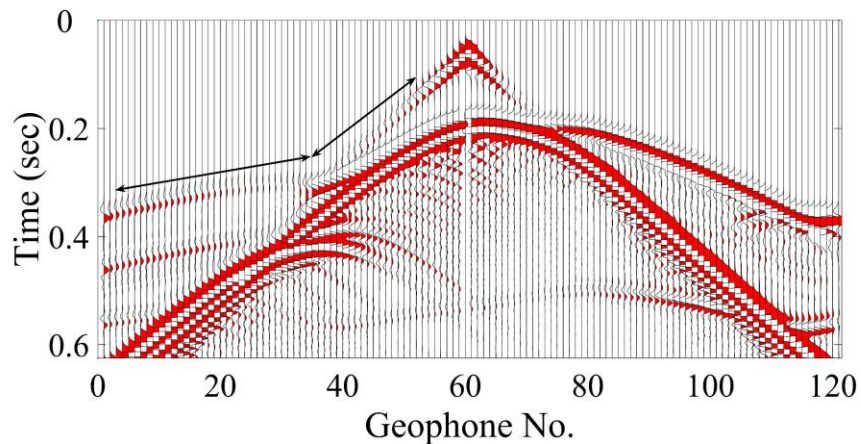


Figure 6.10: Synthetic shot gather 60 after 20 times increased amplitude as compared to the original data.

For the field data, many traces were either dead or very noisy, so, I did mute them first and then interpolated them using the common offset gather interpolation method.

6.5 PARAMETERS DETERMINATION FOR SYNTHETIC AND FIELD DATA

Different parameters need to be estimated before applying the DBSCAN method for the picking. These parameters include peak frequency, time-period, epsilon, the minimum number of points estimations, sample and trace window size, layers' velocity, velocity range, and velocity step determination for linear moveout image.

6.5.1 PARAMETERS DETERMINATION FOR SYNTHETIC DATA

For the synthetic data, I know the source peak frequency (30 Hz), and refraction layers' velocity (1500 m/sec). From peak frequency, I calculated the time-period (~33 msec). The sample window should be ~1-3 periods whereas the trace window value must be smaller (5-10 traces) irrespective of the noise in the data. The sample window should be at least 1 period so that it has one crest and one trough in the window so that a good cluster can be formed. A bigger sample and trace window gives bad picking as more noise and noisy traces will be present in the window which could lead to a bad initial guess or final picking. For epsilon and minPts, by hit and trial method, I found that epsilon and minPts should be 4 and 35 respectively for a velocity step of 10 m/sec. Generally, a smaller velocity step is good to choose. Nevertheless, if we increase the velocity step value, we need to keep in mind that the number of points in the LMO image will decrease and their pairwise distance will increase. So, we will have to adjust epsilon and minPts values again. As a general guideline, for noisy data, the epsilon value should be small with many minPts and vice versa for low-noise data. As we know, for the synthetic data, our first break arrivals are generated from the top of the second layer and

its velocity is 1500 m/sec, we should select our velocity range as 1500+500 m/sec and 1500-500 m/sec to apply for LMO image (Table 1).

6.5.2 PARAMETERS DETERMINATION FOR FIELD DATA

For the field data, by using Fourier transform, the average peak frequency came out to ~ 30 Hz. I calculated the time-period as ~ 33 msec. From different shot gathers, I observed that there are two layers from where the first breaks are generated. As an example, shot no. 25 is shown in Figure 6.11. Two black colors single-headed arrows show two different refractions' slopes. Therefore, I estimated the velocities of both refraction layers using the slope of the first break. Taking the average from different shots, I found that the average first layer velocity is ~ 1000 m/sec and for the second layer is ~ 4200 m/sec. Hence, I applied two different velocities for each layer's first arrival traveltimes picking with a difference of 500 m/sec from the estimated value of each layer to build the LMO image (Table 1).

Again, for epsilon and minPts, using the hit-and-trial method, I found that epsilon and minPts should be 4 and 35 respectively for a velocity step of 10 m/sec, and generally, a smaller velocity step should be chosen.

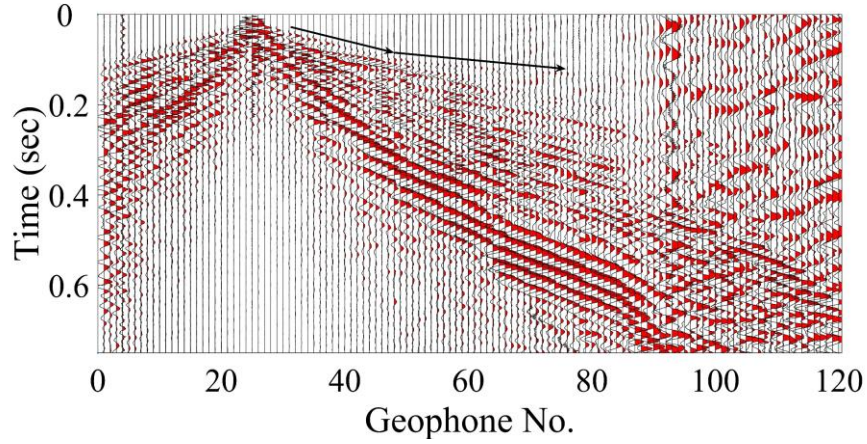


Figure 6.11: Two black arrows show two different slopes of the refraction layers.

Table 1: Parameters selected for different datasets. These are the same for the initial/final first break.

Data	Parameters					
	Epsilon	MinPts	Trace Window	Sample Window	Velocity Range, and step for LMO (m/sec)	Threshold for LMO
Synthetic shot 60 with no random noise	4	10	10	90	1000-2000, 10	0.0
All synthetic shots with random noise	4	40	10	90	1000-2000, 10	0.4
All Ithra shots	4	35	10	40	-	0.57
Layer 1 refraction	-	-	-	-	550-1500, 10	-
Layer 2 refraction	-	-	-	-	3700-4700, 10	-
Yilmaz shots 6 and 25	4	35	5	50	2800-3800, 10	0.5

6.6 BANDPASS FILTER

Before applying super-virtual refraction interferometry to the field data, we should apply a bandpass filter to remove the low and high-frequency noises from the field data. After analyzing the field data in the frequency domain, the dominant frequency range was around 25-30 Hz. Therefore, the designed parameters for lower cut, lower pass, and high pass, high

cut are 25, 35, 80, and 100 Hz, respectively. A sample shot gather is shown before and after applying the bandpass filter in Figure 6.12a and 6.12b, respectively. We observe that low and high-frequency noises are successfully attenuated to much extent, particularly from the far-offset traces, from the data as shown in Figure 6.12b.

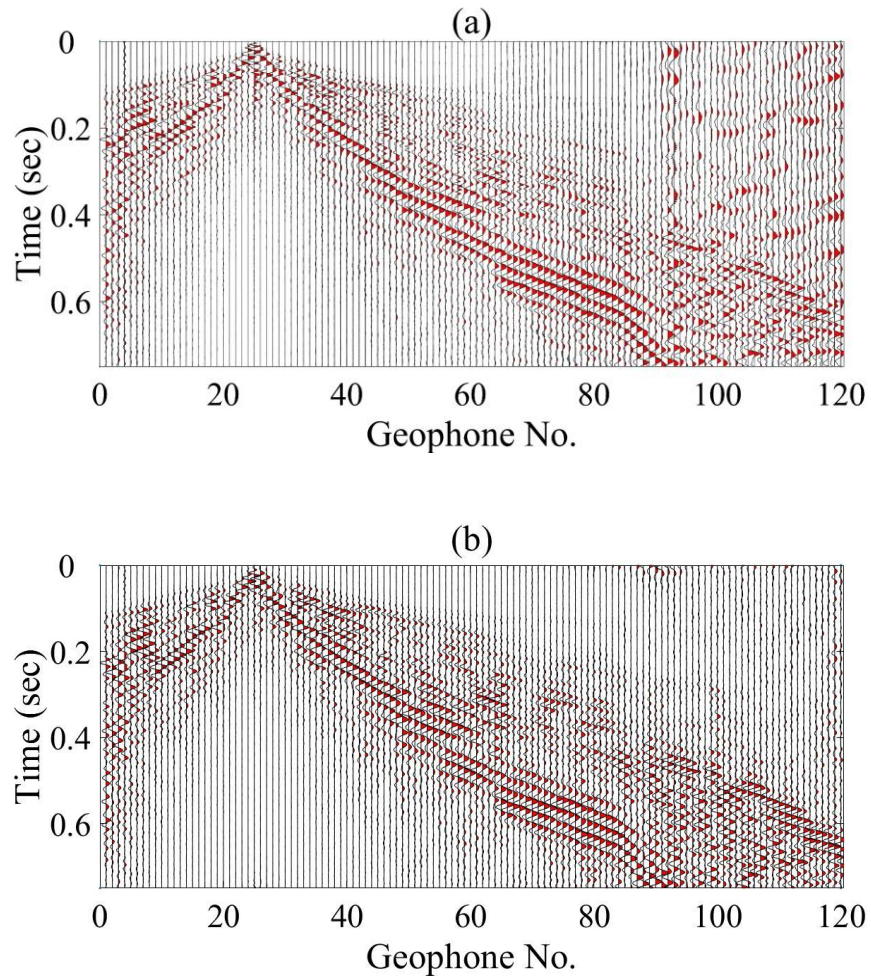


Figure 6.12: Shot gather 25 (a) before band pass filter (b) after bandpass filter.

CHAPTER 7

RESULTS AND DISCUSSION

7.1 RESULTS

The results are discussed into three categories: synthetic, Ithra, and Yilmaz field datasets. I applied the proposed technique to the synthetic data before and after adding random noise, while the field datasets have noise already. The synthetic and field results are shown with only DBSCAN and with DSBCAN + SVI + DBSCAN methodology except for the Yilmaz data where the SNR of the data is very high and only DBSCAN-alone picked the first break traveltimes accurately. Synthetic and field data results are not only compared with the manual pickings but also against the cross-correlation method developed by (Senkaya and Karsli, 2014). To benchmark the results further, Yilmaz's shot gathers 6 and 25 have been used to compare our results with other methods available in the literature for first break pickings.

7.1.1 SYNTHETIC RESULTS WITH NO NOISE

In the noise-free data, very low amplitudes should be muted. Make sure that the muting value is less than the first non-zero value of the first arrivals to get the best results. The manual picking defined as the first non-zero value of the amplitude is shown by magenta cross markers, and the DBSCAN picking is shown by blue circles for the noise-free data in Figure 7.1. As we can see in Figure 7.1, very accurate matching of manual with DBSCAN picking by muting the amplitude values less than 0.0010.

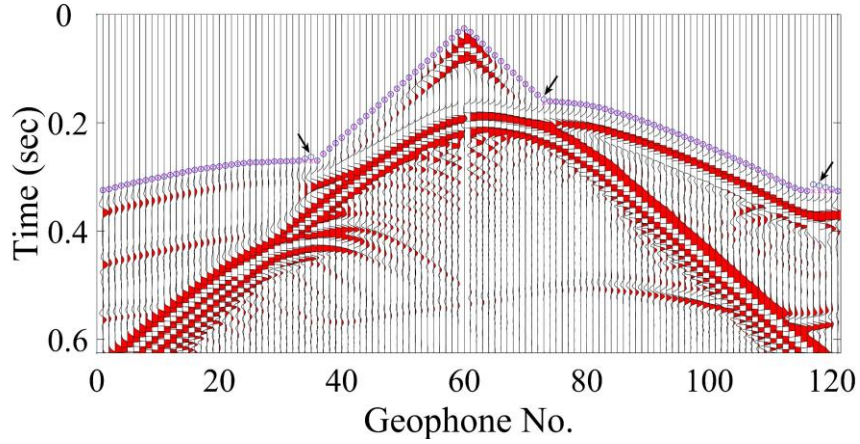


Figure 7.1: DBSCAN pickings (blue circles) compared with manual (magenta cross marker) for synthetic shot gather 60 without noise. Black arrows show mismatched locations of DBSCAN picking with manual picking.

The range of acceptable error for this synthetic data is $T/4 = \pm 8.33$ msec, where T is the dominant time-period of the source wavelet (source peak frequency is 30 Hz). The parameters selected for the noise-free data are given in table 1. For noise-free data, a small epsilon value with a small number of minimum points works well, while for noisy data, a small epsilon value with a large value of minimum points suits better clustering.

All the automatic pickings are matching well with the manual traveltimes except a few values are away from the manual pickings. There are three possible reasons for this deviation. The first and second causes are due to the slope of the first break. The algorithm works fine when the slope of the first break is decreasing i.e., each first break traveltimes is more in time than the previous one. A sudden slope change can occur due to static errors in addition to the subsurface geology. If the slope of the events is not following the decreasing trend and, in the first case, if there are different slopes present in the window (Figure 7.2a), the stacked energy becomes more dispersed/uneven (Figure 7.2b). A good picking is made by the algorithm when the peaks and troughs become parallel to each other resulting in maximum stacked energy for an optimum velocity

value. Otherwise, for each velocity step, no value would result in maximum stacked energy due to the uneven slope of the traces and most stacked energy will be scattered and this leads to a picking value deviated from the actual traveltime (Figures 7.2d).

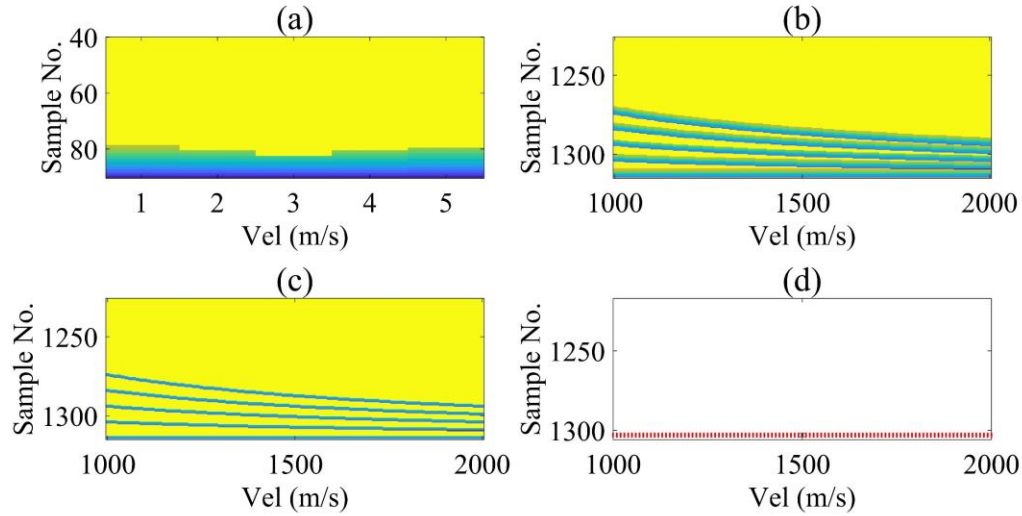


Figure 7.2: a) Non-linear slope trend in the window, b) Stacked LMO image shows scattered energy c) Stacked LMO image after applying the threshold, d) In the end, it gives deviated first break picking value.

In the second case, there is a single slope of the events, but its increasing (Figure 7.3a, black arrow). In this scenario, no data came yet from the first trace (Figure 7.3a), but the cluster has formed (Figure 7.3d) due to the other trace's data in the window. The picked centroid of the cluster is a good pick for the last trace in the window but not for the first trace. However, the algorithm is not sensitive to the slope of the events, and it gives the picked value to the first trace in the window which goes wrong in this scenario. This problem of increasing slope may be fixed if we identify the sloping trend in a window as either increasing or decreasing and assign the picking value accordingly.

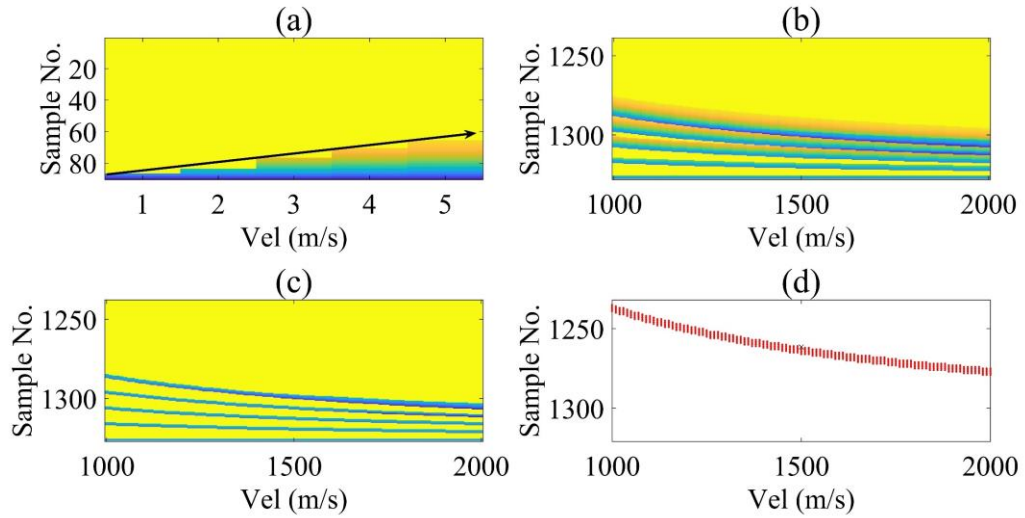


Figure 7.3: a) Increasing linear slope trend in the window, b) Stacked LMO image c) Stacked LMO image after applying the threshold, d) In the end, it gives the wrong first break picking value.

The third reason for the first arrival travetime picking deviation from the accurate value could be explained due to the diminishing trace window, where the remaining traces become less than the length of the selected trace window, and the amplitude values in the stacked gather decrease which causes the decrease in picking accuracy. This error could be removed by flipping the trace window for these traces so that the stacked gather has enough amplitude values to make a correct picking. Or, the last traveltimes picking values can be extrapolated using the previously picked first breaks.

The deviated pickings (Figure 7.1) at the 35-36 and 69 traces are due to the first two reasons explained above, and the deviation of DBSCAN pickings from the manual picking in the end traces of the shot gather 60 is due to all three explanations mentioned above. These deviations will continue to come in synthetic and field results described below depending upon the SNR.

7.1.2 SYNTHETIC RESULTS WITH HIGH-LEVEL RANDOM NOISE AND NO RESTRICTED SCANNING WINDOW

As stated earlier, DBSCAN suffers when it deals with noisy data. I tested the code by adding random noise equal to 10, 20, 30, 40, and 50% of the maximum amplitude of each trace. The added noises are gradually increased from 0% of the maximum amplitude at the zero-offset location to 50% of the maximum amplitude at the maximum source-receiver offset. The result is shown in Figure 7.4 only for 50% noise added to the data.

DBSCAN has made a lot of errors in the pickings as shown in Figure 7.4 (blue points). The first reason to struggle in many instances is the low SNR. This is because the first arrivals are masked by the noise present in the data and the algorithm picking parameters (e.g., epsilon and minPts) can get satisfied to accept a cluster as the picking. The second reason for the wrong picking by the DBSCAN is that it is scanning the whole data for the first break picking from the first sample to the last. However, as we know, the first break generally follows an increasing trend in time, so, there is no need to scan the noisy part to find the first break in a dataset. Therefore, to improve the results, I proposed two ideas.

- a) Restricting the scanning window for picking
- b) Extrapolating the missing picks

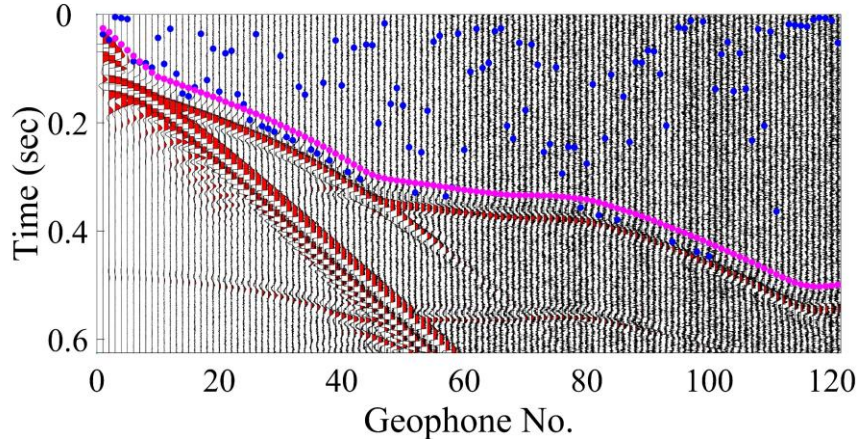


Figure 7.4: DBSCAN picking without restricted scanning window and extrapolation with 50% random noise.

So far, the code is scanning the data starting from the first to the last sample in the data by following a specific size of the sample and trace window and terminates its operation when it finds a cluster/picking. After determining the first break value for a trace, it shifts the trace window by one trace and follows the predefined sample and trace window size to find the picking for the next trace and so on. In this way, DBSCAN picks many outlier values when scanning from the top to the bottom whenever the clustering parameters are satisfied. As we know, the first break has a time offset attribute and is present at specific time samples for each trace and usually follows an increase in time trend for the subsequent traces than earlier. So, the first idea is to restrict the scanning window to find the picking. This not only removed the outliers but also reduced the overall time for the algorithm to pick the traveltimes of the first arrivals (Figure 7.5). However, this will not solve the problem when DBSCAN misses any value in the restricted scanning window. So, to cope with this issue, I proposed the second idea to extrapolate for the missed picks using the probable slope of the first break.

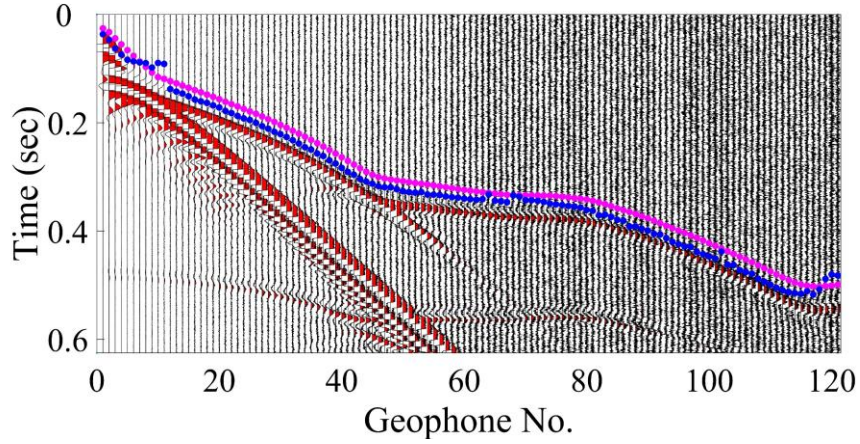


Figure 7.5: DBSCAN picking with restricted scanning window and extrapolation with 50% random noise for synthetic shot gather 1.

Using the restricted scanning window and extrapolation for missed pickings, usually, DBSCAN picks in the area of direct and near-offset arrivals accurately to a great extent than in the far-offset. This is because the SNR is high for direct and near-offset arrivals. For the initial guess of the synthetic data with 50% random noise (Figure 7.5), I restricted the scanning window to 90 samples up and 60 samples down for the next pick. For example, let's assume that my last picking is at 500 msec, then the next scanning window for the picking will start from 410 msec and ends up at 560 msec. When DBSCAN misses a value, it extrapolates its value by using a predefined number of samples. The number of samples to add to the last value for the extrapolation is defined by observing the slope of the first break. In this way, I could improve the results for DBSCAN in terms of eliminating the outlier values and reducing the time to pick the first arrivals by the code as you can see in Figure 7.5.

Consequently, by using this initial guess, I can define the most probable area where the actual picking is located around the DBSCAN picking. After that, I applied super-virtual refraction interferometry to enhance the SNR of the first breaks.

7.1.3 SYNTHETIC RESULTS WITH HIGH-LEVEL RANDOM NOISE AND RESTRICTED SCANNING WINDOW

7.1.4 SYNTHETIC RESULTS

I tested the code for shot gathers 29 and 69. The added noises are gradually increased from 0% of the maximum amplitude at the zero-offset location to 50% of the maximum amplitude at the maximum source-receiver offset. Hence, the first breaks at the far offset traces are not visible nor pickable (Fig. 7.6a and 7.6f).

To apply the DBSCAN method, I used the ϵ , minPts, etc. values listed in Table 1. The DBSCAN is applied to generate an initial picking guess (Figure 7.6a and 7.6f, blue lines). In the presence of random noise, although the initial guess pickings are far away from the true manual picking (Fig. 7.6a and 7.6f, magenta lines), this information is still very useful.

For a successful application of SVI, a muting window around the probable first break area is necessary to avoid artifacts in the output. However, accurate muting window determination is a challenging task in low SNR and often inhibits manual input. To automate this process, I used the DBSCAN initial guess to define the muting window for the SVI (Fig. 7.6b and 7.6g). The muting window size is around four periods (120 samples) up and four periods down from the initial guess. The initial guess has been smoothed using the moving average method to eliminate any outliers.

Then, SVI is applied to the muted data to generate the super-virtual traces with improved SNR of the first breaks (Fig. 7.6c and 7.6h). We can note that the first arrivals SNR has

been improved a lot at the far offsets, particularly, where it was very low before. Finally, I applied the DBSCAN again on the SVI output to pick the first arrivals accurately (Fig. 7.6d and 7.6i, blue line) and compared it with the manual (Figure 7.6d and 7.6i, magenta line). We can observe that the final automatic picking determined by the DBSCAN after applying SVI for both shots matched very well with the manual picking. The final picks are smoothed, and a constant shift of the samples up is applied after the SVI to reduce the error from the manual picking and compensate for the SVI delays.

In Fig. 7.6e and 7.6j, the red and blue lines show the difference between manual and DBSCAN picking before and after applying SVI, respectively. A constant shift up to the smoothed initial guess has been applied to reduce the error from the manual. The green dashed lines are the threshold values for the acceptable error of ± 8.33 msec (peak frequency = 30 Hz, hence, peak period/4 = 8.33 msec). The automatic first arrival's traveltimes picking by DBSCAN before using SVI for shots 29 and 69 were almost 76% and 73% within the threshold, respectively. Whereas, after applying SVI, it improved drastically and all the DBSCAN picking falls within the allowable error.

The cross-correlation method has been applied to both shots to pick the first break traveltimes. After smoothing and applying a constant shift up to the first break traveltimes picked by the cross-correlation method to minimize the error from the manual, the results show that 99% and 100% of the pickings for shots 29 and 69, respectively, are within the threshold compared with the manual.

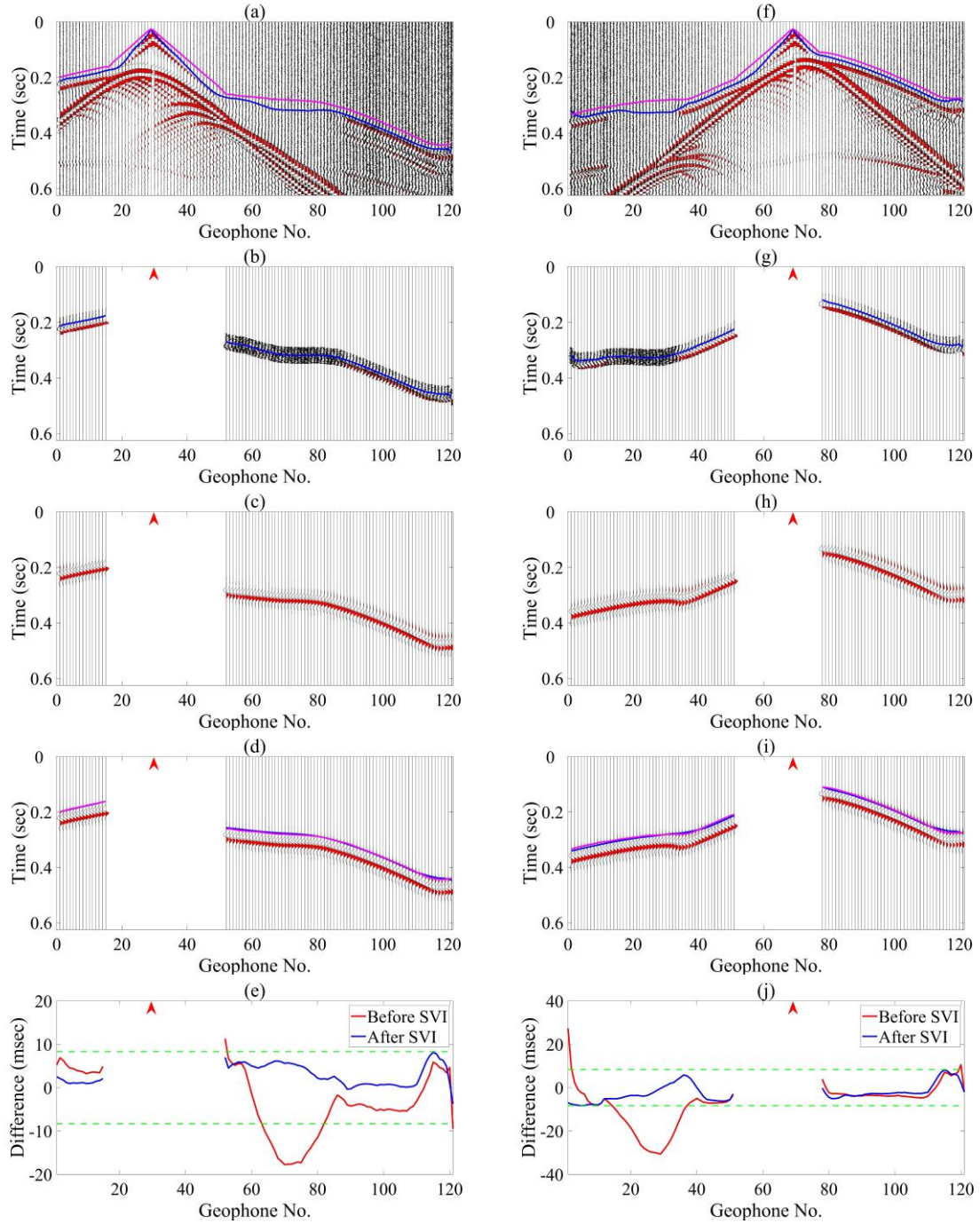


Figure 7.6: a) and f) show shot gather 29 and 69 with added random noises, magenta and blue lines showing manual and DBSCAN initial guess, respectively, b) and g) show the muting around the expected location of the first arrivals using initial guess from DBSCAN, c) and h) showing the enhanced SNR after SVI, d) and i) show the manual picking (magenta line) match with the DBSCAN picking after SVI (blue line), e) and j) presenting the difference between manual and DBSCAN first arrivals traveltime pickings before applying SVI (red line) and after applying SVI (blue line), respectively. The dashed green lines display the threshold values. The red arrow shows the shot location.

Following the same methodology, in Figure 7.7, the histogram shows the differences between manually picked first break traveltimes and DBSCAN's picking after applying SVI for all 121 shot gathers. It can be observed that most of the differences lie within the allowable error (green dashed lines) of ± 8.33 msec which is the $T/4$, where T is the source wavelet dominant time-period and for synthetic data, $T \sim 33$ msec. In detail, for 121 shot gathers, there were 14641 traces by having 121 traces in each of 121 shot gathers. A constant shift up is applied to all the trace's traveltimes. This shift, for each shot, has been determined automatically by calculating the mean difference between manual and final DBSCAN's first breaks traveltimes pickings after applying SVI. After excluding the direct arrival's pickings, 11513 trace's first breaks were available for comparison. The automatic DBSCAN's first break traveltimes picking accuracy came to approximately 98.4% by picking 11324 traces out of 11513 trace's traveltimes within the threshold.

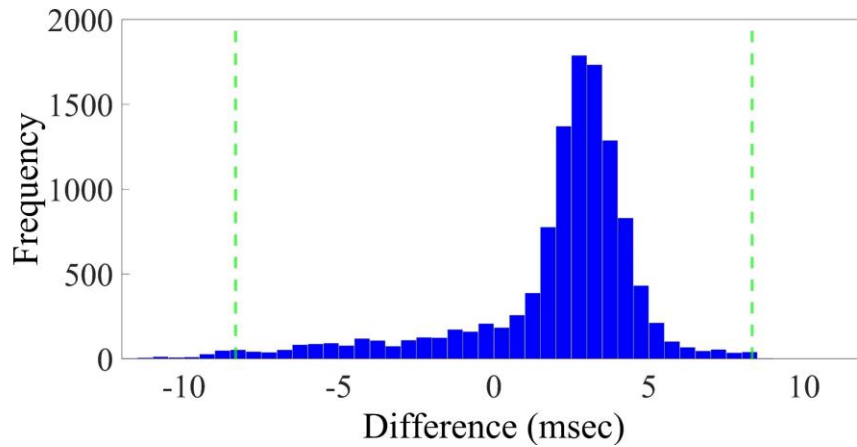


Figure 7.7: Histograms comparing traveltime differences between manual and final DBSCAN's automatic first break's picking after SVI for all shot gathers. Green dashed lines show the threshold values of ± 8.33 msec.

7.2 ITHRA FIELD DATA RESULTS

I applied the same methodology to the Ithra field data. As examples, the shot gathers 25 and 100 are shown in Figures 7.8a and 7.8g with the manual pickings (magenta line). The manual picking is done on the raw data. The noise level in the data is severe in many of the shots as observed. DBSCAN is applied to get an initial guess (Figure 7.8a and 7.8g, blue line) of the data using the parameters given in Table 1. The initial guess has been smoothed using the moving average method to eliminate any outliers. Due to a high level of noise in the data, the DBSCAN was able to give a good guess till trace 73 and 58 for shots 25 and 100, respectively (Figure 7.8b, and 7.8h). To define the window for all traces for the SVI application, the guess was extrapolated (from 73-120, and 58-01) in all the shots by following the slope of the expected first breaks. Then, a band pass filter (low cut=25 Hz, low pass=35 Hz, high pass=80 Hz, high cut=100 Hz) has been employed to suppress the low and high-frequency noises. Then, using the initial guess, the muting window is defined as ~ 1.5 periods up and down the initial guess shown in Figures 7.8c and 7.8i. The results after applying SVI are shown in Figures 7.8d and 7.8j, re-picked by DBSCAN (Figures 7.8e and 7.8k) and compared with the manuals shown in Figures 7.8f and 7.8l. Again, the picks are smoothed, and a constant shift of the samples up is applied. Figures 7.8f and 7.8l show a very accurate match of the DBSCAN pickings (blue lines) with the manuals (magenta lines).

Further, for shot 25, after SVI, the algorithm picking accuracy was enhanced by 23 traces (Figure 7.8f, shown by black arrow) which is $\sim 54\%$ increase in the traveltimes picking as compared to the manual ones, whereas, for shot 100, this enhancement reached to 62%

by picking 27 traces more than the manuals (Figure 7.8l, shown by black arrow). For each shot, manual picking was possible for ~42 traces only.

Although DBSCAN could not pick all the traces traveltime accurately for both shots even after SVI due to the severe presence of random noises in the far offsets, we can observe a good SNR enhancement after SVI for the far offset traces of shot 100 (Figure 7.8l).

In Figures 7.9a and 7.9b, the red and blue lines show the difference between manual and DBSCAN picking before and after applying SVI, respectively. The green dashed lines are the threshold values for the acceptable error of ± 8.33 msec. A constant shift up to the smoothed initial guess has been applied to reduce the error from the manual and compensate for the SVI delays. Results show that for both shots gathers, before and after SVI, all DBSCAN first break traveltime pickings are within the threshold (Figure 7.9a and 7.9b) due to the good SNR near the shot location.

Again, the cross-correlation method (Senkaya and Karsli, 2014) has been tested on both shots. After smoothing, a constant shift up is applied to minimize the error from the manual. The comparison of the first break picking with the manual ones shows ~100% match for both shots. In addition, the algorithm picked ~60% and ~58% traces first break traveltime accurately more than the manuals for shots 25 and 100, respectively. Nevertheless, it is worth mentioning that, although, the cross-correlation method results are in comparison with our methodology, however, the cross-correlation technique takes a lot of time as being a semi-automatic approach whereas our technique takes just 4-5 hours to automatically pick the first breaks of low SNR data.

.

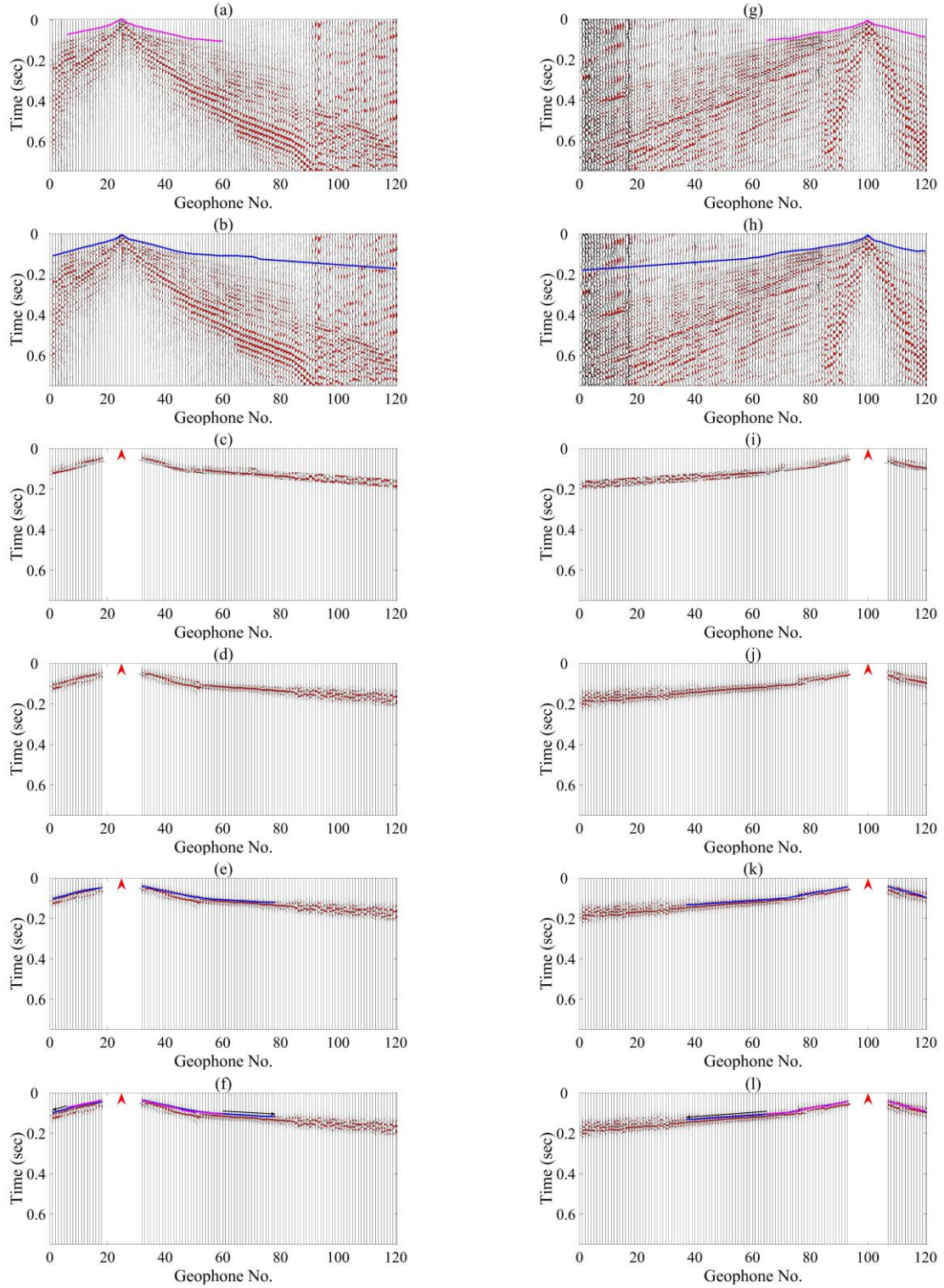


Figure 7.8: a) and g) show Ithra field data shot gathers 25 and 100 with manual pickings (magenta line), respectively, b) and h) showing the DBSCAN initial guess (blue line), c) and i) show the muting around the

expected location of the first arrivals using the initial guess, d) and j) showing the enhanced SNR after SVI, e) and k) show the re-DBSCAN pickings after SVI (blue line), f) and l) representing the matching of manual picking (magenta line) with DBSCAN after SVI (blue line). The black arrow shows enhanced picking than the manual. The red arrow shows the shot location.

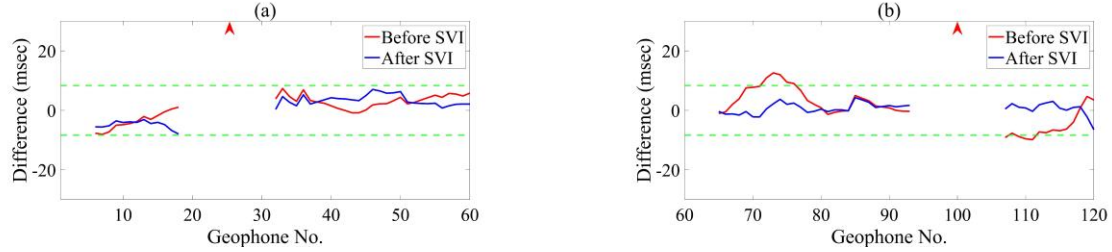


Figure 7.9: a) and b) show the differences between manual and DBSCAN first arrivals traveltime pickings for Ithra shot gather 25 and 100 before applying SVI (red line) and after applying SVI (blue line), respectively.

In Figure 7.10, the histogram shows the differences between manually picked first break traveltimes and DBSCAN's picking after applying SVI for all 120 shot gathers. It shows that most of the differences between raw Ithra data manual pickings and DBSCAN's automatic first arrival's traveltimes pickings after applying SVI lie within the allowable error (green dashed lines) of ± 8.33 msec. The allowable error, again, is calculated using the source wavelet dominant time-period T . In detail, for 120 shot gathers, there were 14400 traces by having 120 traces in each of 120 shot gathers. A constant shift up/down is applied to all the trace's traveltimes. This shift, for each shot, has been determined automatically again by calculating the mean difference between manual and final DBSCAN's first breaks traveltimes pickings after applying SVI. After excluding the direct and manually unpickable arrival pickings, 5925 trace's first breaks were available for comparison. The automatic DBSCAN's first break traveltimes picking accuracy came to approximately 86.4% by picking 4516 traces out of 5925 trace's traveltimes within the threshold.

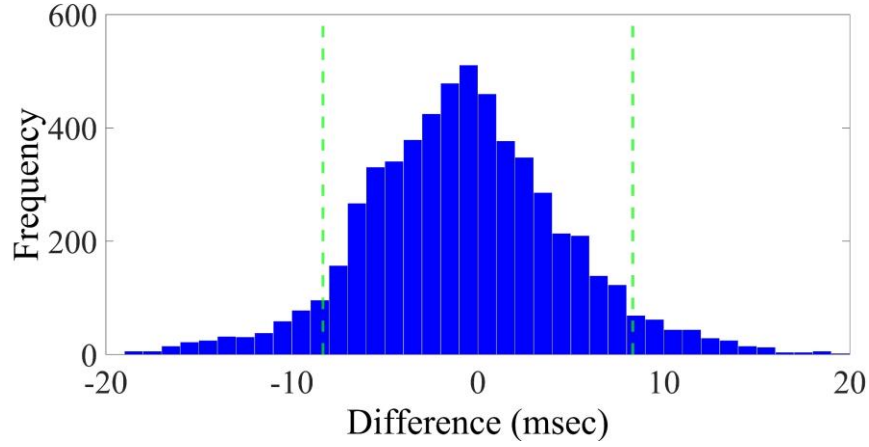


Figure 7.10: Histograms comparing traveltime differences between manual and final DBSCAN's automatic first break's picking after SVI for all shot gathers. Green dashed lines show the threshold values of ± 8.33 msec.

7.3 YILMAZ FIELD DATA RESULTS

To benchmark the results of our method with the others, I used Yilmaz's shot gathers 6 and 25. For the SVI application, we need many shots of data to enhance the SNR, therefore, I could not apply SVI on these shots, however, as we can observe in Figures 6.7 and 6.8, the shots have good SNR, so, I tested only the DSBCAN to pick the first break without applying SVI for the cause of comparing the results with other established methods. To initiate the DBSCAN, the parameters selected are listed in Table 1, and the results are shown in Figures 7.11a and 7.11b after smoothing the initial guess and making a constant shift up/down to decrease the error between manual and DBSCAN pickings.

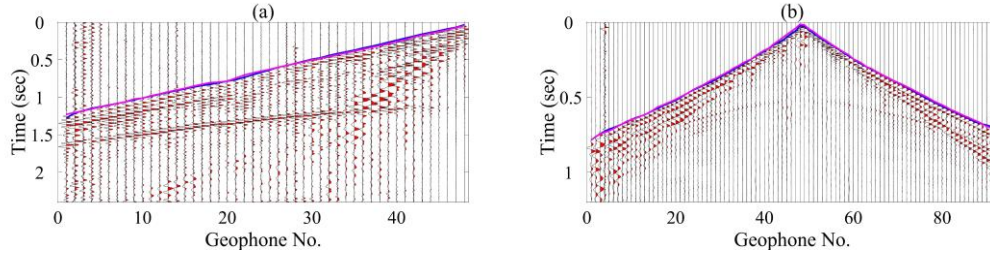


Figure 7.11: Matching of initial guess DBSCAN picking (blue line) with Manual (magenta line), (a) for shot gather 6, and (b) for 25. For shot gather 25, the last 4 traces are muted as they were just noises.

We can note that even the DBSCAN initial guess alone matched very well in both the shots with the manual picking (Figures 7.11a and 7.11b), for shots 6 and 25, respectively, ~91% and 99% first break traveltime pickings determined by DBSCAN lie within a threshold of ± 20 msec compared with the manuals (Figures 12a and 12b).

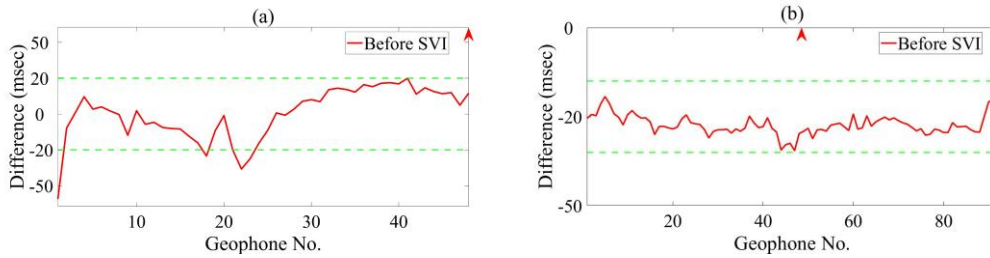


Figure 7.12: a) and b) showing the differences between manual and initial guess of DBSCAN first arrivals traveltime pickings for shot gather 25 and 100, respectively.

Many researchers tested these two open-source shots to show the accuracy of first break pickings for their methods e.g., τ -p transform method (Mousa and Al-Shuhail, 2012) used the Yilmaz shot record 6 while a mix of cross-correlation technique based on the Coppens method (mixed Coppens' method) (Mousa et al., 2011), digital image segmentation using projection onto convex sets (POCS) (Mousa et al., 2011) and pickings based on the maximum difference between adjacent points of the envelope

(MDPE) (Al-Mashhor et al., 2019) used shot gather 6 and 25. These all methods compared their results with the manual pickings. For the τ -p method using shot record 6, the picking accuracy was around 88%. For shot gather 6 and 25, using the mixed Coppens' method, the accuracy came to approximately 46% and 71%, whereas POCS improved these results to 58% and 77%, respectively. The MDPE method further enhanced the results to more than 85% for both shots. All these method's pickings were accepted within a ± 20 msec tolerance of the error from the manual ones except the MDPE where the tolerance was ± 0.1 sec of the manual. However, the performance of our methodology, even without applying the SVI, is better and shows almost 91% and 99% picking accuracy within ± 20 msec of the manuals for shots 6 and 25, respectively.

7.4 DISCUSSION

The unsupervised machine learning technique DBSCAN alone gives good results in the case of noise-free data however suffered in the noise presence. In this research, DBSCAN along with super-virtual refraction interferometry has been employed to automatically pick the first break of the seismic data. The SVI has been used to deal with the low SNR data.

The proposed methodology works with the raw data and requires no type of preprocessing except a basic step e.g., a bandpass filter to suppress the random noises. As we know, many methods have room for improvement when the SNR of the data is low. By using the SVI, the method can be used to pick the first break traveltimes of low SNR

data. To apply a successful SVI, a muting window around the probable first break area is essential, however, this window defining can be very difficult and often inhibit manual input. The results from the synthetic and field data show that in the presence of high-level random noise, using the restricted window approach to search for the picking and extrapolating the missing picks, the DBSCAN gives us an initial guess of the picking. This initial guess guides the muting window for the SVI application to enhance the SNR of the far offsets. So, by using DBSCAN initial guess, I automated the muting window operation for the SVI process. A small muting window results in better enhancing the SNR of the data. In addition, the muting window should be defined as all first arrivals except those coming from the bottom of the same layer that reach within two to three periods of the estimated first breaks time are muted. If for any reason, the refractions from different subsurface layers are not identifiable, the application of SVI will not result in good SNR improvement.

After SVI, due to violating the assumptions of the reciprocity equations, artifacts (side lobes) are generated in the super-virtual refractions. To get rid of these side lobes or ringyness, muting the small amplitudes worked well for the final DBSCAN first break picking.

In the presence of a very high level of random noise, the DBSCAN may not be able to give a good initial guess for all the traces in a shot gather. In that case, keeping the good guess picking values and extrapolating the remaining bad values by using the slope of the

first break is a good approach to define an appropriate muting window for the SVI application.

The method is easy to apply as there are only a few parameters that need to be determined to start the methodology and are given in Table I. Epsilon and minPts for the DBSCAN clustering are determined using the hit and trial method. The test shows that the trace window range should be always 5-10 irrespective of the random noise level in the data, whereas the sample window should be two-three periods. The velocity range for the LMO image should be ± 500 m/sec from the velocity of the refraction layer. The refraction's layer velocity is determined from the first break slope which is straightforward. A small velocity step is suggested for the stacking in the LMO image. A threshold value of 0.5 usually works better and helps a lot in the clustering procedure by muting the low amplitude values of the LMO image.

From the synthetic data, two shots' automatic traveltimes picking of the first break has been picked and compared with the manual picking. The comparison shows that the automatic first arrival's traveltimes picking by DBSCAN after using SVI for shots 29 and 69 increased by 24% and 27%. Further, the cross-correlation method has been applied to both shots to pick the first break traveltimes and it confirms the same results as compared with re-DBSCAN picking, however, cross-correlation is a semi-automatic approach versus our automatic methodology. Further, for all synthetic shot gathers after applying SVI, the histogram confirms 98.4% of the DBSCAN's final first break's traveltimes pickings within the threshold values compared with the manual pickings.

For the Ithra field data, two shots containing a high level of low and high-frequency noise results are shown. These two shots' first break traveltimes were picked using manual and cross-correlation methods. For the Ithra field data, the SVI enhanced the SNR of the data as we observed. For shot 25, re-DBSCAN could pick 65 traces' first break traveltimes in total out of 106 first arrivals whereas the remaining far offset traces are very challenging to pick the first arrivals for any method as the SNR is too low to differentiate the first break even by the human eye (Figure 7.8f). For shot gather 100, re-DBSCAN could pick 69 traces first break traveltimes out of 106, however, although re-DBSCAN could not pick the remaining traces due to severe random noises, yet the first break can be observed for the traces after SVI which shows the robustness of the SVI application. Moreover, with the enhanced picking after SVI, the traveltime tomogram would have more depth of penetration in the subsurface which is an important addition. For all Ithra field data shots gathers after applying SVI, the histogram confirms 86.4% of the DBSCAN's final first break's traveltime pickings within the threshold values compared with the raw manual pickings.

Further, the results of the methodology can be improved if we use other noise suppression techniques along with a bandpass filter.

A big advantage of DBSCAN is that it is an unsupervised machine-learning technique. While the supervised machine learning methods greatly depend on the input training dataset, the unsupervised machine learning methods (e.g., clustering) are much more

flexible and require no training input. Nevertheless, unsupervised machine learning algorithms always require quality control to ensure the best results.

So far, the methodology worked well in the presence of random noises but is not applicable in the presence of coherent noises. Lastly, the methodology works accurately in the case of zero or minimum phase wavelet but needs to be tested on the mixed phase wavelet.

CHAPTER 8

FUTURE PROSPECTS AND CONCLUSION

DBSCAN is an unsupervised machine learning approach that does not require any labeled data, whereas supervised machine learning algorithms require labeled data as a primary input, thus making the DBSCAN much more flexible in its use. DBSCAN along with the SVI approach has been tested with zero-phase and minimum-phase wavelets. It would be a good research direction to test this practice with mixed-phase wavelets.

Unsupervised machine learning techniques can be used to create the training dataset for the supervised machine learning approaches (Hofmann, 2001). So, there is the possibility to use DBSCAN along with SVI to pick the first break and provide that picked data as input for the supervised machine learning techniques. This would reduce the time to do the manual picking to make the labeled data for the supervised methods to pick the first break. Our methodology can also be used along with the supervised methods to improve the results further.

In this research, I focused on using DBSCAN for first arrival traveltime picking in the presence of random noise. Synthetic and Ithra field examples show that DBSCAN alone does not work accurately in case of random noises and low SNR data. However, the low-quality DBSCAN picking, using the restricted window and extrapolation for missed picks, can be used as an initial guess to find the expected location of the first breaks. This initial guess can be used to mute around the expected location of the first breaks which is a critical input for SVI. After applying SVI, the DBSCAN method is applied again to

accurately pick the first breaks, which is improved drastically. The results of both synthetic and Ithra field data have been compared with the manual and cross-correlation picking methods. For the synthetic data compared with the manual, the methodology confirms an accurate automatic first break picking of 100% first break traveltimes for shots 29 and 69, whereas 98.4% for all 121 shot gathers. For the Ithra field data compared with the manual, two shots with high random noises have been picked using the proposed approach. Results show that the automatic first break not only matches with the manual picking but also the reliable picking increased by 54% and 62% for shots 25 and 100, respectively. The comparison with the cross-correlation method picking confirms almost comparable results. Whereas, for all 120 shot gathers, the final DBSCAN's first break's traveltime pickings compared against raw manual pickings gives almost 86.4% accuracy. Further, the results of first break picking using only DBSCAN for Yilmaz shot gather 6 and 25 have been compared with the manual, mixed Coppens', τ -p, POCS, and MDPE methods. The comparison shows that our methodology can pick the first break far better than these methods even without applying the SVI. The success of this methodology is very promising as I am using an unsupervised machine learning technique that doesn't require any labeling of data.

REFERENCES

- Abdullah, Abobakr Al-Mashhor. 2020. "First arrival travel time picking using two different methods: MDPE and DBSCAN." MS Geophysics, King Fahd University of Petroleum and Minerals.
- Ahmadi, Mohammad Ali, and Zhangxing Chen. 2019. Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petrophysical logs: *Petroleum* **5**, no. 3, 271-284.
- Akram, Jubran, and David W Eaton. 2016. A review and appraisal of arrival-time picking methods for downhole microseismic data: *Geophysics* **81**, no. 2, KS71-KS91.
- Al-Anezi, Ghunaim T, Abdullah M Al-Amri, and Haider Zaman. 2012. Investigation of the weathering layer using seismic refraction and high-resolution seismic reflection methods, NE of Riyadh city: *ASEG Extended Abstracts* **2012**, no. 1, 1-4.
- Al-Hagan, Ola, Sherif M Hanafy, and Gerard T Schuster. 2014. Iterative supervirtual refraction interferometry: *Geophysics* **79**, no. 3, Q21-Q30.
- Al-Mashhor, Abdullah A, Abdullatif A Al-Shuhail, Sherif M Hanafy, and Wail A Mousa. 2019. First Arrival Picking of Seismic Data Based on Trace Envelope: *IEEE Access* **7**, no., 128806-128815.
- Al-Shuhail, Abdullatif A. 2015. Improving automatic first-arrival picking by supervirtual interferometry: examples from Saudi Arabia: *Arabian Journal of Geosciences* **8**, no. 10, 8731-8740.
- Al-Shuhail, Abdullatif A, and Abdulaziz M Al-Shaibani. 2013. Characterization of Sabkha Jayb Uwayyid, eastern Saudi Arabia using seismic refraction profiling: *Arabian Journal of Geosciences* **6**, no. 3, 845-855.
- Al-Shuhail, Abdullatif, SanLinn I Kaka, and Michael Jervis. 2013. Enhancement of passive microseismic events using seismic interferometry: *Seismological Research Letters* **84**, no. 5, 781-784.
- Aleardi, Mattia. 2015. Seismic velocity estimation from well log data with genetic algorithms in comparison to neural networks and multilinear approaches: *Journal of Applied Geophysics* **117**, no., 13-22.
- Alkhalifah, Tariq, and Ilya Tsvankin. 1995. Velocity analysis for transversely isotropic media: *Geophysics* **60**, no. 5, 1550-1566.
- Almadani, Sattam, Elkhedr Ibrahim, Kamal Abdelrahman, Abdulaziz Al-Bassam, and Awad Al-Shmrani. 2015. Magnetic and seismic refraction survey for site investigation of an urban expansion site in Abha District, Southwest Saudi Arabia: *Arabian Journal of Geosciences* **8**, no. 4, 2299-2312.
- Araya-Polo, Mauricio, Taylor Dahlke, Charlie Frogner, Chiyuan Zhang, Tomaso Poggio, and Detlef Hohl. 2017. Automated fault detection without seismic processing: *The Leading Edge* **36**, no. 3, 208-214.
- Baer, M, and U Kradolfer. 1987. An automatic phase picker for local and teleseismic events: *Bulletin of the Seismological Society of America* **77**, no. 4, 1437-1445.

- Bell, Anthony J, and Terrence J Sejnowski. 1995. An information-maximization approach to blind separation and blind deconvolution: *Neural computation* **7**, no. 6, 1129-1159.
- Benediktsson, Jon A, Philip H Swain, and Okan K Ersoy. 1990. Neural network approaches versus statistical methods in classification of multisource remote sensing data: *IEEE Transactions on Geoscience and Remote Sensing*, no.
- Bharadwaj, Pawan, Gerard Schuster, Ian Mallinson, and Wei Dai. 2012. Theory of supervirtual refraction interferometry: *Geophysical Journal International* **188**, no. 1, 263-273.
- Bharadwaj, Pawan, Gerard T Schuster, and Ian Mallinson. 2011. "Super-virtual refraction interferometry: Theory." In *SEG Technical Program Expanded Abstracts 2011*, 3809-3813. Society of Exploration Geophysicists.
- Bharadwaj, Pawan, Xin Wang, Gerard Schuster, and Kirk McIntosh. 2013. Increasing the number and signal-to-noise ratio of OBS traces with supervirtual refraction interferometry and free-surface multiples: *Geophysical Journal International* **192**, no. 3, 1070-1084.
- bin Waheed, Umair, Saleh Al-Zahrani, and Sherif M Hanafy. 2019. "Machine learning algorithms for automatic velocity picking: K-means vs. DBSCAN." In *SEG Technical Program Expanded Abstracts 2019*, 5110-5114. Society of Exploration Geophysicists.
- Boschetti, Fabio, Mike D Dentith, and Ron D List. 1996. A fractal-based algorithm for detecting first arrivals on seismic traces: *Geophysics* **61**, no. 4, 1095-1102.
- Calderón-Macías, Carlos, Mrinal K Sen, and Paul L Stoffa. 1998. Automatic NMO correction and velocity estimation by a feedforward neural network: *Geophysics* **63**, no. 5, 1696-1707.
- Çelik, Mete, Filiz Dadaşer-Çelik, and Ahmet Şakir Dokuz. 2011. Anomaly detection in temperature data using DBSCAN algorithm: 2011 international symposium on innovations in intelligent systems and applications, IEEE, 91-95.
- Chi-Durán, Rodrigo, Diana Comte, Marcos Díaz, and Jorge F Silva. 2017. Automatic detection of P-and S-wave arrival times: new strategies based on the modified fractal method and basic matching pursuit: *Journal of Seismology* **21**, no. 5, 1171-1184.
- Coppens, F. 1985. First arrival picking on common-offset trace collections for automatic estimation of static corrections: *Geophysical Prospecting* **33**, no. 8, 1212-1231.
- Cox, Mike. 1999. Static corrections for seismic reflection surveys. Society of Exploration Geophysicists.
- Dong, Shuqian, Jianming Sheng, and Gerard T Schuster. 2006. "Theory and practice of refraction interferometry." In *SEG Technical Program Expanded Abstracts 2006*, 3021-3025. Society of Exploration Geophysicists.
- Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise: kdd226-231.

- Fajaryanti, R, HM Manik, and C Purwanto. 2018. Application of multichannel seismic reflection method to measure temperature in Sulawesi Sea: IOP Conference Series: Earth and Environmental Science, IOP Publishing, 012044.
- Gan, G, C Ma, and J Wu. 2007. Data clustering: Theory, algorithms, and applications. vol. 20: Siam: Gershenson, C.(2004). Self-organizing traffic lights. arXiv: <http://arxiv.org/abs/nlin/0411066>, no.
- Gelchinsky, B, and V Shtivelman. 1983. Automatic picking of first arrivals and parameterization of traveltimes curves: Geophysical Prospecting **31**, no. 6, 915-928.
- Hamada, GM, and MA Elshafei. 2010. Neural network prediction of porosity and permeability of heterogeneous gas sand reservoirs using NMR and conventional logs: Nafta **61**, no. 10, 451-465.
- Han, Song, Yujin Liu, Yubing Li, and Yi Luo. 2021. First Arrival Traveltime Picking Through 3-D U-Net: IEEE Geoscience and Remote Sensing Letters **19**, no., 1-5.
- Hanafy, Sherif M, and Ola Al-Hagan. 2012. Super-virtual refraction interferometry: an engineering field data example: Near Surface Geophysics **10**, no. 5, 443-449.
- Hanafy, Sherif M, Ola AlHagan, and Feras Al-Tawash. 2011. Super-virtual refraction interferometry: Field data example over a colluvial wedge: 2011 SEG Annual Meeting, OnePetro.
- Hatherly, PJ. 1982. A computer method for determining seismic first arrival times: Geophysics **47**, no. 10, 1431-1436.
- Helmy, Tarek, Anifowose Fatai, and Kanaan Faisal. 2010. Hybrid computational models for the characterization of oil and gas reservoirs: Expert Systems with Applications **37**, no. 7, 5353-5363.
- Hollander, Yaniv, Amin Merouane, and Orhan Yilmaz. 2018. Using a deep convolutional neural network to enhance the accuracy of first-break picking: 2018 SEG International Exposition and Annual Meeting, OnePetro.
- Huahui, Zeng, Wang Xiao, Feng Huiyuan, Yang Wei, and Liu Jie. 2018. Technology Method and Strategy for Massive High-density Seismic Data Merging Processing: International Geophysical Conference, Beijing, China, 24-27 April 2018, Society of Exploration Geophysicists and Chinese Petroleum Society, 1711-1714.
- Jia, Bohan, and Qiyu Huang. 2020. DE-CapsNet: A diverse enhanced capsule network with disperse dynamic routing: Applied Sciences **10**, no. 3, 884.
- Jia, Yongna, and Jianwei Ma. 2017. What can machine learning do for seismic data processing? An interpolation application: Geophysics **82**, no. 3, V163-V177.
- Kang, Tae-Seob. 2014. Seismic interferometry and ambient noise tomography in East Asia: Science reports of Niigata University (Geology), no.
- Karri, Nidhi Arora, Mohd Yousuf Ansari, and Anurag Pathak. 2018. Identification of Seismic Zones of India using DBSCAN: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), IEEE, 65-69.
- Kaufman, Leonard, and Peter J Rousseeuw. 2009. Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons.

- Kim, Yuji, and Nori Nakata. 2018. Geophysical inversion versus machine learning in inverse problems: *The Leading Edge* **37**, no. 12, 894-901.
- Li, Jing, Kai Lu, Sherif Hanafy, and Gerard Schuster. 2018. Two robust imaging methodologies for challenging environments: Wave-equation dispersion inversion of surface waves and guided waves and supervirtual interferometry+ tomography for far-offset refractions: *Interpretation* **6**, no. 4, SM27-SM37.
- Liao, Qingqing, Donald Kouri, Dip Nanda, and John Castagna. 2011. "Automatic first break detection by spectral decomposition using minimum uncertainty wavelets." In *SEG Technical Program Expanded Abstracts 2011*, 1627-1631. Society of Exploration Geophysicists.
- Liu, Yi, Kees Wapenaar, Joost van der Neut, and Børge Arntsen. 2014. Combining inter-source seismic interferometry and source-receiver interferometry for deep local imaging: 2014 SEG Annual Meeting, OnePetro.
- Lu, Kai, and Shihang Feng. 2018. Auto-windowed super-virtual interferometry via machine learning: A strategy of first-arrival traveltime automatic picking for noisy seismic data: SEG 2018 Workshop: SEG Maximizing Asset Value Through Artificial Intelligence and Machine Learning, Beijing, China, 17-19 September 2018, Society of Exploration Geophysicists and the Chinese Geophysical Society, 10-14.
- Lu, Ping, Matt Morris, Seth Brazell, Cody Comiskey, and Yuan Xiao. 2018. Using generative adversarial networks to improve deep-learning fault interpretation networks: *The Leading Edge* **37**, no. 8, 578-583.
- Ma, Yue, and Yi Luo. 2018. Automatic first-arrival picking with Reinforcement Learning: International Geophysical Conference, Beijing, China, 24-27 April 2018, Society of Exploration Geophysicists and Chinese Petroleum Society, 493-497.
- Mousa, Wail A, and Abdullatif A Al-Shuhail. 2012. Enhancement of first arrivals using the τ -p transform on energy-ratio seismic shot records: *Geophysics* **77**, no. 3, V101-V111.
- Mousa, Wail A, Abdullatif A Al-Shuhail, and Ayman Al-Lehyani. 2011. A new technique for first-arrival picking of refracted seismic data based on digital image segmentation: *Geophysics* **76**, no. 5, V79-V89.
- Nicolson, Heather, Andrew Curtis, Brian Baptie, and Erica Galetti. 2012. Seismic interferometry and ambient noise tomography in the British Isles: *Proceedings of the Geologists' Association* **123**, no. 1, 74-86.
- Nicolson, Heather Johan. 2011. Exploring the Earth's subsurface with virtual seismic sources and receivers, no.
- Pan, Shulin, Ziyu Qin, Haiqiang Lan, and José Badal. 2019. Automatic first-arrival picking method based on an image connectivity algorithm and multiple time windows: *Computers & Geosciences* **123**, no., 95-102.
- Peraldi, R, and A Clement. 1972. Digital processing of refraction data study of first arrivals: *Geophysical Prospecting* **20**, no. 3, 529-548.

- Qadrouh, AN, JM Carcione, M Alajmi, and MM Alyousif. 2019. A tutorial on machine learning with geophysical applications: *Bollettino di Geofisica Teorica ed Applicata* **60**, no. 3.
- Ross, Zachary E, Men-Andrin Meier, and Egill Hauksson. 2018. P wave arrival picking and first-motion polarity determination with deep learning: *Journal of Geophysical Research: Solid Earth* **123**, no. 6, 5120-5129.
- Saad, Omar M, and Yangkang Chen. 2021. Capsphase: Capsule neural network for seismic phase classification and picking: *IEEE Transactions on Geoscience and Remote Sensing* **60**, no., 1-11.
- Sabbione, Juan I, and Danilo Velis. 2010. Automatic first-breaks picking: New strategies and algorithms: *Geophysics* **75**, no. 4, V67-V76.
- Sabor, Kawtar, Damien Jougnot, Roger Guerin, Barthélemy Steck, Jean-Marie Henault, Louis Apffel, and Denis Vautrin. 2021. A data mining approach for improved interpretation of ERT inverted sections using the DBSCAN clustering algorithm: *Geophysical Journal International* **225**, no. 2, 1304-1318.
- Sabour, Sara, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules: *Advances in neural information processing systems* **30**, no.
- Samuel, A. L. 1959. Some studies in machine learning using the game of checkers: *IBM J. Res. Dev.* **3**, no. 3, 210–229. <https://doi.org/10.1147/rd.33.0210>.
- Schmidhuber, Jürgen. 2015. Deep learning in neural networks: An overview: *Neural networks* **61**, no., 85-117.
- Schuster, Gerard Thomas. 2009. *Seismic interferometry*. Vol. 1. Cambridge university press Cambridge.
- Senkaya, Mustafa, and Hakan Karsli. 2014. A semi-automatic approach to identify first arrival time: the cross-correlation technique (CCT): *Earth Sciences Research Journal* **18**, no. 2, 107-113.
- Senkaya, Mustafa, and Hakan Karsli. 2014. A semi-automatic approach to identify first arrival time: the cross-correlation technique (CCT): *Earth Sci. Res. J.* **18**, no. 2, 107-113.
- Vaezi, Y, and K DeMeersman. 2014. Supervirtual S-wave refraction interferometry for converted wave statics and near-surface S-wave velocity model building: 76th EAGE Conference and Exhibition 2014, European Association of Geoscientists & Engineers, 1-5.
- WANG, Fu-Yun, Cheng-Bin ZHAO, Shao-Ying FENG, Ji-Fa JI, Xiao-Feng TIAN, Xue-Qiang WEI, Yi-Qing LI, Ji-Chang LI, and Xin-Sheng HUA. 2015. Seismogenic structure of the 2013 Lushan MS7. 0 earthquake revealed by a deep seismic reflection profile: *Chinese Journal of Geophysics* **58**, no. 9, 3183-3192.
- Wang, Li-Xin, and Jerry M Mendel. 1992. Adaptive minimum prediction-error deconvolution and source wavelet estimation using Hopfield neural networks: *Geophysics* **57**, no. 5, 670-679.

- Wibisono, S, MT Anwar, A Supriyanto, and IHA Amin. 2021. Multivariate weather anomaly detection using DBSCAN clustering algorithm: Journal of Physics: Conference Series, IOP Publishing, 012077.
- Yalcinoglu, L, and C Stotter. 2018. Can Machines Learn To Pick First Breaks As Humans Do?: First EAGE/PESGB Workshop Machine Learning, European Association of Geoscientists & Engineers, 1-3.
- Yilmaz, Öz. 2001. Seismic data analysis: Processing, inversion, and interpretation of seismic data. Society of exploration geophysicists.
- Yuan, Sanyi, Jiwei Liu, Shangxu Wang, Tieyi Wang, and Peidong Shi. 2018. Seismic waveform classification and first-break picking using convolution neural networks: IEEE Geoscience and Remote Sensing Letters **15**, no. 2, 272-276.
- Zeiler, Matthew D., and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks, Cham, Springer International Publishing, 818-833.

VITAE

Name: Muhammad Awais

Nationality: Pakistani

Date of Birth: 29/07/1995

Email: mawais.com@gmail.com

Address: Room 111, Building 819, Student Housing, KFUPM,
Dhahran, 31261, Saudi Arabia

Village Bal Shahan, Post Office Dhamthal, 51660, Tehsil
Zafarwal, District Narowal, Punjab, Pakistan

Academic Background: University of Engineering and Technology, Lahore
B.S. in Geological Engineering,
June 2018.

King Fahd University of Petroleum and Mineral (KFUPM),
M.Sc. in Geophysics,
October 2020.

Publications: Awais, Muhammad, Sherif M. Hanafy, Saifullah Shafiq,
Al-Shuhail, A. A., Umair bin Waheed, "First Break
Traveltime Picking Using Unsupervised Machine Learning

and Super-virtual Refraction Interferometry.” In *IEEE Transaction on Geoscience and Remote Sensing*.”

(Submitted).

Awais, Muhammad, Sherif M. Hanafy, "Enhancing the accuracy of DBSCAN's first arrival traveltime picking using super-virtual refraction interferometry." In *Second International Meeting for Applied Geoscience & Energy*, pp. 1915-1919. Society of Exploration Geophysicists and American Association of Petroleum Geologists, 2022.

Awais, Muhammad, and Sherif M. Hanafy. "Land-streamer data: Fast and accurate seismic technique for shallow engineering applications." In *Sixth International Conference on Engineering Geophysics, Virtual, 25–28 October 2021*, pp. 147-149. Society of Exploration Geophysicists, 2021.