Evaluation and Application of Machine Learning Techniques to Data Conditioning Problems in Microseismic Data

by

Michael J. Nava

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor James W. Rector III, Chair
Professor Carl Boettiger
Professor Joan Walker

Fall 2020

Evaluation and Application of Machine Learning Techniques to Data Conditioning Problems in Microseismic Data

Abstract

Evaluation and Application of Machine Learning Techniques to Data Conditioning Problems in Microseismic Data

by

Michael J. Nava

Doctor of Philosophy in Engineering - Civil and Environmental Engineering

University of California, Berkeley

Professor James W. Rector III, Chair

Hydraulic fracturing has evolved dramatically over the past decades. A number of new techniques have emerged in order to maximize production from organic-rich shale. For example, multistage fracturing, dynamically varying pumping parameters, horizontal drilling and finely-tuned perforation shots have all led to incremental improvements in the industry. With these engineering advancements, so too has the ability to monitor microseismic fractures expanded. An added benefit, or potentially an unintended consequence, of this new era of high frequency, high precision acoustic monitoring equipment is the generation of large scale digital data. With any real data set, there will inevitably be data conditioning problems that exist. Whether missing values, corrupt data, or poor experimental design and execution, there will be some constraint or obstacle that inhibits the cultivation of knowledge and insights.

The objective of this dissertation is to identify and understand where those limitations exist, to understand the genesis of those constraints - whether they arise from some physical limitation or from common data recording issues - and then apply an interdisciplinary approach to overcome those limitations.

To this end, we identify limitations caused by a typical, cost-effective microseismic monitoring geometry and pivot to understand and characterize microseismic events through spectral analysis. We build features that provide insight into the nature of microseismicity present in the data, which would otherwise elude us. Next, we incorporate information that is typically lost in the presence of high amplitude resonance and leverage this newly found data to identify specific microseismic attributes to make marked improvements on event location estimates. Through the inclusion of head waves and the use of inversion techniques, we reduce the uncertainty of microseismic event locations significantly. This is a fundamental step toward understanding the behavior of hydraulic fractures far beneath the surface of the earth.

Next, we turn to data science to continue to overcome data quality issues present in the data from a hydraulic fracturing project in the Marcellus shale. Specifically, machine learning and deep learning methodologies are applied to the data in order to recover *meaningful* information. The benefits of this are twofold. First, this work provides a data-driven approach to imputation through various learning methods. Second, it provides an understanding of the limitations and computational time required for various learning methods. This information will aid in the decision making of engineers who desire a more accurate solution or an accurate solution that can be used in real-time analysis.

Finally, we culminate the dissertation with an exploration into the ability to leverage ensemble learning methods to overcome poorly conditioned data sets with the objective of improving automated analysis steps. Specifically, we create an extensible computational paradigm that enables the automatic picking of waveform first arrivals. This is typically an arduous, time-consuming analysis step that suffers from inconsistent picks based on subjective assessment. Moving away from a human-in-the-loop system enables more transparency and reproducibility. Additionally, the total time for end-to-end analysis of first arrivals is dramatically decreased. Given the extensibility of this framework, expanding the use of the system to include full waveform classification is an appropriate next step.

To Emily, Alessandra, and Sofia

Contents

List of Figures

List of Tables

Acknowledgments

I would first like to thank my advisor, Professor James W. Rector, for his unyielding support. His guidance has been invaluable in my pursuit of knowledge. His approach to a difficult problem has inspired me to find calm in the moments between uncertainty and understanding.

I am grateful for Shelley Okimoto for helping me navigate the tumultuous waters of PhD life. Without her help, the work that follows would be lost. Berkeley and the Civil and Environmental Engineering department are incredibly lucky to have someone so dedicated to her craft.

A special thanks to my co-author, Dr. Zhishuai Zhang, for his friendship and help making nebulous ideas become concrete concepts throughout this process. It is always better to embark on a difficult journey with someone next to you.

Thank you to Dr. Eric Munsing for helping me conquer a difficult part of this endeavor. It takes a truly kind person to help a complete stranger. Your guidance and words of reassurance helped me focus on the task at hand, rather than the unknown.

I am very appreciative to my committee members, Professors Joan Walker, Carl Boettiger, Marta González, and Scott Moura for taking the time to provide valuable feedback on my work and for reminding me to always appreciate the essence of the world around me.

I would also like to thank the Geophysics Department at Stanford University and Lawrence Berkeley National Lab for the opportunity to present my work and for sharing their wealth of knowledge.

My academic career began at the United States Naval Academy, and it was here that I learned the value of being able to think critically despite volatile conditions. Professor Joshua Radice taught me the value of pushing forward in the face of unbeatable odds. Alex Angelillo, Vidal Rodriguez, Thomas Alessi, Colin Doherty, and Joseph Travers hold a special place in my thoughts. With your support and friendship, I was able to continue when the world seemed to be insisting that I reconsider. Though we were always a world away from each other, I never felt alone. Ethan Harvey, you continue to set the example of true commitment to accomplishing a goal no matter the obstacles. You reminded me that sometimes it is necessary to swim with everything I had without saving anything for the swim back to shore.

I am grateful to Evan Hafer and Mat Best at Black Rifle Coffee Company for providing the fuel needed to push through late nights and early mornings and for helping to create a community for veterans to come together to stay in the fight.

To my parents: through your selflessness and sacrifice, I am the man I am today. To my brother: your relentless encouragement continues to keep me grounded and motivated to face the next challenge head on. To my wife and daughters: your unconditional love and support is the reason I am able to wake up in the morning. All that I am is because of you and for you.

Chapter 1

Introduction

It is not the critic who counts, not the man who points out how the strong man stumbled, or where the doer of deeds could have done them better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood; who strives valiantly; who errs, who comes short again and again, because there is no effort without error and shortcoming; but who does actually strive to do the deeds; who knows great enthusiasms, the great devotions; who spends himself in a worthy cause; who at the best knows in the end the triumph of high achievement, and who at the worst, if he fails, at least fails while daring greatly, so that his place shall never be with those cold and timid souls who neither know victory nor defeat.

– Theodore Roosevelt

1.1 Motivation

Hydraulic fracturing has been used to increase permeability of unconventional reservoirs for some time and, in recent years, has been instrumental in the shale gas revolution (King et al., 2012). Recent technological advances have enabled the successful execution of hydraulic fracturing projects in a way that significantly improves the ability to capture organic-rich shale. Specifically, the drilling of horizontal boreholes used for creating fractures in horizontally distributed shale layers has been a critical driver for the economic feasibility of many unconventional reservoirs (Maxwell, 2014). Additionally, recent advances in microseismic analysis through distributed surface monitoring arrays, as well as downhole and crosswell monitoring geometries, have enabled the capturing of high resolution data.

While there certainly has been a dramatic improvement in the understanding of fracture orientation, source mechanism, fracture network, and event location estimation, there still exists a very real constraint in the field of hydraulic fracturing, which is primarily driven by business needs (Eisner et al., 2007; Maxwell, 2014; Zhang et al., 2017a,b). The influence

of global markets and predicted project viability from an economic standpoint have a very real impact on the ability to capture real, high quality data. Although acoustic recording and storage solutions have been rapidly improving in the digital age, there still exists an intrinsic trade off between optimal monitoring geometry and optimal treatment geometry. Specifically, downhole monitoring wells are typically drilled in a direction that parallels the treatment well (the well that is pressurized in order to create hydraulic fractures) in order to enable the reuse of the well in subsequent phases of the hydraulic fracturing project. This is done in an effort to minimize financial losses, which accompany horizontal drilling efforts. While this reduces overall cost and helps to improve economic viability of a given hydraulic fracturing project, this orientation of the monitoring well generates a hard physical constraint that negatively impacts the ability to adequately understand the physical changes occurring in and around the treatment zone. Details of limitations such as limited aperture are discussed in more detail in Chapter 3, and non-optimized monitoring geometries are discussed in Chapter 4.

In an effort to overcome these limitations, we turn to the field of data science, which leverages machine learning and artificial intelligence methodologies to recover meaningful information and garner insights that would otherwise elude us.

1.2 Dissertation Organization

This dissertation seeks to highlight the interdisciplinary approach undertaken to understand the essence of hydraulic fracturing and microseismic analysis and to overcome hard limitations that arise from the collection of real data in an engineering arena where real world costs and strategic planning constraints must be considered.

- Chapter 2 provides background on the goals and challenges of hydraulic fracturing, an exploration of microseismic monitoring and analysis, and an overview of common monitoring geometries to include downhole and surface monitoring approaches. The following chapters build on this foundation in order to identify and understand limitations that exist based on common monitoring geometries, economic constraints in hydraulic fracturing project design, and data corruption or loss through either capturing or transmission deficiencies.

- Chapter 3 investigates the use of analysis in the spectral domain to overcome the limitations imposed by limited aperture, a common disadvantage to typical monitoring geometry in hydraulic fracturing processes[1]. A typical microseismic monitoring configuration contains two horizontally drilled boreholes – one treatment well and one observation well. This configuration, while cost-effective, leads to an inability to execute moment tensor inversion through traditional means. However, through careful analysis in the spectral domain, parameters like center frequency and bandwidth can

[1] A version of this work was published as (Nava et al., 2015)

be used in tandem with knowledge of process parameters to better understand microseismic source characteristics (Nava et al., 2015).

- Chapter 4 focuses on microseismic data acquired from a geophone array deployed in the horizontal section of a well drilled in the Marcellus Shale near Susquehanna County, Pennsylvania[2]. Head waves were used to improve event location accuracy as a substitution for the traditional P-wave polarization method. We identified that resonances due to poor geophone-to-borehole coupling hinder arrival-time picking and contaminate the microseismic data spectrum. The traditional method had substantially greater uncertainty in our data due to the large uncertainty in P-wave polarization direction estimation. We also identified the existence of prominent head waves in some of the data. These head waves are refractions from the interface between the Marcellus Shale and the underlying Onondaga Formation. The source location accuracy of the microseismic events can be significantly improved by using the P-, S-wave direct arrival times and the head wave arrival times. Based on the improvement, we have developed a new acquisition geometry and strategy that uses head waves to improve event location accuracy and reduce acquisition cost in situations such as the one encountered in our study (Zhang et al., 2017a).

- Chapter 5 explores the idea of imputing corrupt or missing data through the use of machine learning methods. Corrupt or missing data, whether due to unavoidable physical constraints or from data recording issues that lead to information loss, are prevalent in nearly every seismic data set. There are a number of imputation techniques that attempt to overcome this problem in general; however, there has been limited work evaluating the applicability of machine learning methodologies for imputation on microseismic data sets. This chapter considers data from a hydraulic fracturing microseismic monitoring experiment that took place in the Marcellus Shale near Susquehanna County, Pennsylvania. One significant cause of data corruption is the presence of large amplitude resonance energy on non-axial sensor components that inhibit the identification of first arrival times and compressional amplitudes for both direct and head waves. We evaluated the performance of various learning techniques used to impute missing or corrupt data. After performing k-fold cross-validation, notable improvement is seen and a significant portion of missing values are recovered with minimal error. As a result, a data set, complete with imputed variables, can be used to leverage a number of machine learning and deep learning techniques to gain more insight to aid in subsequent analysis steps (Nava et al., 2020b).

- Chapter 6 attempts to improve first arrival picking, which is a critical step in understanding hydraulic fracturing through microseismic monitoring. Typically, this is performed by a subject matter expert and can be incredibly time intensive. Alternatively, automatic first arrival picking techniques can be applied; however, this commonly injects greater error into the analysis process. Time series classification is an area of

[2]This chapter presents a modified version of a previously published work.

artificial intelligence and machine learning that has not been applied to microseismic data until recently. Through the use of ensemble learning methods, we propose a new method for classifying compressional waves and then applying standard picking methods to make improvements on overall accuracy. We apply this method on 249 traces from a hydraulic fracturing project and create a unique group cross-validation method that is well-suited for time series data. Extreme Gradient Boosting (XGBoost) with dropout results in a classification accuracy of 94.9% and enables the reduction in mean absolute error from 126 ms to 23.8 ms on real microseismic data from the Marcellus Shale. Dynamic parameterization and an extensible framework enable the potential for multiclass classification to identify shear wave arrivals with minimal effort (Nava et al., 2020a).

While these chapters are designed to be self-contained, the overarching goal of the dissertation is to understand the nature of the hard constraints that exist in the field of hydraulic fracturing and microseismic analysis that typically arise from noisy or corrupt data. Then we attempt to apply new methods with roots in signal processing, artificial intelligence, and data science in order to overcome those limitations. In order to accomplish this, an interdisciplinary approach is necessary. As such, Figure 1.1 shows the areas of research that intersect to form the body of this work.

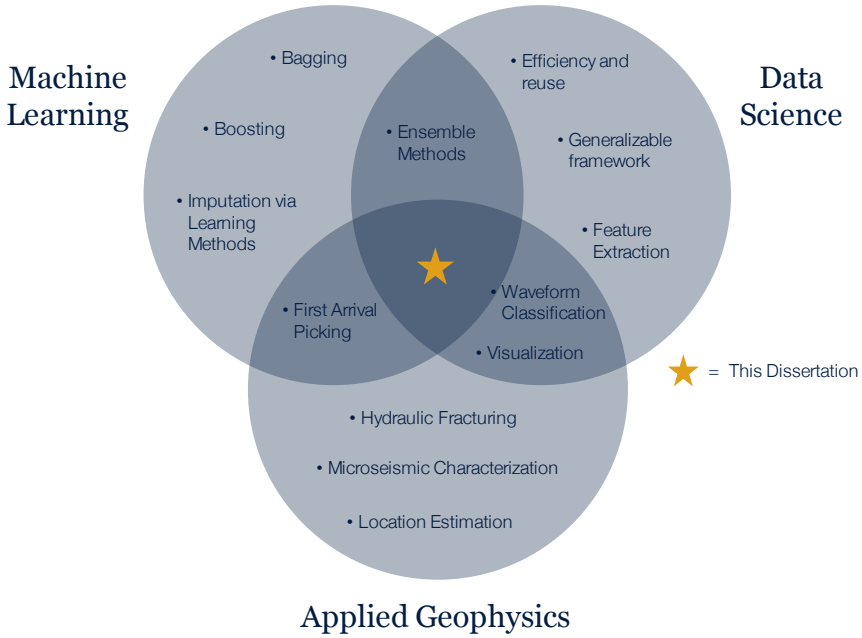

Figure 1.1: Overview of the areas of research this dissertation explores. This work exists at the intersection of applied geophysics, machine learning, and data science.

Chapter 2

Background

The time will come when diligent research over long periods will bring to light things which now lie hidden. A single lifetime, even though entirely devoted to the sky, would not be enough for the investigation of so vast a subject... And so this knowledge will be unfolded only through long successive ages. There will come a time when our descendants will be amazed that we did not know things that are so plain to them... Many discoveries are reserved for ages still to come, when memory of us will have been effaced.

– Seneca

2.1 Hydraulic Fracturing

Hydraulic fracturing is the process of injecting fluid at pressure that exceeds the minimal principal stress of a formation to create cracks and fractures with the purpose of capturing natural gas and other organic-rich material (King et al., 2012). In the past, there were a number of specific conditions that would need to be met before a potential conventional reservoir would be considered potentially viable. For example, hydrocarbon source rocks would need to be located, reservoir quality rocks would be identified, and surveys would be performed in order to locate a trapping mechanism. Next, wells would be drilled in order to capture valuable material (Alexander et al., 2011). With the introduction of hydraulic fracturing, hydrocarbon source rocks that were previously ignored are now much more viable.

The general workflow for executing a hydraulic fracturing project is:

1. Identify organic-rich shale
2. Perform initial survey to understand geologic attributes in the area of interest
3. Drill vertically and encase vertical borehole in concrete to protect water table and other near surface features

4. Reach kick-off point (some vertical distance below the surface) and turn the drill bit to drill horizontally

5. Enclose small length of the borehole (this is referred to as a "stage") and execute perforation shots to create initial fractures in the preferred direction. This is typically done from the toe of the well and continues in the direction of the kick-off point.

6. Complete all stages of the hydraulic fracturing project and withdraw tooling from the well. Retrieve hydrocarbon material.

Figure 2.1 shows an example of a typical hydraulic fracturing project that employs a horizontal treatment well. Each grouping of yellow lines represents a single stage in the overall project. The point at which the borehole is turned vertically is referred to as the *kick-off point*, and the end of the well (at the left of the image) is referred to as the *toe* of the well.

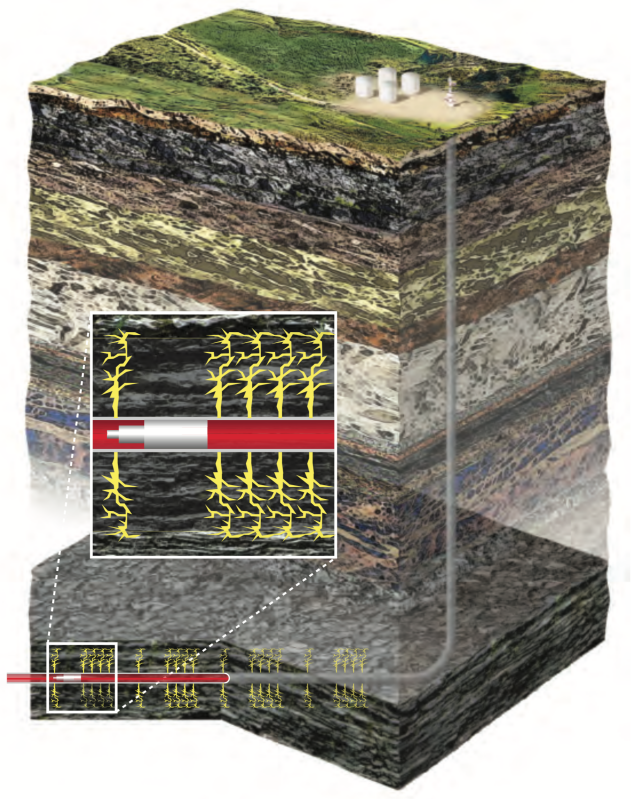

Figure 2.1: Depiction of typical hydraulic fracturing project that incorporates the use of horizontally drilled borehole and multistage treatment design. Note the relatively small surface footprint (Alexander et al., 2011).

While there is some variability in the true depths of each hydraulic fracturing project, it can be seen in Figure 2.2 that there is a great distance between the horizontal well and near surface features like water tables. Most water tables exist at depths less than 1000 ft from the surface (King et al., 2012). It should be noted that the average water table depth in the area where the data were collected for this work was approximately 67 ft.

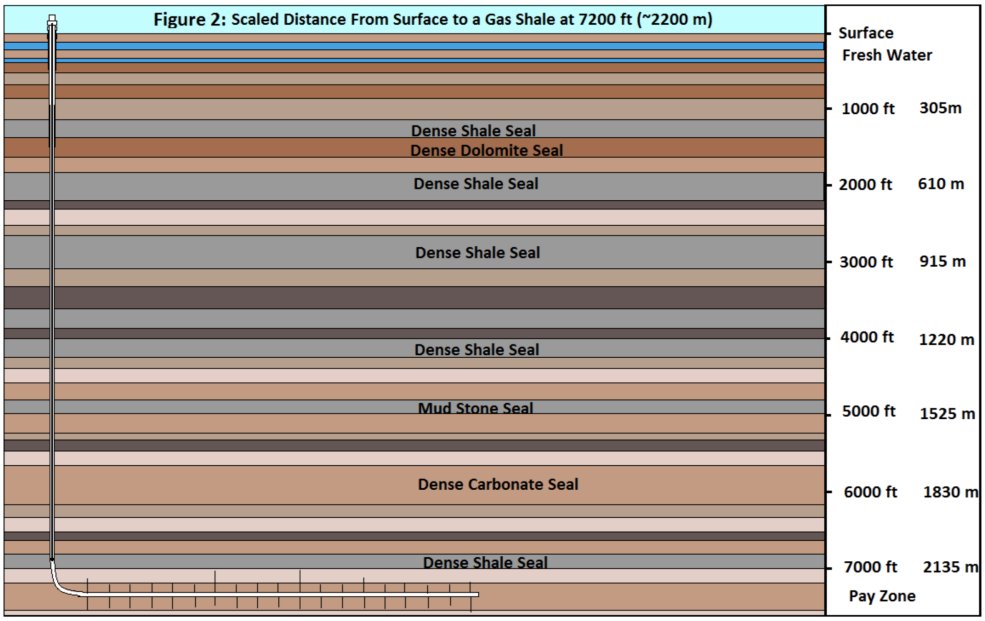

Figure 2.2: Cross-sectional view of a typical hydraulic fracturing project, drawn to scale. Shows vertical distance between pay zone and water table (King et al., 2012).

2.2 Microseismic Monitoring

There are a number of human activities that can induce small earthquakes, or microseismic events, in the subsurface; however, this work focuses on microseismic events caused from hydraulic fracturing. Microseismic analysis, in this arena, is the identification and characterization of these small scale earthquakes caused from the fracturing of rock in the pay zone and surrounding areas. A number of methods exist to understand microseismics; however, a crucial first step involves understanding microseismic event location. In this regard, there exists a significant body of work that includes least-square traveltime inversion (Richards and Aki, 1980; Rutledge and Phillips, 2003), time-reverse imaging (Artman et al., 2010; Artman and Witten, 2011), simultaneous inversion with Bayesian inference (Zhang et al., 2017b), and full-waveform inversion (Song and Toksöz, 2011).

Although event location estimation is a common processing step, which is fundamental in understanding the nature of fracturing in the subsurface, strict limitations exist that

diminish the ability to successfully analyze the data recorded from the hydraulic fracturing project, a topic that is covered in the remainder of this dissertation.

Two forms of microseismic monitoring are in common practice today, surface monitoring and downhole monitoring. The data utilized in this work come from downhole monitoring.

Surface Monitoring Geometry

Surface monitoring arrays have been used to monitor hydraulic fracturing projects for some time. There are distinct advantages to surface monitoring over downhole monitoring. For example, surface monitoring arrays typically have a much larger aperture through which inversion of microseismic source mechanism is possible. This comes as a direct result from the large azimuthal coverage that accompanies these geometries. However, these advantages come at a cost. Specifically, surface monitoring arrays typically contain thousands to tens of thousands of geophones, or acoustic recording sensors.

Additionally, the use of surface monitoring dramatically increases the surface footprint of a hydraulic fracturing project. As a result, there are commonly issues regarding land permits that limit the orientation and contiguous nature of the sensor arrays (Harris and Bacon, 2015). Figure 2.3 demonstrates both the large scale nature of surface monitoring arrays as well as the difficulty that arises from non-permitted regions surrounding the treatment well.

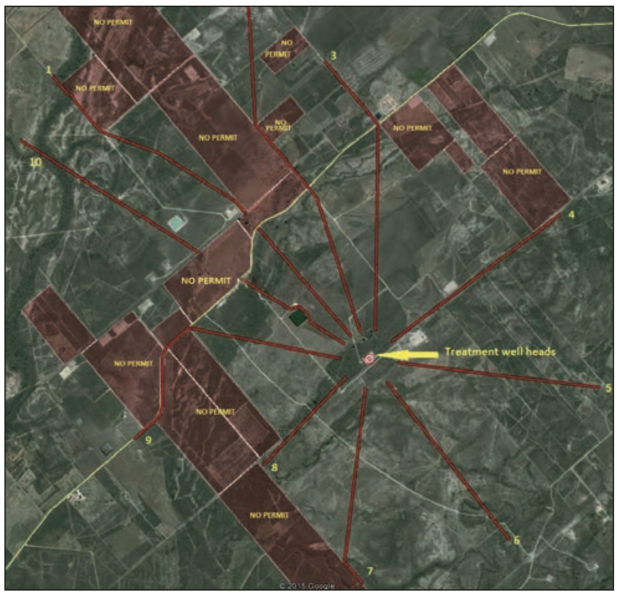

Figure 2.3: Satellite imagery showing a common surface monitoring geometry. While effective in monitoring microseismic activity, the overall area is large and number of sensors required is much greater than downhole monitoring. Limitations such as permitting exist in surface monitoring and areas in red denote where this occurs (Harris and Bacon, 2015).

Downhole Monitoring Geometry

Downhole monitoring is another method of recording acoustic signals from a hydraulic fracturing project and has been widely used for microseismic analysis. There are distinct advantages to downhole monitoring. For example, a common pitfall encountered in surface monitoring is an inability to capture meaningful signal from lower magnitude microseismic events. This phenomenon stems as a direct result from the poor signal-to-noise ratio (S/N) due to the combination of large source-receiver distances, low magnitude seismic signal, and scattering. Conversely, in downhole monitoring, S/N is significantly better due to the fact that the source-receiver distance is drastically smaller. Many monitoring approaches move the geophone array to remain perpendicular from the current treatment stage in order to further minimize the source-receiver distance to improve S/N. This orientation can be seen in Figure 2.4. Here, it can be seen that for a given fracturing stage, the geophone array, which is located in an observation well that parallels the treatment well, is located directly across from the stage being monitored. Also note that the typical source-receiver distance is 500 - 1500 ft. This approach to optimizing S/N was employed in the hydraulic fracturing project under consideration in this dissertation.

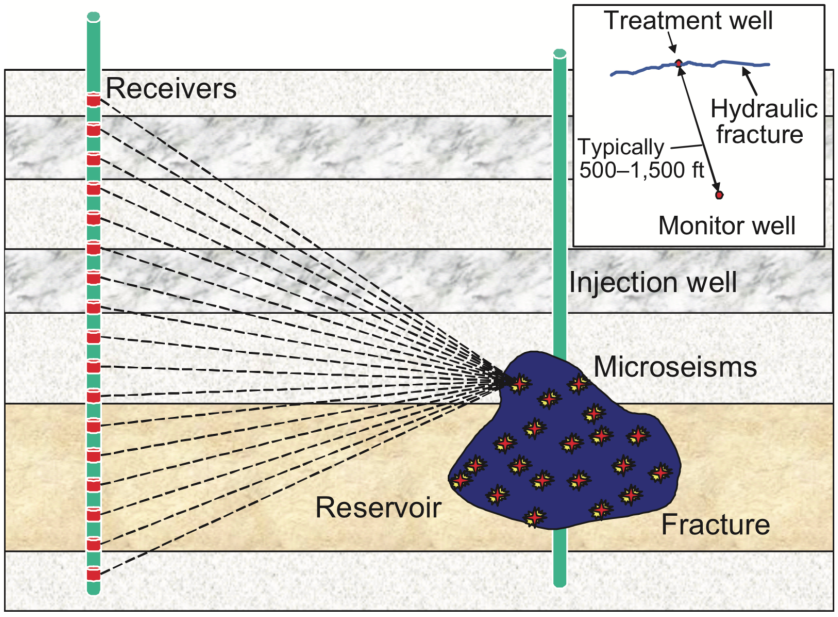

Figure 2.4: Depiction of typical microseismic downhole monitoring geometry. The injection well, also known as treatment well, is located on the right of the image. Parallel to this well is the observation well, which is where acoustic sensors called geophones are located. Note that each microseismic event is recorded on multiple geophones at different times. This is a fundamental requirement for understanding microseismic location and source mechanism (Warpinski et al., 2009).

Microseismic Uncertainty

A disadvantage of downhole monitoring is directly related to the small source-receiver distance. Namely, because the geophones are located perpendicular to the stage being monitored, and at a relatively short distance, there is a loss of azimuthal coverage. This leads to the problem of *limited aperture*, where the ability to perform analysis tasks like moment tensor inversion is no longer possible due to the small solid angle (Vavryčuk, 2007).

Additionally, despite efforts to improve S/N with downhole monitoring geometries, poor S/N remains an ever-present issue in real data collected from hydraulic fracturing projects. Moreover, inaccurate velocity models, insufficient azimuthal coverage of geophones, as well as general data transmission and recording issues all lead to uncertainty in microseismic analysis (Eisner et al., 2009; Maxwell, 2009). The effects of these uncertainties are significant and motivate the following chapters in this dissertation. Figure 2.5 shows how uncertainties in event locations increase with distance from the monitoring well (Warpinski et al., 2009). This phenomenon is common in downhole monitoring geometries and is present in the data that are investigated in this dissertation. Chapter 4 discusses methods we investigated and developed to decrease event location uncertainties in hydraulic fracturing microseismic events.

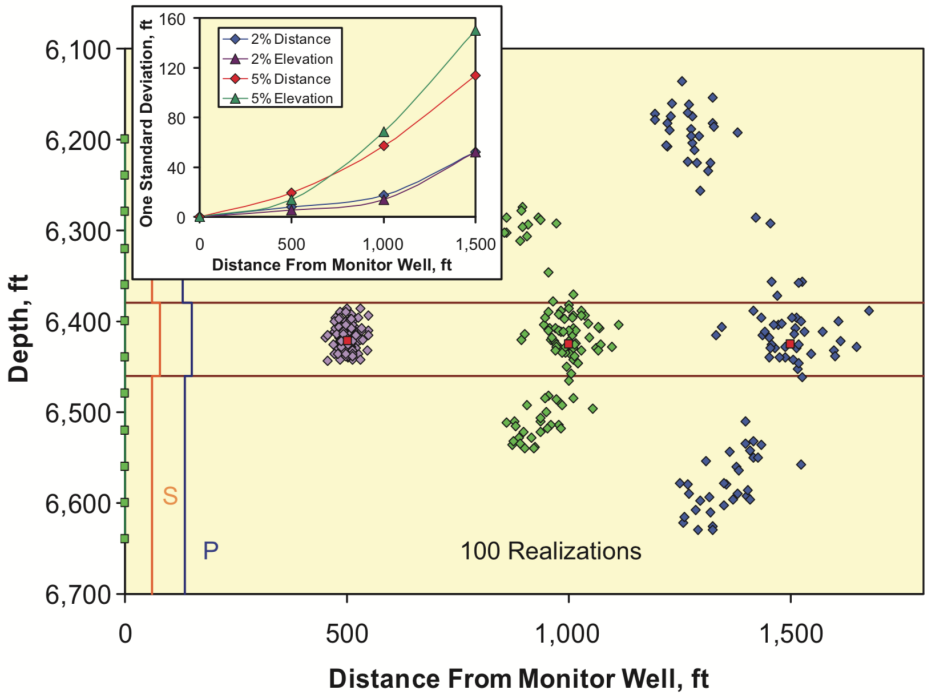

Figure 2.5: Depiction of microseismic location uncertainty that is common in microseismic analysis. Uncertainty can be attributed to a velocity structure that is inaccurate, data quality issues, low magnitude microseismic events, or in general, the presence of non-unique solutions (Warpinski et al., 2009).

Chapter 3

Characterization of Microseismic Events through Spectral Analysis

Those who dare to fail miserably can achieve greatly

– John F. Kennedy

3.1 Introduction

In this chapter[1], we begin with an investigation into data loss due to a physical constraint that stems from a common microseismic monitoring geometry used in the hydraulic fracturing industry. There has been a significant increase in the amount of hydraulic fracturing projects in the United States as a result of a number of technical and economic factors. One of the main technological advances that has enabled hydraulic fracturing projects to be completed, which were previously economically infeasible, is the ability to drill horizontal boreholes. There are many advantages to this approach over the traditional vertical borehole method. For example, a much larger treatment zone in an area of interest can be produced as a direct result of the project geometry. Specifically, due to the orientation of shale formations, a much larger pay zone can be realized by drilling for a greater distance within a horizontal formation.

In order to monitor the microseismic activity resulting from these types of hydraulic fracturing processes, surface arrays or crosswell monitoring arrays are utilized. Surface arrays can be an effective tool for monitoring microseismic activity since they can provide large azimuthal coverage. However, given that the magnitudes of events resulting from hydraulic fracturing typically range from -1 Mw to -4 Mw, and that the depth of fracturing is usually one or more miles below the surface, signal-to-noise (S/N) can become a difficult problem to overcome. As such, there is usually the need for both a very large number of acoustic sensors

[1] A version of this work was published as (Nava et al., 2015)

(6,000-24,000 geophones) and a large area at the surface (1-3 miles) to achieve coherent monitoring of microseismic events (Duncan and Eisner, 2010).

Downhole monitoring with a horizontal observation well requires significantly fewer acoustic sensors to achieve good S/N; however, there are also a number of disadvantages to this approach. For example, there is increased uncertainty when determining microseismic event location. This comes as a direct result of the survey geometry. Specifically, since the monitoring array is parallel to the treatment well, location estimates rely on hodogram angle of inclination for depth determination (Maxwell, 2014). The main disadvantage of crosswell monitoring, however, is an inability to perform moment tensor inversion with a single monitoring well (Vavryčuk, 2007). This constraint is due to the small solid angle as a result from the close proximity of geophones and accompanying limited azimuthal coverage of the treatment zone. This is referred to as the limited aperture problem. In an effort to overcome this restriction, we turn to the spectral domain.

3.2 Hydraulic Fracturing Project Overview

The hydraulic fracturing project was performed in Susquehanna County, Pennsylvania in the Marcellus shale formation using the horizontal drilling technique previously discussed. Two horizontal boreholes were drilled – one treatment well and one observation well. The treatment well was used to inject fracture fluid at high pressure in order to exceed the treatment zone's minimum principal stress in an attempt to create new fractures. The newly created fractures increase the permeability and porosity of the zone of interest for the retrieval of hydrocarbon-rich material. A second horizontal borehole was drilled parallel and approximately at the same depth in order to house an array of geophones for measuring acoustic emissions. The treatment well was approximately 5,600 ft in length; the observation well was approximately 4,400 ft in length and the distance between the wells was approximately 720 ft.

There were eighteen fracturing stages in the project progressing from the toe of the well to the heel of the well (Figure 3.1). In order to monitor the acoustic emissions from microseismic events, the geophone array was moved six times in an effort to minimize viewing distance (Table 3.1).

Reducing the viewing distance by moving the geophone array is important as it improves S/N; however, reducing the distance between source and receiver also limits the azimuthal coverage of the events. As such, there is a reduced ability to perform moment tensor inversion as a result of the basic monitoring geometry.

3.3 Methods

Analysis of raw data showed many instances of large amplitude ringing. Given that the geophones were not locked into place, or clamped to the borehole casing, these high frequency

Table 3.1: Description of geophone locations and associated stages.

Hydraulic Fracturing Stage	Geophone Array Location
1-9	1
10-11	2
12-13	3
14-15	4
16-17	5
18	6

artifacts were likely caused by tube waves propagating through the borehole (Gaiser et al., 1988). In order to minimize the negative effects of this artifact, low pass and band pass Butterworth filters were applied and a location-based noise characterization and removal schema was developed. This approach considered the root mean square (rms) of each channel of the geophone array for all events. Then the average rms was found for each location of the geophones by only considering the events that occurred at each monitoring location. This is an important step because with each move of the geophone array, the noise signature changes due to a number of factors. For instance, at the first location the geophones could all be oriented in the same manner; however, after being pulled to the next monitoring location, any of the geophones could have shifted in transit. A location-specific approach to noise minimization accounts for these inconsistencies (Figure 3.2).

After processing the raw data in order to minimize noise from poor coupling of geophones, first arrivals were picked and a 100 ms Tukey (tapered cosine) window was applied to the data in order to capture various waveforms. Specifically, a window to capture compressional waves, a window to capture shear waves, and also a combined window capturing both waveforms were applied to the processed data. This approach gives information regarding the spectral content of each waveform and also the overall event for later analysis.

A Fourier Transform was performed on the three windowed wave types for each trace, which yielded a spectral response for each of the eleven geophones. In order to reduce the amount of data to interpret, an average spectral response of all traces for each event was calculated to show a representative spectral response on a total event basis.

With a representative spectral response for each event, we begin to focus on properties like bandwidth and center frequency to gain some intuition regarding relationships between event spectra and source mechanism. In order to classify bandwidth, the global maximum of each signal was identified and the associated prominence calculated. At one-half the prominence, the width of the signal is noted. This is done for all event spectra and we are left with a relative measure of bandwidth for the three wave types. An example of a broadband event is shown in Figure 3.3.

Additionally, the center frequency for each event was determined using the centroid

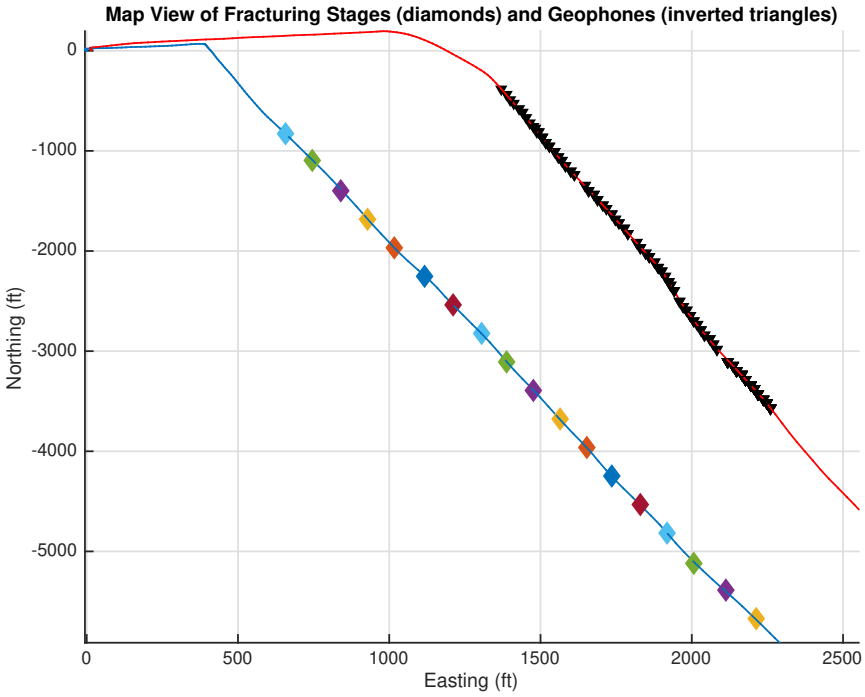

Figure 3.1: Map view of hydraulic fracturing project in the Marcellus shale. The blue line indicates the treatment well and the red line indicates the observation well. Diamonds represent the average locations of perforations for each of the eighteen stages. Inverted triangles represent the six locations of the geophone array.

method (Bracewell and Bracewell, 1986). In order to find the frequency at which the majority of the signal energy is located, the first moment, or centroid, of the event spectra is found by:

$$f_c \coloneqq \langle x \rangle = \frac{\int_{-\infty}^{\infty} x f(x) dx}{\int_{-\infty}^{\infty} f(x) dx} \tag{3.1}$$

An understanding of where the signal energy is located in the frequency domain is important as it can give information about slip distance, Q determination, and other source parameters (Beresnev, 2001; Brune, 1970; Eaton, 2011, 2014; Maxwell and Cipolla, 2011). With bandwidth and center frequency measurements for windowed compressional waves, windowed shear waves, and also a combined window, event characteristics can be seen.

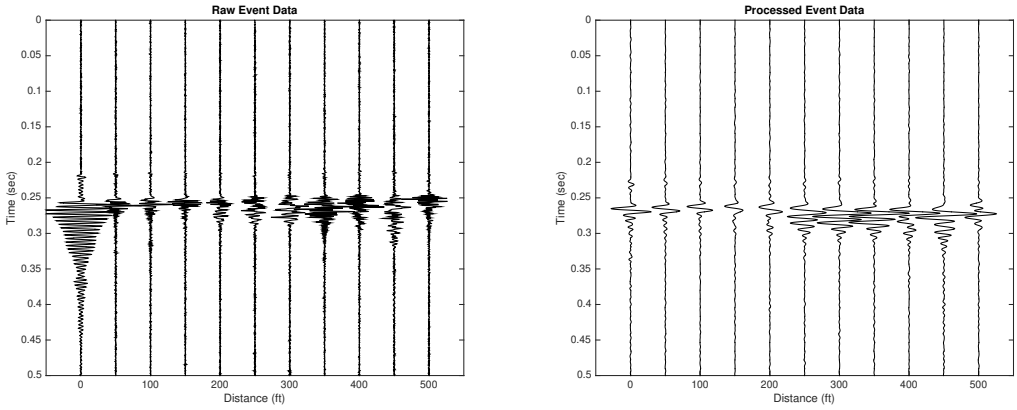

Figure 3.2: Seismogram showing raw data containing large amplitude artifact believed to be tube wave energy (left) and processed data capturing the same microseismic event (right).

3.4 Results

As a preliminary step, the mean center frequency and mean bandwidth were calculated in order to see if there were apparent statistical trends in the data. Compressional wave mean center frequency and bandwidth are lower then the shear wave parameters. Furthermore, the combined window mean center frequency and bandwidth are both closer to the mean of the shear wave (Table 3.2). As such, it can be inferred that the majority of microseismic energy may be associated with shear openings and less from tensile events.

Table 3.2: Mean values of bandwidth and center frequency for the three types of applied windows.

	Mean Bandwidth (Hz)	Mean Center Frequency (Hz)
Compressional Wave	61.9	86.0
Shear Wave	72.5	92.3
Combined Window	66.0	886

Considering the event spectra, we see that there is a large amount of variation between events throughout the hydraulic fracturing project (Figure 3.4a). After sorting these event spectra by bandwidth, it can be seen that there is still variation within narrowband events (Figure 3.4b). Finally, after sorting event spectra by center frequency, less variation can be seen (Figure 3.4c). Moreover, in this last view, it is evident that the majority of broadband events are located near the middle of the range. This is due to the fact that the centroid method considers the frequency at which the majority of signal energy is located. As such, broadband events typically have center frequencies near the mean.

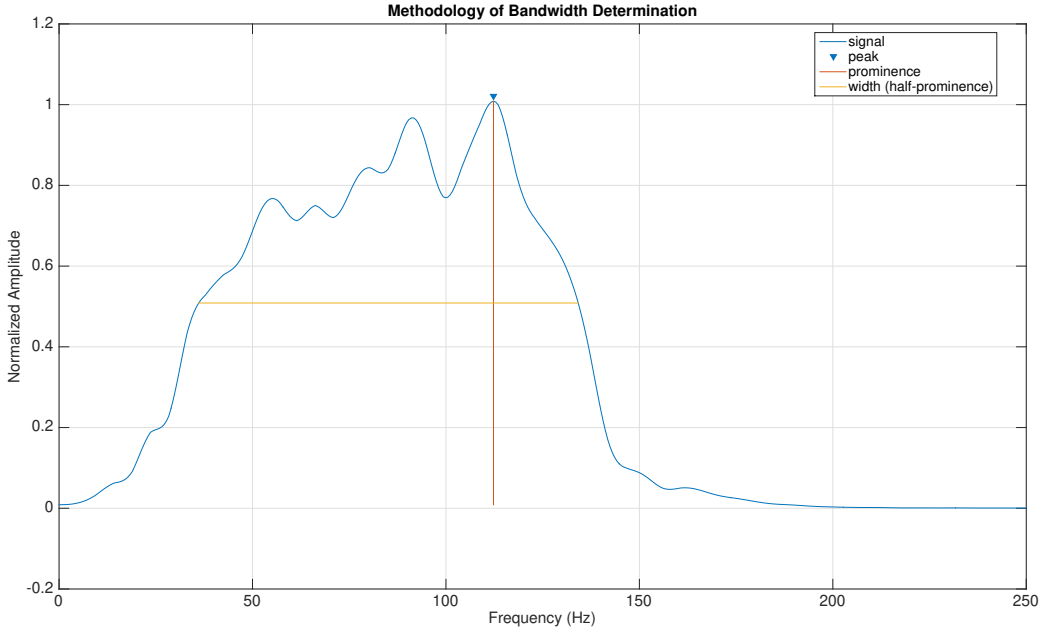

Figure 3.3: Determination of bandwidth. Inverted triangle represents global maximum, vertical line shows prominence of the signal, and horizontal line represents the width measured at one-half the prominence.

In an effort to more effectively interpret the data, scalar values of center frequency and bandwidth were plotted as a function of time. This enabled correlation between spectral properties and process parameters like surface pressure, slurry flow rate, and proppant concentration. Figure 3.5 shows spectral and process parameters as a function of time.

Interesting relationships can be seen between both bandwidth and center frequency when compared to event magnitude in the seventh stage of the hydraulic fracturing project. For example, there is an indication that bandwidth has an inverse relationship with event magnitude. As such, there is an implication that narrowband events are accompanied by greater magnitude. Additionally, center frequency appears to vary proportionally to magnitude. Consequently, it seems that events with the largest magnitude are narrowband events with high center frequencies.

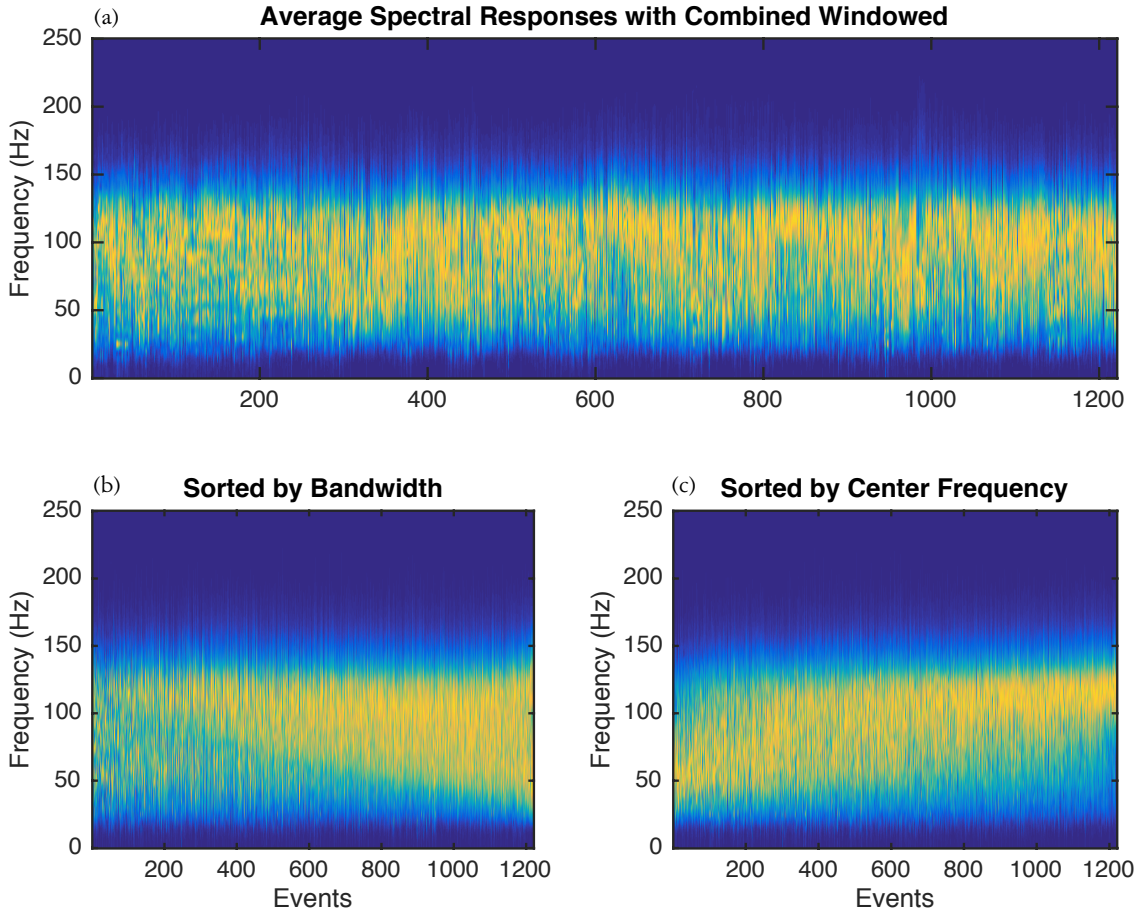

Figure 3.4: Combined window event spectra for all events in the hydraulic fracturing project. Color represents normalized amplitude where blue is lowest and yellow is greatest. Unsorted events (top) show large variation between neighboring events. Bandwidth-sorted events (bottom left) show variation between narrowband events. Center frequency-sorted events (bottom right) show broadband events located in the middle – near the mean.

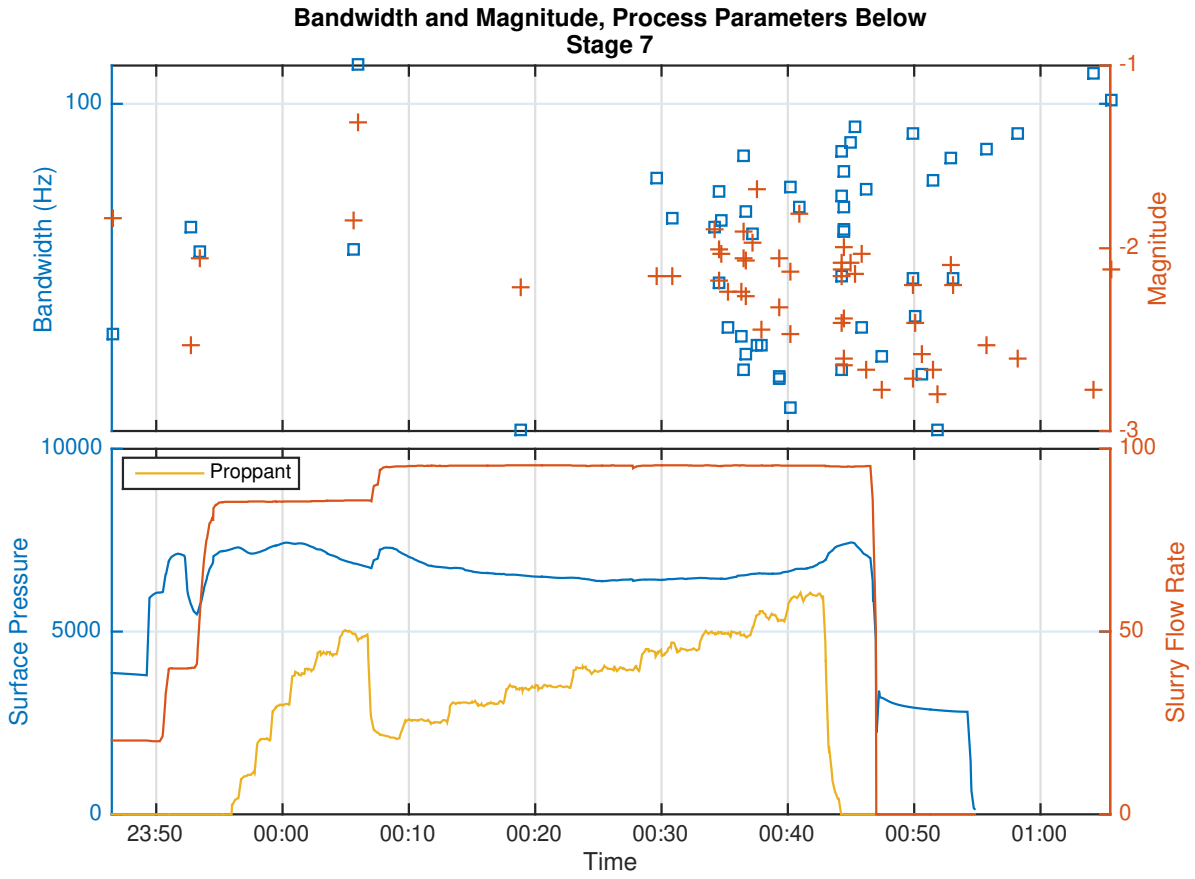

Figure 3.5: Bandwidth (blue diamond) and event magnitude (red plus) as a function of time shown on top. Note that at the end of the stage, it is clear that there is an inverse relationship between bandwidth and magnitude. Process parameters are same as above.

Another method of analysis is similar to the S/P amplitude ratio method traditionally used to understand source mechanism. Here, the event bandwidth is considered. In an effort to determine the main component of source energy, we investigate the ratio of windowed shear wave bandwidth to windowed compressional wave bandwidth. Since the mean bandwidth was higher for shear waves and lower for compressional waves, we conclude that a larger bandwidth ratio indicates a shear wave dominated event. Conversely, a lower bandwidth ratio would indicate that the event is dominated by compressional energy. A spatial plot displaying the locations of microseismic events is shown where color and size both indicate S/P bandwidth ratio (Figure 3.6).

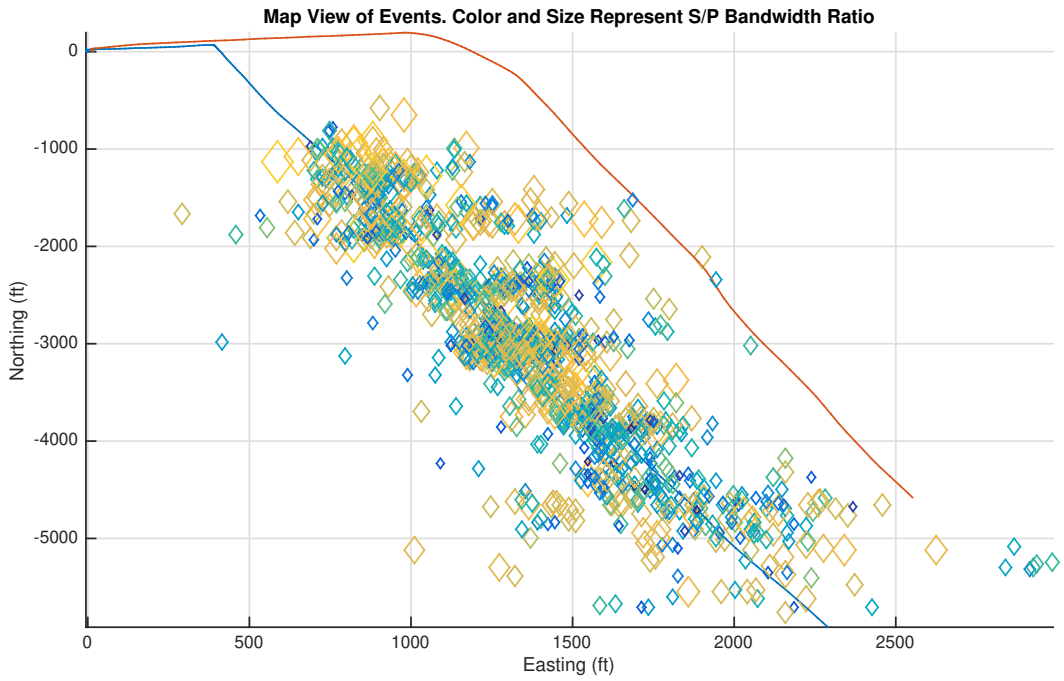

Figure 3.6: Map view of treatment zone. Diamonds indicate locations of microseismic events. Color and shape both represent S/P bandwidth ratio, where blue is smallest and yellow is largest. Large, yellow diamonds represent shear-dominated events.

While it may be difficult at this stage to determine whether high shear regions exist, it is possible to better understand fault plane orientation in the area of interest (Warpinski et al., 2010). Despite the limitations imposed on source mechanism determination as a result of survey geometry, shear and compressional wave-dominated events are seen distributed throughout the treatment zone.

3.5 Conclusion

In an effort to overcome the restrictions imposed on moment tensor inversion as a result of a single monitoring array configuration, analysis in the spectral domain is performed. Preliminary analysis reveals relationships between spectral parameters and both source and process characteristics. For example, after windowing compressional and shear waves, compressional waves are, in general, more narrowband in nature. Shear waves are predominately broadband events with a higher center frequency. Considering the ratio of these determined parameters gives an indication of events dominated by shear energy, which may lead to a better understanding of shear opening events in the treatment zone.

Chapter 4

Location Estimation in the Marcellus Shale

I can live with doubt and uncertainty and not knowing. I think it is much more interesting to live not knowing than to have answers that might be wrong. If we will only allow that, as we progress, we remain unsure, we will leave opportunities for alternatives. We will not become enthusiastic for the fact, the knowledge, the absolute truth of the day, but remain always uncertain ... In order to make progress, one must leave the door to the unknown ajar.

– Douglas Adams

4.1 Introduction

The previous chapter presented an approach that attempted to create features from the data considered that were otherwise unavailable in order to recover meaningful information for analysis. This chapter [1] follows a similar methodology in the sense that we overcome poorly conditioned data in order to recover geophysical attributes that enable additional analysis and improved results. Microseismic monitoring has been widely used for hydraulic fracturing monitoring and characterization since its initial implementation (Cipolla et al., 2011; Eisner et al., 2007; Maxwell, 2014; Warpinski et al., 2009). Microseismic acquisition can use either surface or downhole deployments (Duncan and Eisner, 2010; Maxwell et al., 2010). Shallow wells (typically below the water table) are also used for situations where downhole monitoring is inadequate (Cladouhos et al., 2013). For downhole microseismic monitoring, it is common to have only one nearby well available for microseismic monitoring (Warpinski et al., 2009). To assist in overcoming the aperture limitations imposed by the acquisition geometry, three-component geophones are deployed, which makes polarization

[1] A version of this work was published as (Zhang et al., 2017a)

analysis feasible (Yuan and Li, 2016, 2017). Moreover, multiple phase identification, and full-waveform inversion of microseismic signal are also possible in some environments (Belayouni et al., 2015; Song and Toksöz, 2011; Zhang et al., 2015). In a borehole seismic survey, a geophone can record the ground motion accurately only if it is well-coupled to the well borehole. Unfortunately, this is usually not the case due to a lack of locking force (Gaiser et al., 1988; Sleefe et al., 1995). The poor coupling may lead to severe resonance in seismic waveforms and is common in microseismic surveys (Sleefe et al., 1995). Gaiser et al. (1988) conducted an experiment to study the resonance of geophones in a vertical well used for vertical seismic profiling (VSP). In their experiment, a geophone was locked in borehole with a horizontal locking force to imitate a typical VSP condition. They found that the geophone was subject to severe resonance issues in the horizontal (radial with respect to the borehole axis) component that is perpendicular to the locking arm and the locking force direction when there are only two points of contact with the borehole well. In the cases where cylindrical geophones are deployed in horizontal wells, as is common in microseismic monitoring, there is only one point of contact with the borehole wall. The only coupling force between the geophone and borehole in this situation is usually the gravitational force of the geophone. As such, the resulting waveform shows even more severe resonance due to the lack of locking force. Bandpass filters have been designed and applied in previous research to mitigate the effect of downhole geophone resonance (Nava et al., 2015); however, this is based on the assumption that the resonance frequency is known and different from the microseismic spectrum.

Microseismic surveys with a single monitoring well and location estimation with only P- and S-wave arrival times result in event locations with ambiguity due to the limited coverage of acquisition geometry (Warpinski et al., 2005). An additional constraint on event location usually comes from direct P-wave polarization (Dreger et al., 1998; Eisner et al., 2009; Li et al., 2014). Three-component data are necessary for P-wave polarization direction estimation. The major challenges in using three-component data are the unknown orientation of downhole geophones, poor coupling between geophone and borehole wall, and anisotropic/multiple arrival effects in the P-wave polarization estimation (Coffin et al., 2012; Du et al., 2013; Gaiser et al., 1988; Maxwell, 2014). These challenges make the uncertainty in the P-wave polarization estimation relatively large and is usually a major source of microseismic event location uncertainty (Eisner et al., 2009; Maxwell, 2009). A perforation cluster, each of which usually consists of four to five shots and spread around 0.3 m (1 ft) length, can be treated as point source and used for geophone orientation calibration. In this chapter, we refer to perforation cluster as *perforation shot*, which is considered infinitely small in dimensions when compared with the microseismic event location uncertainty. However, depending on the stimulation design, perforation may not have been conducted or recorded by the geophones.

When the seismic source and receiver are both located at nearly the same depth in low velocity shale, head wave arrivals can often be observed (Coffin et al., 2012; Zimmer, 2010). Researchers have recognized the possible presence of head waves before direct arrival. There are numerous examples in the crosswell (Dong and Toksöz, 1995; Parra et al., 2006, 2002) and

microseismic (Maxwell, 2010; Zimmer, 2010, 2011) literature where the head wave arrival is the first arrival. However, the head wave is often of weak amplitude and is commonly regarded as contamination of the direct arrival since it can impact the polarization estimation of the direct P-wave or be misinterpreted as the direct P-wave (Wilson et al., 2003). Synthetic studies using head waves have been conducted; however, there are few studies using field data on the improvement in event location obtained by using available head waves (Zimmer, 2010, 2011). Our analysis on microseismic data acquired in the Marcellus Shale shows that head waves convey useful information and can be used to constrain microseismic event location as a substitution for the P-wave polarization.

In this chapter, we first present the theoretical background of this study. We then give an overview of the microseismic survey in the Marcellus Shale. Next, we present and analyze the resonance in microseismic data acquired in the downhole survey. Subsequently, we show the head waves observed in the Marcellus Shale and use them to constrain microseismic event location as a substitution for direct P-wave polarization. Finally, we propose a new acquisition geometry to improve the traditional microseismic acquisition practice based on the location accuracy improvement due to the use of head wave arrival times.

4.2 Methods

Resonance Due to Poor Coupling

Geophone-borehole coupling is a concern in borehole geophysics surveys. The ground motion can be accurately recorded only if the geophone has no internal resonance and is well coupled to the borehole (Gaiser et al., 1988). However, due to operational limitation, this ideal situation is usually not achieved. In a borehole seismic survey, a geophone is coupled to the borehole with a locking mechanism, which is usually a locking arm in one direction. According to Gaiser et al. (1988), in a vertical well bore, the impulse response of a geophone is related to the contact width of a geophone with the borehole wall, the locking force, and the weight of the geophone. The resonance is usually most severe in the horizontal component that is perpendicular to the locking force direction. For a geophone placed in a horizontal well, the only coupling force between the geophone and wellbore is usually the gravitational force of the geophone itself. This can make the resonance due to poor geophone-borehole coupling even more severe.

The recorded noise-free seismogram due to a microseismic event or perforation shot can be expressed as the convolution of source wavelet, earth impulse response, and geophone response (including resonance due to poor coupling):

$$x(t) = w(t) * e(t) * r(t), \tag{4.1}$$

where $x(t)$ is the recorded seismogram, $w(t)$ is the source wavelet, $e(t)$ is the earth impulse response, and $r(t)$ is the receiver (geophone) response.

Its equivalent form in the frequency domain is

$$X(\omega) = W(\omega)E(\omega)R(\omega), \tag{4.2}$$

where $X(\omega)$, $W(\omega)$, $E(\omega)$, and $R(\omega)$ are the frequency domain representation of $x(t)$, $w(t)$, $e(t)$, and $r(t)$, respectively.

Deconvolution of Microseismic Signal

The effect of a receiver resonance can be attenuated with receiver channel consistent deconvolution (Claerbout, 1992; Yilmaz, 2001). The deconvolution improves the compactness of a seismic wavelet and can help in the identification of seismic phases by recovering the impulse response of the earth. Under the assumption that the impulse response of the earth, $e(t)$, is random ($|E(\omega)|$ is constant in the frequency domain), the seismogram has the same amplitude spectrum, $|X(\omega)|$, with the amplitude of the convolution of the source wavelet and the geophone response, $|W(\omega)R(\omega)|$. An additional minimum phase assumption enables the determination of an optimum Wiener filter, which can recover the impulse response of the earth from the recorded seismogram (Yilmaz, 2001). This can be used to remove the geophone resonance, thus, improve the identification of the multiple arrivals.

Head Wave

The generation mechanism of head waves in the Marcellus can be seen in Figure 4.1, which is a common acquisition configuration in shales. If the velocity of a nearby layer (the Onondaga Formation in this case) is larger than the shale, and assuming both source and receiver are located in the shale, head waves will be generated when the angle of incidence is equal to a critical angle $arcsin(\frac{V_1}{V_2})$, where V_1 and V_2 are the velocities of the low and high velocity layer, respectively, as shown in Figure 4.1. The head wave will then travel along the formation interface until the point where it refracts back to the original low velocity layer with angle of emergence at the critical angle. P-P-P, S-S-S, and S-P-P, and P-P-S converted head waves are potentially identifiable. In practice the three latter head waves are difficult to identify because they occur after the first arrival. Also, a dip-slip microseismic focal mechanism, which is often thought to be the dominant rock breaking mechanism, will preferentially generate P-P-P arrivals (Rutledge and Phillips, 2003). The direct arrival amplitude is inversely proportional to the distance that the seismic ray traveled from the source due to geometrical spreading, while head wave amplitude is approximately inversely proportional to the square of this distance (Červenỳ and Ravindra, 1971). Thus, the head wave will decay faster than the direct arrival and usually has smaller amplitude. As in refraction seismology, though the head wave travels a longer path than the direct arrival, it arrives before the direct arrival past the cross-over distance. Figure 4.2 shows traveltime versus source/receiver separation for the configuration in Figure 4.1.

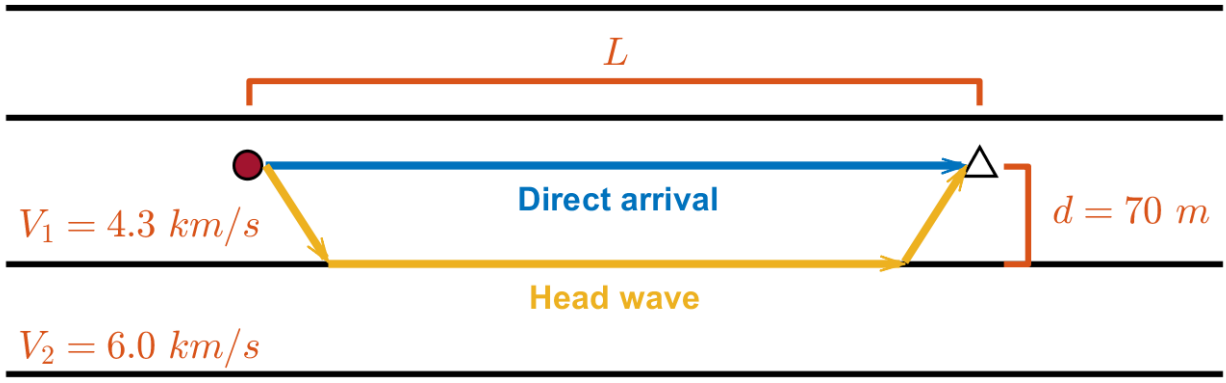

Figure 4.1: A common configuration for a head wave. Due to the low velocity nature of shale, the head wave is commonly identified when there is a nearby high velocity layer.

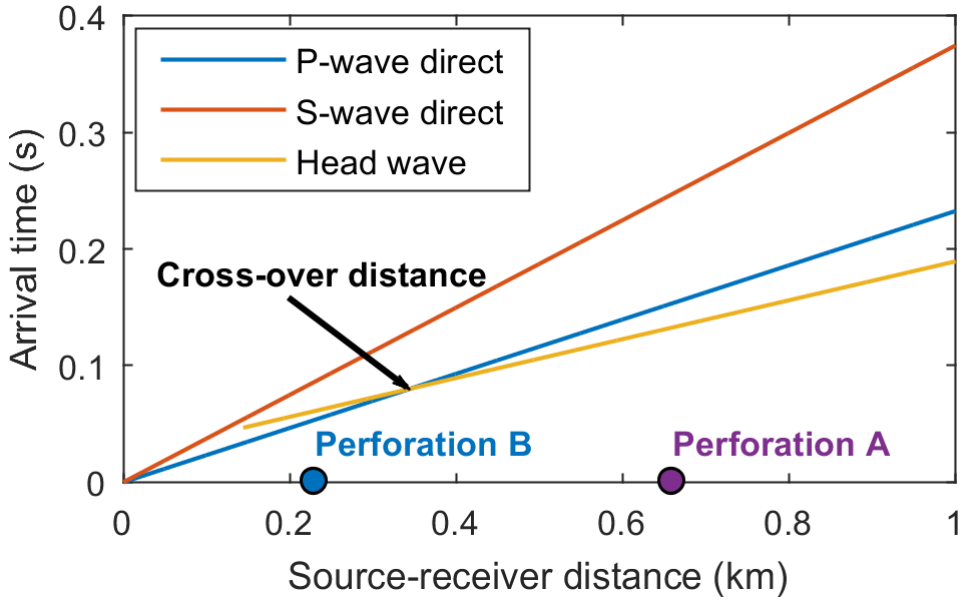

Figure 4.2: Arrival time of various phases as a function of the source-receiver distance. When the source-receiver distance is larger than the crossover distance, the head wave can overtake the direct arrival to be the first arrival. Perforation A and Perforation B are two shots with a source-receiver distance larger and smaller than the crossover distance, respectively.

Event Location Estimation and Velocity Model Calibration

The velocity model calibration and microseismic event location estimation were conducted with a microseismic event location program we previously developed (Zhang et al., 2017b). It aims to minimize the misfit between the observations, which include arrival times and polarization directions, and the model predictions of these observations. An objective function is minimized iteratively with a Gauss-Newton method (Zhang et al., 2017b). The standard deviation of arrival time picking uncertainties is assumed to be 1 ms for all phases and P-wave polarization uncertainty is assumed to be 6°. Similarly, the velocity model can be calibrated with perforation data by minimizing the objective function with respect to velocity model parameters instead of the microseismic event locations and origin times.

4.3 Hydraulic Fracturing Project Overview

The hydraulic fracturing project was carried out in the Marcellus Shale in Susquehanna County, Pennsylvania, within the Susquehanna River Basin. The Marcellus Shale is a Middle Devonian age unit of marine sedimentary shale that contains largely untapped natural gas reserves. It underlies the Mahantango Formation (siltstone and shale) and overlies the Onondaga Formation (limestones and dolostones). Its natural gas trend is the largest source of natural gas in the United States. The Marcellus Shale in the studied area has a thickness of roughly 46 m (150 ft) and the average porosity and permeability are 0.08 and 600 nanodarcy, respectively.

A multiple well pad that includes seven nearly parallel horizontal wells is the site of field acquisition (Salehi et al., 2013). The trajectories of the lateral wells are normal to the maximum in situ horizontal stress orientation. The horizontal distances between two nearby lateral wells are approximately 152 m (500 ft) and the average horizontal wellbore length is 1109 m (3640 ft). The true vertical depths (TVDs) of the wells are approximately 1981 m (6500 ft). The target zone of the wells lies along the lower portion of the Marcellus Shale. One of the major purposes of the hydraulic fracturing project was to evaluate the potential to increase stimulation efficiency (increased production, reduced water consumption per unit of gas produced, and reduced environmental footprint) by varying the pump rate. Microseismic data has been acquired and analyzed. Surface microseismic tools were deployed in an approximately 7.8 km^2 (3 square miles) area and 93 stimulation stages were monitored. Downhole geophones were placed in one of the horizontal wells and 62 stimulation stages were monitored. A previous study observed increased microseismicity during hydraulic fracturing in stages with frequent pump rate changes, which suggests better stimulation efficiency (Ciezobka et al., 2016).

Our study is focused on two wells, a monitor well and a stimulation well, as shown by Figure 4.4. The lengths of the horizontal portion of the two wells are 1350 m (4430 ft) and 1700 m (5577 ft), respectively. The average distance between the horizontal portions of the two wells is around 220 m (722 ft). The stimulation started from the toe and continues until

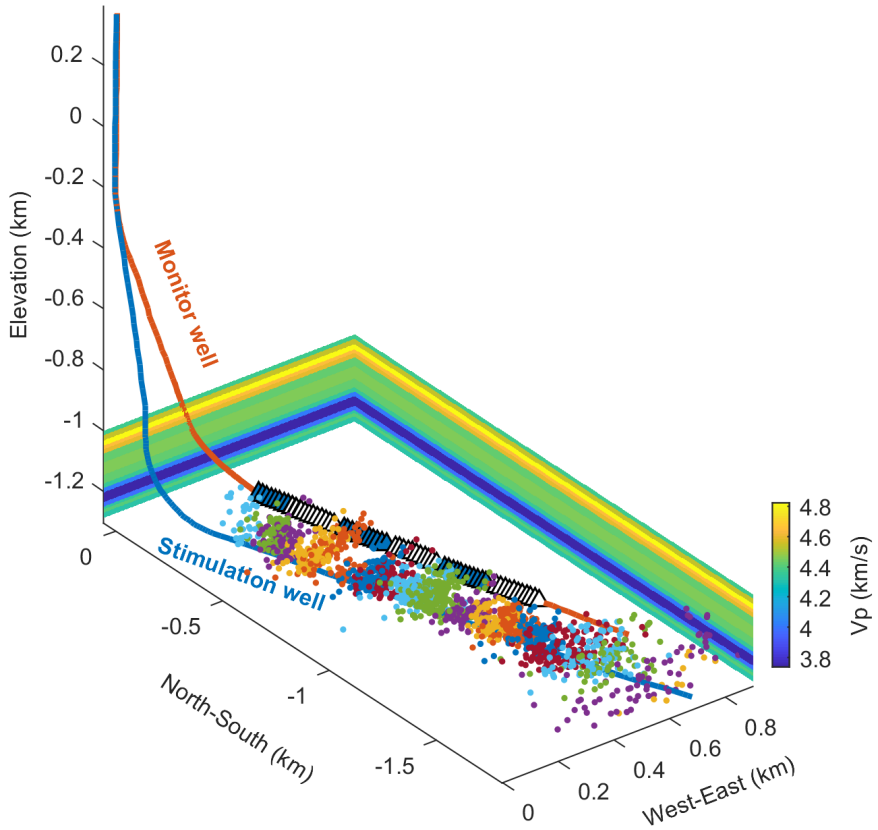

Figure 4.3: Microseismic survey geometry. The microseismic event locations (dots) were located conventionally using P-, S-wave arrival times and P-wave polarization directions. The alternating white and blue geophone arrays are different locations of the same array that is used to monitor the stimulation. The stimulation stages and their corresponding geophone array positions are shown in Figure 3.1. Microseismic events are color-coded according to their associated stimulation stages.

reaching the heel of stimulation well. It consists of 18 stages with an interval of 91 m (300 ft), as shown by Figure 4.4. We refer to the stimulation stages as Stage 1 to Stage 18 from the toe to the heel of the well. Among these stages, nine were designed to have variable pump rate and nine used the traditional constant rate design. Each stage consists of four perforation shots with a perforation interval of 21 m (70 ft). We refer to the shot on the side of the toe as Perforation 1 and the shot on the side of the heel as Perforation 4 in each stimulation stage. The fracture stages alternated along the horizontal wellbore to account for changes in the reservoir and natural fractures.

The microseismic survey was conducted with an array of 11 three-component 10 Hz geophone tools. The tool spacing in the array was 15.2 m (50 ft). The geophone on the side

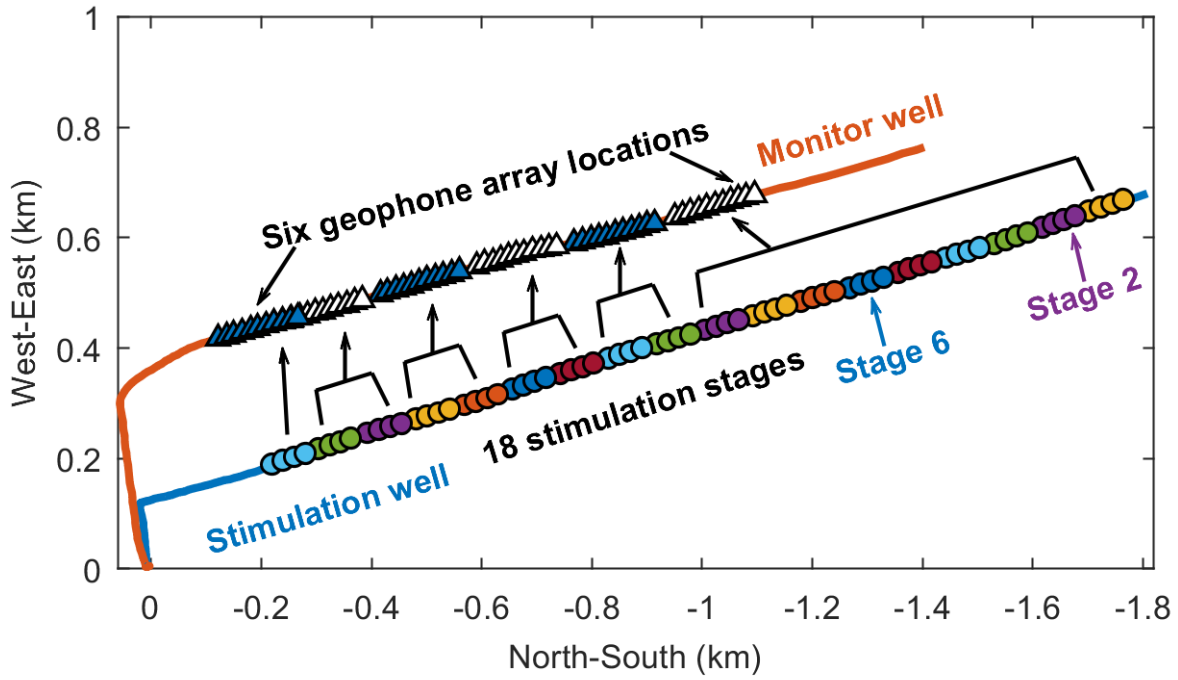

Figure 4.4: Map view of the acquisition geometry. The stimulation was performed in 18 stages and the microseismic signal was recorded by an array of 11 geophones in the nearby monitoring well. The geophone array was moved according to the stimulation stage location to reduce errors due to large event to receiver distances.

of the toe is referred to as Geophone 1 and the geophone on the side of the heel is referred to as Geophone 11. The tools were deployed via tractor in the horizontal section of the borehole, and the only coupling between the tool and the borehole wall was due to gravity. As is typical in these types of surveys, the tool array was moved along the monitor well bore to be roughly across from the stimulated zone in the treatment well, thereby reducing travel path length to improve S/N and event location accuracy.

A total of 1842 events were detected and processed during the 18 stimulation stages. The number of events in each stage is shown in Table 4.1. In addition to these microseismic events, perforation shots from Stage 2, 6-9, 12-14, and 17-18 were recorded by the geophone array and used for velocity model calibration and location uncertainty analysis. An isotropic 1D velocity model was created based on a sonic log from the vertical section of the stimulation well and then calibrated with perforation shots, as shown in Figure 4.3. The geophone orientations were estimated using the P-wave polarization directions from the perforation shots. P-, S-wave arrival times were manually picked and used for the initial microseismic event location. P-wave polarization directions were also used to constrain microseismic event locations. The microseismic event locations obtained from this analysis are shown in Figure

4.3 and are color-coded with their corresponding stimulation stages.

Table 4.1: Number of microseismic events in each stage.

Stage	Number of Events	Stage	Number of Events
1	11	10	224
2	66	11	168
3	63	12	94
4	93	13	141
5	130	14	101
6	106	15	120
7	141	16	80
8	120	17	70
9	80	18	34

4.4 Data Analysis

Figure 4.5 and Figure 4.6 show a typical perforation shot (the second perforation shot) and a typical microseismic event waveform from stimulation Stage 6, respectively. Examination of the microseismic data acquired in this survey shows frequency resonance in both the axial (with respect to the borehole) and radial components of the data. The perforation shot data are also affected by channel-dependent resonances. By visual inspection, it can be seen that the characteristic of the resonance is dependent on the channel instead of the source mechanism.

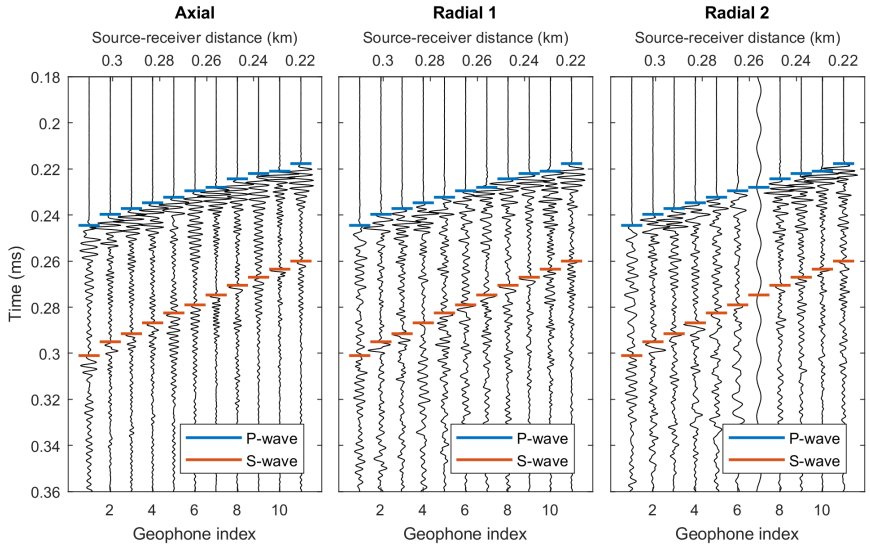

Figure 4.5: Waveforms of a typical perforation shot from stimulation Stage 6. The waveforms of a perforation shot are usually P-wave dominated due to the source mechanism of perforation shot. Severe resonance effect in waveforms can be observed, especially in the axial component.

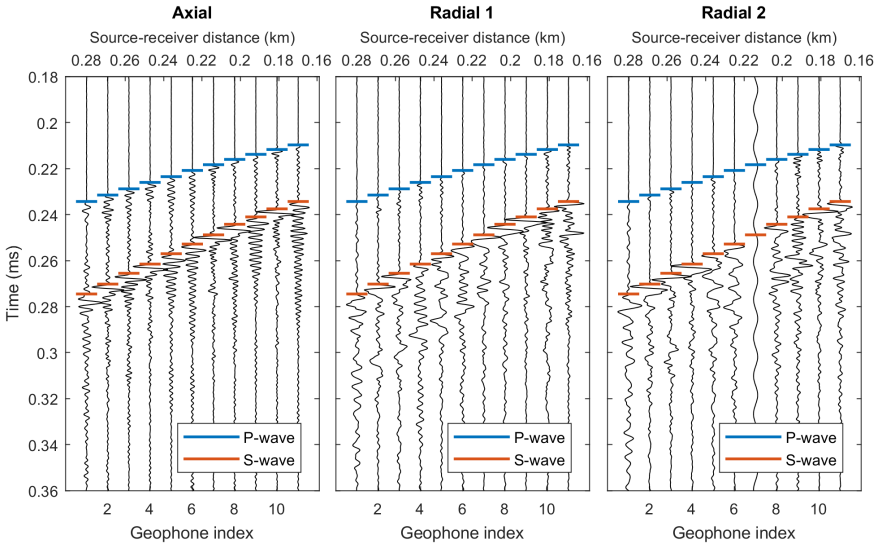

Figure 4.6: Waveforms of a typical microseismic event from stimulation Stage 6. The waveforms of a microseismic event are usually S-wave dominated.

Spectrum of the Resonance

The spectrum of the resonance can be seen from a short-time Fourier transform (STFT) of the three component waveforms recorded by Geophone 5 as shown by Figure 4.7. For the axial component, the resonance frequency is around 420 Hz. The first radial component has resonance frequencies of 120 Hz and 440 Hz. The second radial component resonates at 120 Hz and 340 Hz. Gaiser et al. (1988) show that the resonance due to poor geophone-borehole coupling is mainly on the radial component instead of the axial component. This is the character of the resonances at frequencies around 120 Hz and 340 Hz. The fact that the only coupling force between the geophone and the wellbore is the gravitational force of the geophone in the horizontal well is likely the reason for the resonance in both radial components. The resonance above 400 Hz is polarized on the axial and the first radial components and may result from the resonance of the geophone themselves. Resonance will create problems for tasks such as Q value estimation, waveform inversion, and P-wave polarization direction estimation. In the presence of resonance, additional processing procedures should be performed such as the relative spectrum analysis introduced by Zhang et al. (2016).

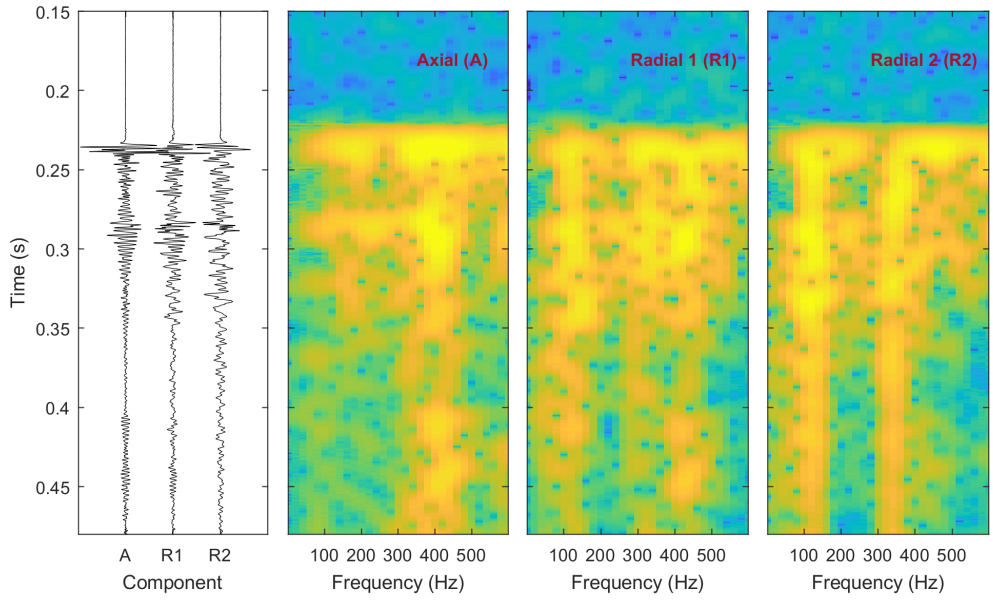

Figure 4.7: STFT of a typical three-component waveform generated by a perforation shot. For the axial component, the resonance frequency is around 420 Hz. The first radial component has resonance frequencies of 120 Hz and 440 Hz. The second radial component resonates at 120 Hz and 340 Hz. The resonance around 120 Hz may be due the poor coupling between geophone and wellbore. The resonance above 400 Hz may result from the geophone themselves.

Deconvolution of Microseismic Signal

The presence of resonances in microseismic signals may negatively impact the identification of seismic phases. We performed a spiking deconvolution to remove the receiver signatures in these waveforms. An optimum Wiener filter was designed using the average autocorrelation of the four perforation shots in Stage 6. The waveforms before and after deconvolution are shown in Figure 4.8. From the comparison, we can see a significant suppression of the resonance following the P- and S-wave arrivals after the deconvolution. This suppression prevents the later phases from being contaminated by resonance due to earlier arrivals. For instance, it can be difficult to determine the S-wave arrival times on Geophone 5 and 9 in Figure 4.8a due to their preceding resonance. After the removal of the resonance (Figure 4.8b), it is significantly easier to pick those arrivals on Geophone 5 and 9. In addition, we also find two weak, yet clear phases after the deconvolution denoted by multiple 1 and multiple 2 in Figure 4.8b. These two arrivals can hardly be identified in the original data.

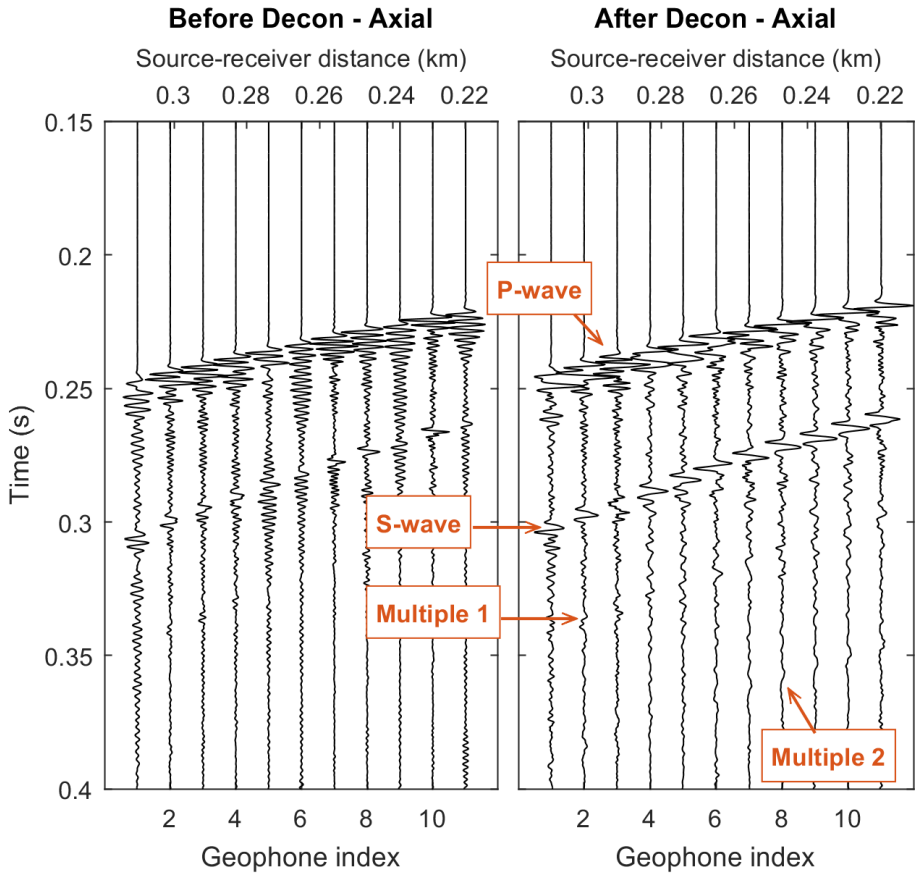

Figure 4.8: Deconvolution result of the axial component. The deconvolution successfully suppressed the resonance in the original data. In addition, it enhances multiple arrivals that are hardly identified in the original waveform.

4.5 Event Location Result

Due to the azimuthal ambiguity in microseismic event location using only P- and S-wave arrival times, P-wave polarization is commonly used to constrain the azimuthal direction of microseismic events. However, the effect of resonance on the downhole geophones may result in large uncertainty in P-wave polarization estimation. In addition, the orientations of downhole geophones will require calibration using information from perforation shots, which may be unavailable. Due to the low velocity nature of shale, the head wave is commonly identified in microseismic surveys (Maxwell, 2010; Zimmer, 2010, 2011). Like many other microseismic surveys, we observed head waves in the Marcellus Shale. Figure 4.9a shows the axial component of the waveforms for perforation shot 4 in Stage 2 (Perforation A in Figure 4.10). The head wave arrivals have low amplitude and high velocity moveout as annotated by the yellow picks in Figure 4.9a. However, as shown by Figure 4.9b, the waveform for perforation shot 3 in Stage 6 (Perforation B in Figure 4.10) shows no identifiable head wave since its source-receiver distance is smaller than the cross-over distance. In this section, we use the head wave arrival times as a substitution for the P-wave polarization to constrain the microseismic event locations.

For a microseismic event at a distance of L from the observation geophone array, the location uncertainty due to uncertainty in polarization will be on the order of αL, where α is the uncertainty of P-wave polarization estimation. A common value of $\alpha = 6^{\circ}$ and $L = 400$ m (1312 ft) will result in a location uncertainty of 42 m (138 ft). This is a value significantly larger than the location uncertainty resulting from arrival time picking uncertainty, which is usually on the order of several meters. Additional uncertainty usually comes from velocity model uncertainty; however, it is common for both methods.

Velocity Model Calibration

Since the original velocity model is a model based on sonic logs and calibrated with perforation shots, it is limited to the TVD of the kickoff point (sonic logs are not typically run in the horizontal section). According to this provided model, the head wave will not take over the direct P-wave to be the first arrival as observed in the waveform within the offset ranges in this study. To calibrate the velocity model, perforation shots were used and P-, S-, and head wave arrival times were picked. From the calibrated velocity model, we found that Marcellus velocities near the stimulated interval were close to the one provided by the contractor. The calibration also reveals the existence of a high velocity ($V_p = 6.01 km/s$) formation, Onondaga Formation, underlies approximately 70 m (230 ft) below the geophone array. However, there was no velocity information in the original model due to lack of sonic logs.

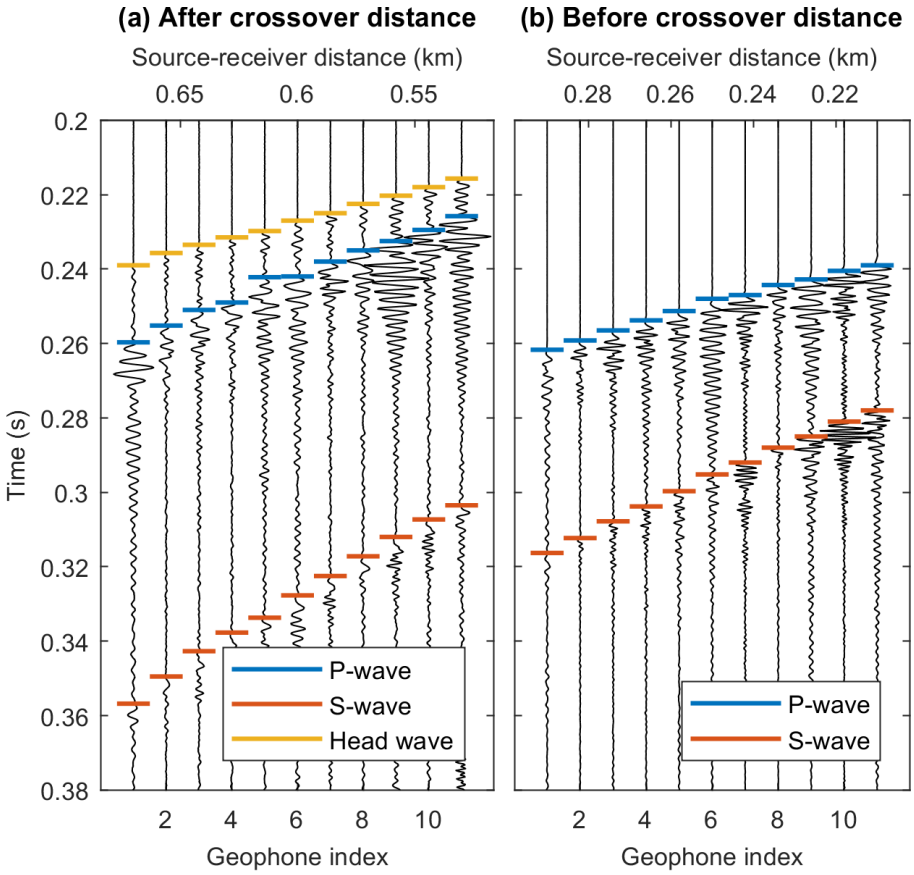

Figure 4.9: The axial component of the waveforms of perforation shots after (a) and before (b) the cross-over distance. Head waves can be easily identified based on their low amplitude and high velocity moveout from waveform (a). The head waves arrive after the direct P-wave; thus, cannot be identified in waveform (b). The location of the perforation shots are shown in Figure 4.10.

Finite Difference Simulation

To further verify the existence of head waves and the calibrated velocity model, we conducted a finite difference simulation to investigate the wave propagation of microseismic signals with SW4, a 3D elastic forward modeling code (Petersson and Sjogreen, 2013). The code implements a fourth order accurate method in space and time. The focal mechanism of the source is assumed to be a vertical crack with a moment tensor proportional to

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\nu} - 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.3}$$

where ν is Poisson's ratio.

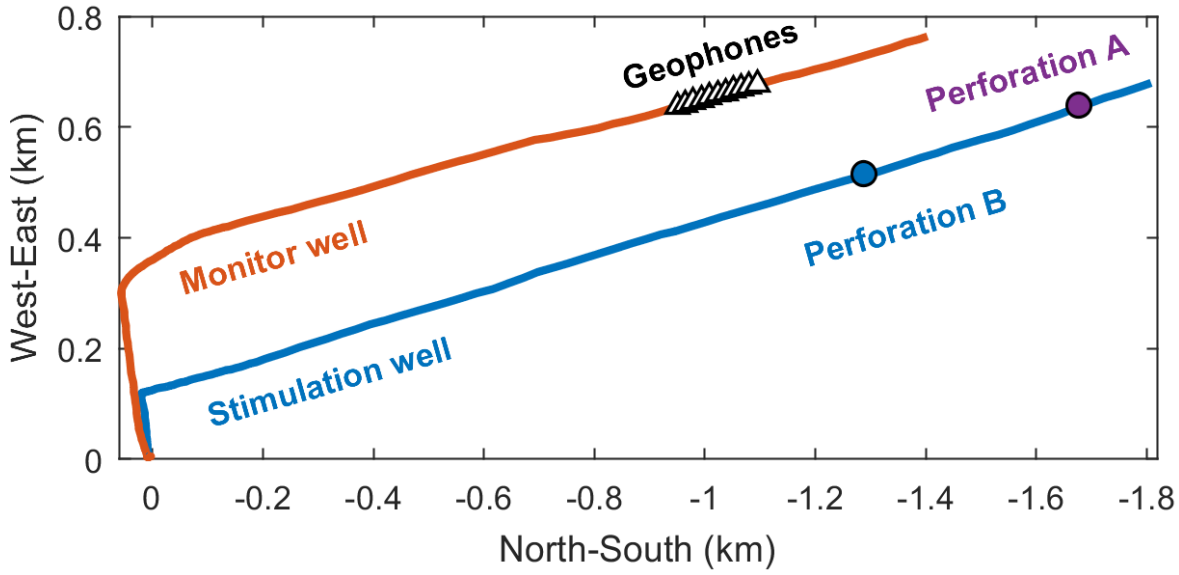

Figure 4.10: The locations of two perforation shot whose waveforms are shown by Figure 4.9.

The source time function is assumed to be a Ricker wavelet with peak frequency at 100 Hz. The existence of head waves can be verified by the comparison between field and synthetic waveform as shown by Figure 4.11. The arrival time of the head wave in field data matches that of the synthetic result well. In addition, the low amplitude ratio between P- and head wave is also verified by the synthetic simulation. The differences in the S-wave in the V_x and V_y components may be due to the lack of knowledge of the source mechanism of the actual event for the finite difference simulation.

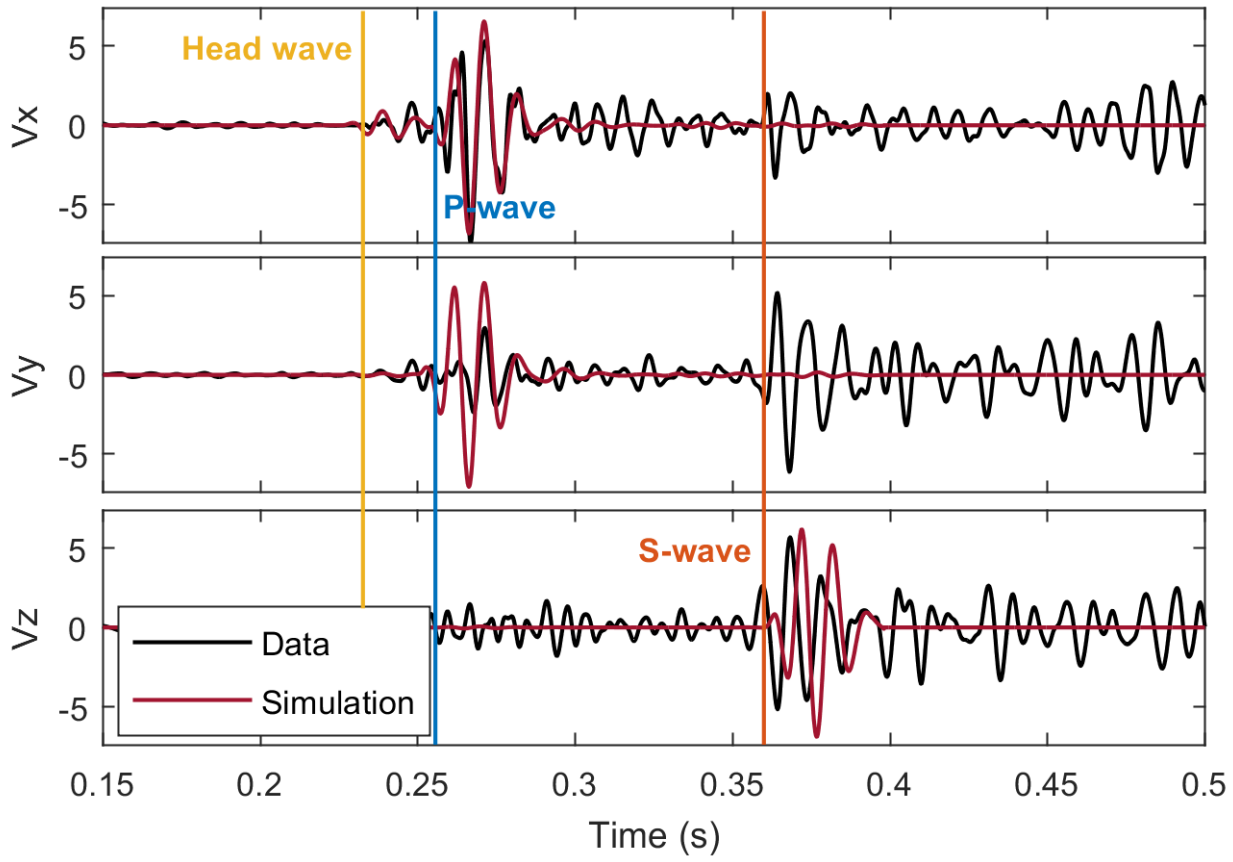

Figure 4.11: Comparison between synthetic and field waveform. The synthetic waveform matches the field data relatively well, which verifies the existence of head wave. The difference between the S-wave in the x and y components may be due to the unknown source mechanism of the actual event for simulation.

Perforation Shot Location

To quantify our event location estimation uncertainty, we located the perforation shots in Stage 2 with a Jackknife technique (Miller, 1974). That is, for each perforation shot, its location is estimated with the velocity model calibrated with the other three perforation shots. Since the velocity model was not calibrated with the perforation shot to be located, these perforation shots in Stage 2 can be treated as normal microseismic events and used for location uncertainty analysis. Our location result of the four perforation shots along with their true location is shown in Figure 4.12. What is also shown is the location result with the traditional method, which used direct arrivals and P-wave polarization directions.

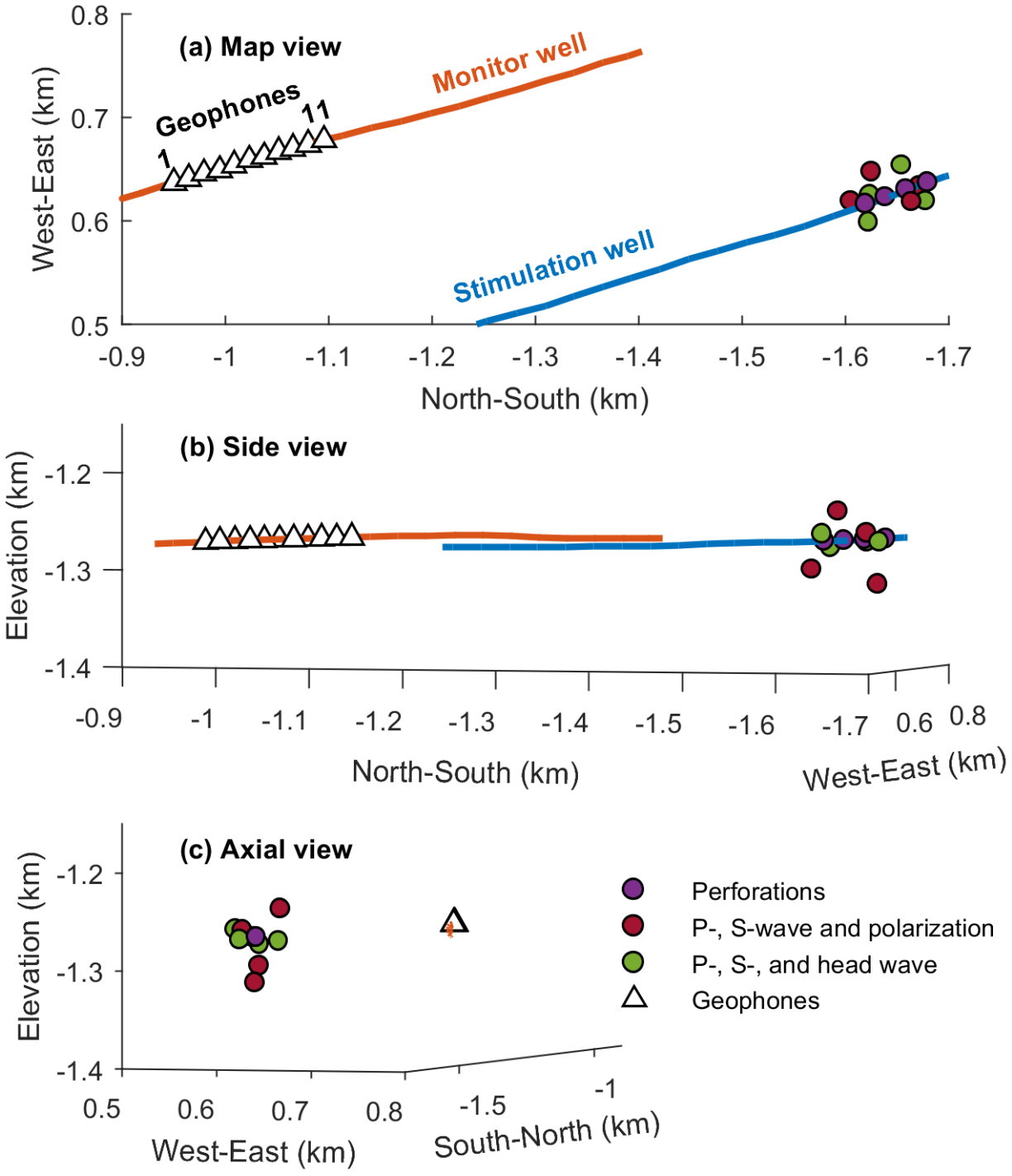

Figure 4.12: Comparison of estimated perforation shot locations and the true perforation locations. Location estimation using head wave arrival times gives a RMS error of 19 m while the traditional method using P-wave polarizations gives a RMS error of 52 m.

From the comparison, we found the method using head wave arrivals instead of P-wave polarizations gives a root mean square (RMS) error of 19 m (62 ft) while the traditional method with P-wave polarizations and P-, S-wave arrival times gives a RMS error of 52 m (171 ft). Given the limited acquisition geometry and relatively large source-receiver distance in this survey, the method using head wave arrival times gives a plausible result while the traditional method using P-wave polarization directions leads to relatively large location uncertainty.

Relocation of Events in the Second Stage

A map view of the microseismic event locations estimated with the traditional P-wave polarization method is shown in Figure 4.13. Note that the microseismic event locations in Stage 2 are significantly more scattered than those in later stages. One possible explanation to this scattering is because of the larger stimulated reservoir volume associated with Stage 2 stimulation. However, an alternative explanation is simply because of the larger event location uncertainties in Stage 2 events due to the longer travel paths of seismic rays.

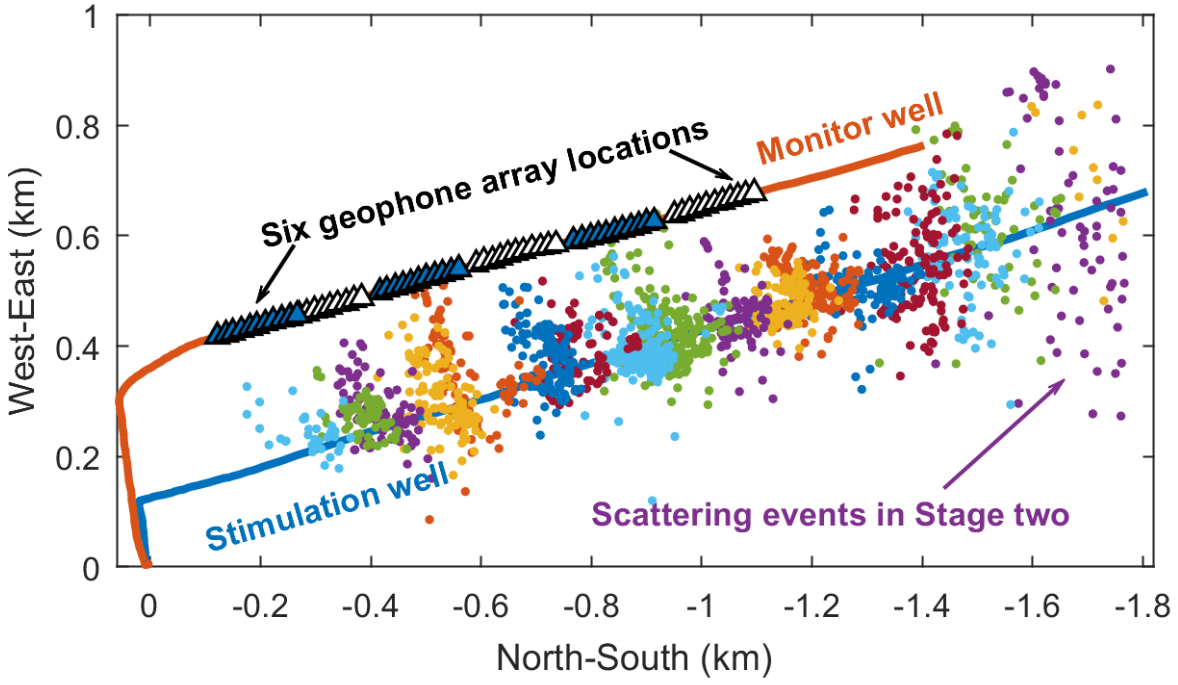

Figure 4.13: Map view of microseismic event locations processed using P-, S-wave arrival times and P-wave polarizations. The event locations in Stage 2 are much more scattered than those in later stages.

We relocated these events using direct P-, direct S- and head wave arrivals without

polarization as shown in Figure 4.14. The relocated events are much less scattered than the result estimated with the traditional location method. This pattern is more consistent with the microseismic event patterns in the later stimulation stages and indicates the effectiveness of using head wave arrival times in microseismic event locations to improve event location accuracy.

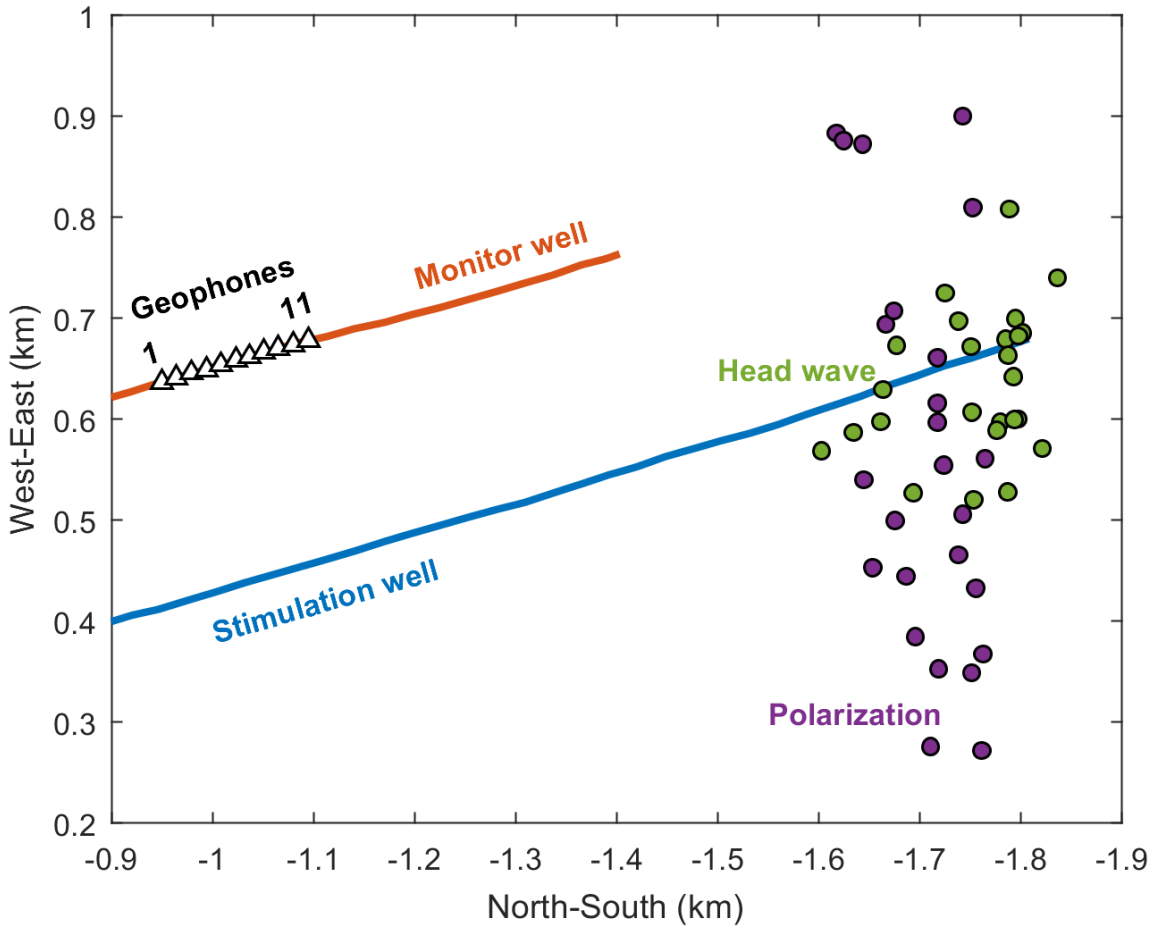

Figure 4.14: The microseismic event locations estimated with P-, S-, and head wave arrival times are less scattered and more consistent with other stimulation stages when compared with the microseismic event locations processed using the traditional location method.

4.6 Discussion

The microseismic event location methodology developed in this study relied on head wave availability. However, the head waves exist only if a high velocity layer is present in the

vicinity of the stimulation zone and the observation geophones. Even so, they can hardly be identified if they arrive after the direct arrivals, which is the case when the source-receiver distance is smaller than the cross-over distance.

When the source-receiver distance is smaller than the cross-over distance such as the data in Figure 4.9b, which comes from perforation shot B in Figure 4.10, head waves will arrive after the direct P-waves (Figure 4.2). In this case, it will be more difficult to pick head wave arrivals, and conventional methods of event location using P-wave polarization directions may be required to constrain the event locations. Traditional acquisition practices place the geophone array as close as possible to the stimulation zone. However, our analysis shows this practice may result in loss of information with multiple arrivals. We would propose to place the geophone array farther than a cross-over distance for single horizontal well monitoring as shown by Figure 4.15. This acquisition geometry will enable the identification of multiple arrivals and will therefore improve microseismic event location accuracy. Moreover, fewer moves (perhaps no moves whatsoever) may be required to provide accurate location information. Significant reductions in acquisition cost and wellbore risk might be achieved with this geometry without sacrificing accuracy and in some situations perhaps improve location accuracy.

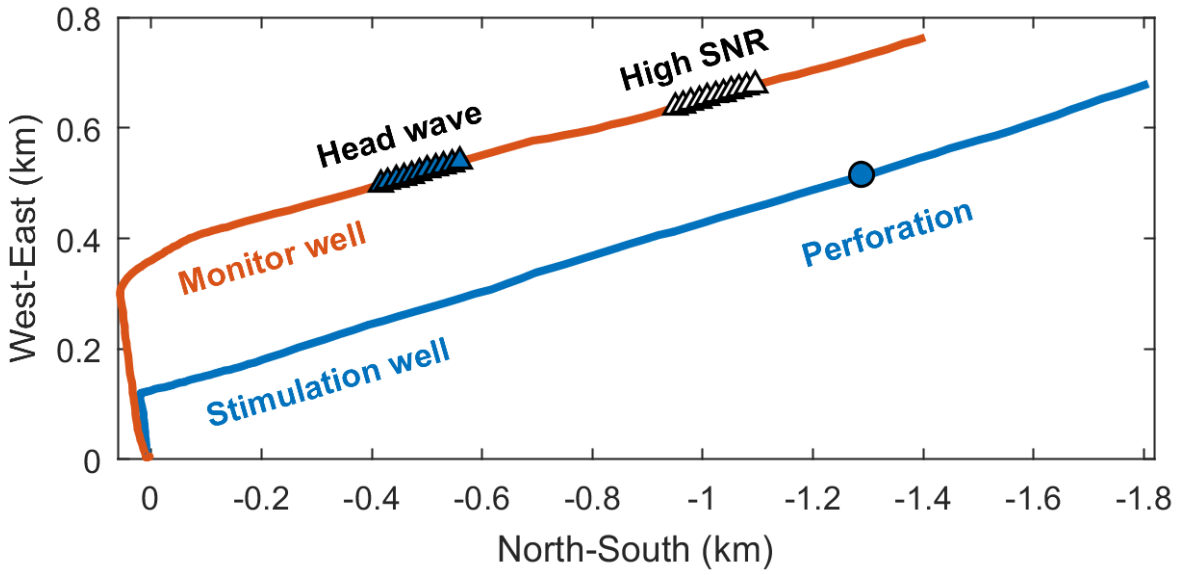

Figure 4.15: Traditional acquisition geometry aims at improving S/N by decreasing source-receiver distance (white geophone array). Our study shows that one can monitor hydraulic stimulation with geophone array that is farther than a cross-over distance (blue geophone array) for head wave observation. This acquisition practice will be able to avoid large location uncertainty due to using P-wave polarization as well as to reduce acquisition cost.

4.7 Conclusion

Resonance due to poor geophone-borehole coupling is commonly observed in downhole microseismic surveys. Deconvolution is successful in removing resonance and improves the identification of multiple arrivals. However, it will not help to improve the P-wave polarization estimation, which is traditionally used to constrain microseismic event location in single monitoring well observation. The existence of head waves in microseismic survey of Marcellus Shale is observed and verified. The location result of perforation shots using the developed method verified that, whenever available and identifiable, accounting for head wave arrival time as a substitution for P-wave polarization indeed improves the microseismic location accuracy. Based on the developed method, we propose an improved acquisition geometry for single horizontal well hydraulic fracturing monitoring, which enables us to improve the identification of multiple arrivals, utilize the head wave as the first arrival, and improves microseismic event location accuracy as well as reduce acquisition cost.

Chapter 5

Recovering Compressional Wave Amplitudes via Machine Learning

I must study politics and war, that my sons may have the liberty to study mathematics and philosophy... in order to give their children the right to study painting, poetry, and music.

– John Adams

5.1 Introduction

In this chapter, as well as Chapter 6, we continue to strive for understanding of where poorly conditioned data negatively impact our ability to garner insights. However, we incorporate a more interdisciplinary approach through the use of machine learning and artificial intelligence. Here, we rely on the knowledge gained from previous work in order to shift focus to data-driven solutions while relying on data science to augment our work that relied on more purely geophysical techniques.

Recent advances in drilling technology have led to a significant increase in the exploration of unconventional resources via hydraulic fracturing (fracking) in shale plays in order to recover natural gas and other hydrocarbons. Monitoring of microseismic events can be performed downhole or at the surface. Downhole monitoring provides significantly better resolution and signal-to-noise ratio (S/N) than surface monitoring approaches (Maxwell et al., 2010). However, one of the limitations of downhole monitoring is that in order to minimize the source-receiver distance, and thereby improve S/N, it is necessary to sacrifice azimuthal coverage of the monitoring area. As such, moment tensor inversions have significant uncertainty (Nava et al., 2015). Furthermore, when a single observation well is used, there is a large reliance on multiple compressional (P) and shear (S) wave arrival times for location estimation and other analytic objectives (Warpinski et al., 2009). Moreover, it has been

shown that the P-wave amplitude is fundamental to invert focal-plane mechanisms in cases where azimuthal coverage is limited (Kuang et al., 2017).

While there are a number of works on imputation and the use of machine learning models for imputation tasks, there are a limited number of papers focusing on geophysics and far fewer on microseismic analysis (Gill et al., 2007; Haukoos and Newgard, 2007; Kondrashov and Ghil, 2006). Cawley and Talbot (2010) explore interpolation of missing data in seismic traces with nonstationary prediction-error filters (PEF). The PEFs are first estimated and are then used to fill missing trace bins as part of linear least squares. This approach also enables the separation of noise and real signal. Kondrashov and Ghil (2006) focus on the use of Singular Spectrum Analysis (SSA) to fill in missing information in both space and time in a number of synthetic data sets as well as data sets from oceanographic, hydrology, and space physics. Multiple methods were compared and an effective cross-validation method was employed to validate the results of the gap-filling approach on various signals. Gill et al. (2007) employ artificial neural networks (ANN) and support vector machines (SVM) to predict groundwater levels over a short-term period at a specific well field and explore the overall impact of missing data on these learning algorithms. A local least squares method of imputation was used and the effect of varying amounts of missing data was explored. Finally, performance of each learning algorithm was compared over the range of missing data, up to 30% of the overall data set. The results showed that for groundwater estimation, the SVM algorithm performed well despite large amounts of missing data. However, a known drawback to SVM algorithms, like other kernel methods, is the risk of sensitivity to overfitting, which makes reuse of a model difficult across various data sets (Cawley and Talbot, 2010). Additionally, SVM algorithms are typically employed for classification problems and not regression tasks for estimation of continuous variables.

Ensemble methods like random forest are particularly robust to overfitting and can be used for classification, regression, and survival analysis (Breiman, 2001). Moreover, random forest models are one of the few machine learning techniques that can be successfully executed with input data that contain missing values. Additionally, recurrent neural networks, particularly long short-term memory networks (LSTM), have the advantage over traditional ANN in that they are well-equipped to deal with nonlinear, time-dependent data (Monner and Reggia, 2012). Multivariate Imputation by Chained Equations (MICE) is an incredibly powerful method of imputation, and an open-source package in R, that enables imputation of more than one variable for both categorical and numerical data (Buuren and Groothuis-Oudshoorn, 2010). These methods are data-driven and are agnostic regarding the nature of the data. As such, these learning models can be used for a number of learning objectives to include imputation of microseismic parameters.

In this chapter, we explore the applicability of machine learning and deep learning methods for the explicit purpose of imputing missing information from a real microseismic data set. The main motivation is to realize the benefit of data-driven approaches in the hydraulic fracturing domain that do not rely on signal processing techniques, but rather identify relationships from aggregate parameters common to typical microseismic workflows in use today. We begin with an overview of the hydraulic fracturing project where the data set

under investigation was gathered, provide a short discussion on common sources of data corruption in microseismic data sets, and explore traditional imputation methodologies. Then, four imputation approaches are presented: stage-specific median imputation, random forest imputation, Multivariate Imputation by Chained Equations (MICE), and long short-term memory networks (LSTM). Model performance for each imputation approach is presented and compared when applied to a real data set. Finally, we conclude with a discussion of the various applications of the best performing methods investigated.

5.2 Microseismic Survey in Marcellus Shale

The hydraulic fracturing project considered in this chapter was executed in the Marcellus Shale located in the Susquehanna River Basin in Pennsylvania, where the principal material found is Middle Devonian aged sedimentary shale. The Marcellus Shale is flanked by the Mahantango Formation (above), which is mostly siltstone and shale, and the Onondaga Formation (below), which is limestones and dolostones (Zhang et al., 2017a). It is one of the largest in the world in both volume and production content. As such, it is one of the greatest sources of natural gas in the United States. The average thickness of the Marcellus Shale in the survey area was approximately 46 m (150 ft) with a porosity of 0.08 and permeability of 600 nanodarcy.

Survey Geometry

Figure 5.1 shows two horizontal wells that were considered, a treatment well and an observation well with horizontal lengths of 1700 m and 1350, respectively. There were 18 stimulation stages beginning at the toe and ending at the heel of the treatment well with an interval of 91 m. Stage two is considered for analysis based on the survey geometry and data quality (Zhang et al., 2017a).

The downhole monitoring array consisted of 11 three component, 10 Hz geophones with 11.2 m (50 ft) spacing between each geophone. In order to accommodate the relocation of the sensor array, the geophones were deployed by tractor, which allowed for subsequent moves for the last nine stages of the project. This is common in downhole monitoring projects in order to improve S/N by locating the geophones perpendicular to the current stage, thereby minimizing the source-receiver distance. There were a total of 1842 contractor-identified microseismic events over 18 stages. While all of these events satisfied some threshold for quality and are categorized as microseismic events, it is clear that there is a non-negligible amount of missing data.

5.3 Sources of Data Loss

There are many causes of data loss inherent to microseismic monitoring processes. For example, scattering of microseismic energy due to near field and far field effects is a naturally

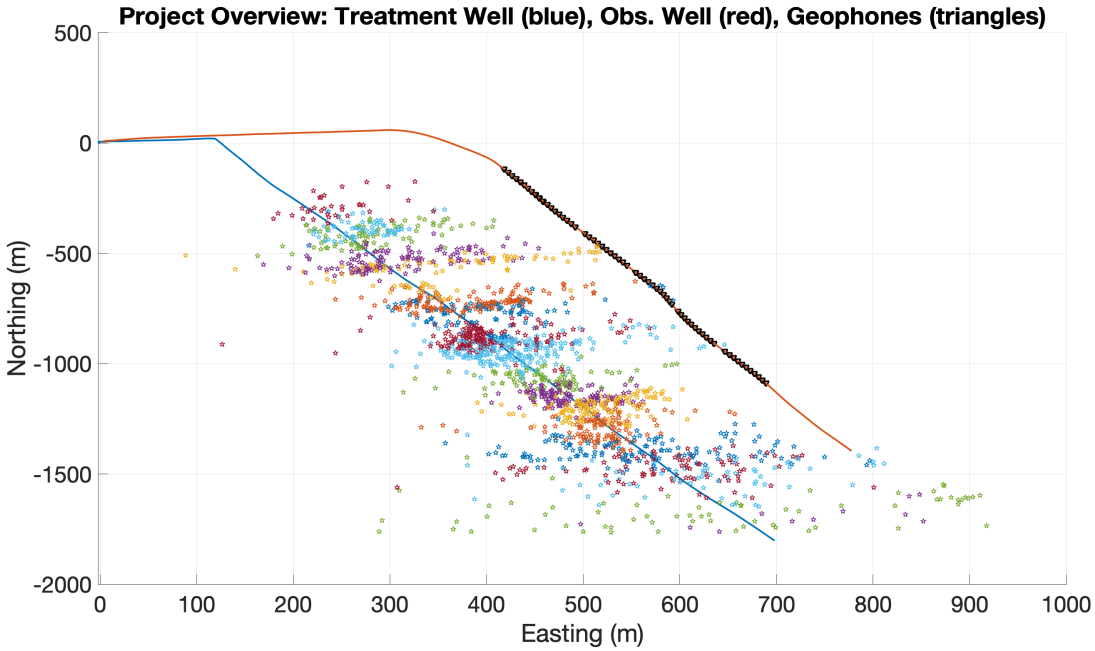

Figure 5.1: Map view of hydraulic fracturing geometry showing fracture stages and geophone locations. Inverted triangles show the different locations of the geophone array in the observation well, shown in red. The locations of microseismic events are shown around the blue treatment well color-coded for each stage.

occurring phenomenon that physically inhibits the recording of fracture events (Schoenberg, 1980). Additionally, shear waves, unlike compressional waves, cannot propagate through fluid-filled regions. As such, microseismic energy from a fracture occurring on the far side of a fracture network filled with fracking fluid will lose the majority of the shear component when recorded (Quintal et al., 2012).

The orientation of geophones can also have a significant impact on data loss and this effect may be amplified during each relocation of the geophone array. Moreover, resonance due to poor coupling between the geophone and borehole has been shown to reduce data quality, shown in Figure 5.2. Specifically, there is an inverse correlation between locking force, or the force exerted in order to maximize contact between the geophone and borehole wall, and the presence and amplitude of resonance (Gaiser et al., 1988). Furthermore, the area of contact between geophone and borehole wall, as well as geophone weight, are significant factors in the presence of resonant energy. Resonance is a destructive signal that is captured in (5.1) as part of the geophone response. Equation (5.1) describes a noise-free microseismic event and is the convolution of source wavelet, impulse response of the earth, and finally, geophone response as follows:

$$x(t) = w(t) * e(t) * r(t) \tag{5.1}$$

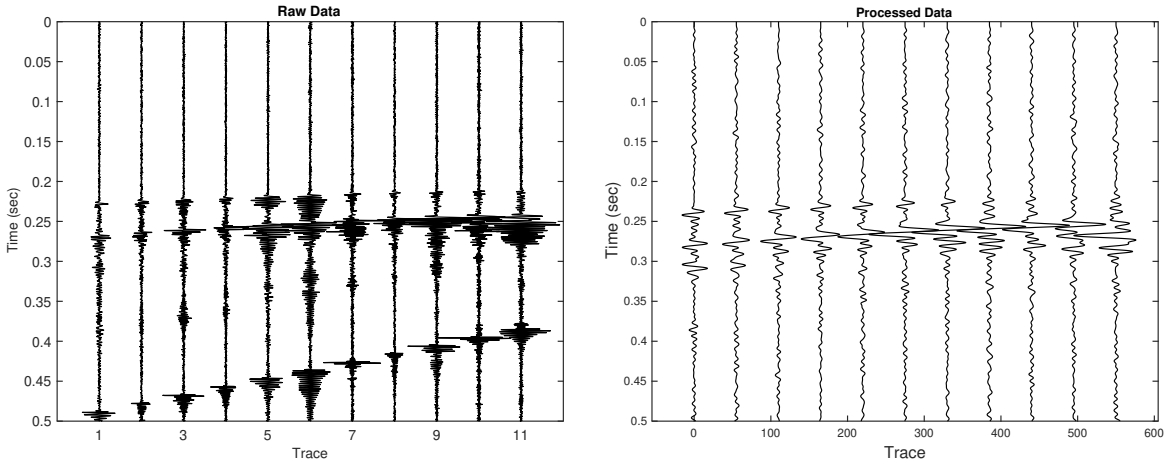

Figure 5.2: Real microseismic event recorded from the Marcellus Shale. Raw event (left) shows presence of resonant noise. Processed event (right) shows that this noise is effectively removed.

where $x(t)$ represents the recorded seismogram, $w(t)$ denotes the source wavelet, $e(t)$ denotes the earth impulse response, and $r(t)$ is the receiver response. Here, it is important to note again that the receiver response, or geophone response, contains the resonance due to poor coupling. An example of noise added from this resonant signal and the absence of resonance after processing can be seen in Figure 5.2. There are sophisticated methods for overcoming resonance due to poor geophone coupling in downhole monitoring; however, these methods, while powerful, are still limited by monitoring geometries (Zhang et al., 2017a).

5.4 Machine Learning Model Selection

While there are numerous machine learning techniques that are widely available, few are specifically designed for imputation tasks. To determine the best approach, we follow the workflow described in Figure 5.6. Four methods are considered in this study; however, we omit specific details regarding Long Short-Term Memory network (LSTM) since the results are similar to that of random forest. Results are included for completeness.

Data Preparation

In order to establish ground truth for examining performance of the imputation methods, it is necessary to first understand the distribution of missing values, Figure 5.3, and then remove them. Next, we systematically remove known values from the remaining data set and consider these in later steps for testing. Here, it is critical to ensure that the synthetic missing values have a proportional distribution with respect to the real missing values. This

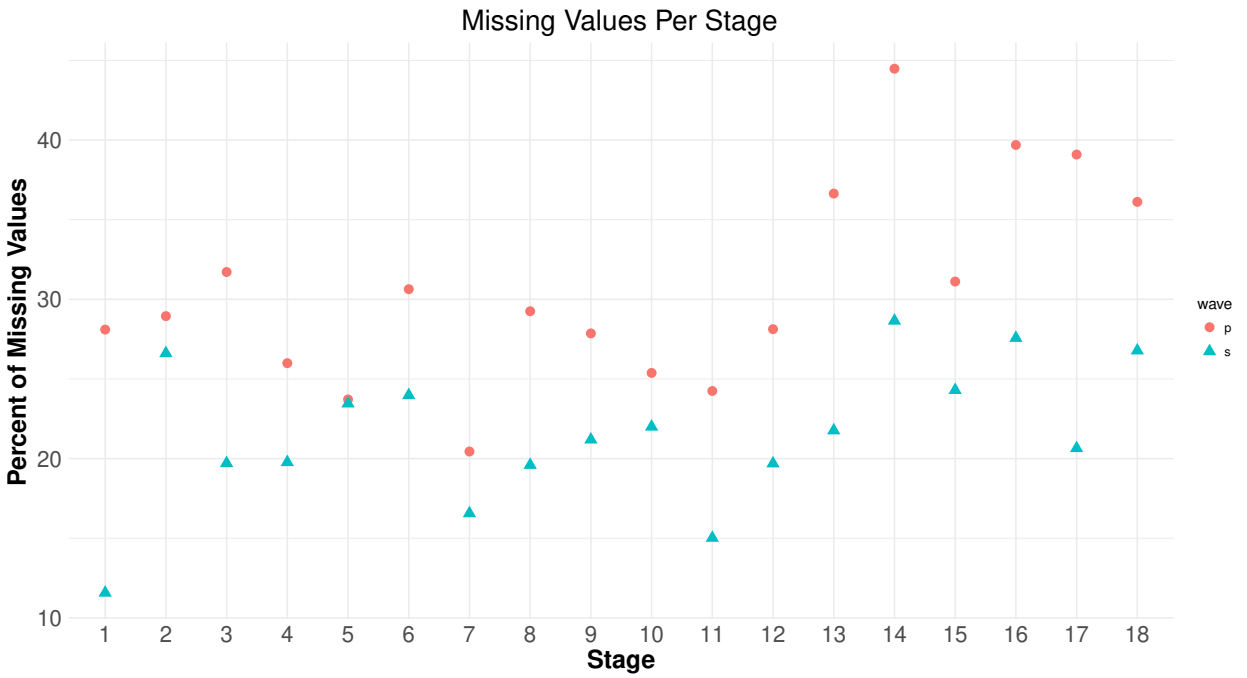

Figure 5.3: Percent of missing values for P and S wave amplitudes. Note that for every stage, there are more missing values of P wave amplitudes than S wave amplitudes. This leads to the likely conclusions that these values are not captured due to the relatively lower amplitudes of that wave type.

newly created data set with artificially missing values is used to train and test the selected learning methods. This approach is necessary due to the nature of the problem, since it is impossible to reliably assess performance on missing data and can be seen in Figure 5.4.

Stage-Specific Median Imputation

One of the most basic techniques used to impute missing information is to simply choose a value statistically derived from the whole data set that aims to minimize bias (Haukoos and Newgard, 2007). Typically, the mean, median, or mode is used to replace missing values in a given data set. However, due to the non-stationarity of the complete data set, it is necessary to segment values into homoscedastic subsets with stable statistical measures over time to minimize error. Given that each stage occurs in a different location, which leads to changing source-receiver pathways, it is important to consider each stage as a new data set with new real conditions that may lead to missing data. It can be seen in Figure 5.5 that median values vary significantly between each stage of the hydraulic fracturing project. As such, each stage was considered for the determination of median values for imputation.

Programmatically, this approach is straightforward, requires little computational overhead, and executes quickly. A disadvantage to this approach is that, though reduced dra-

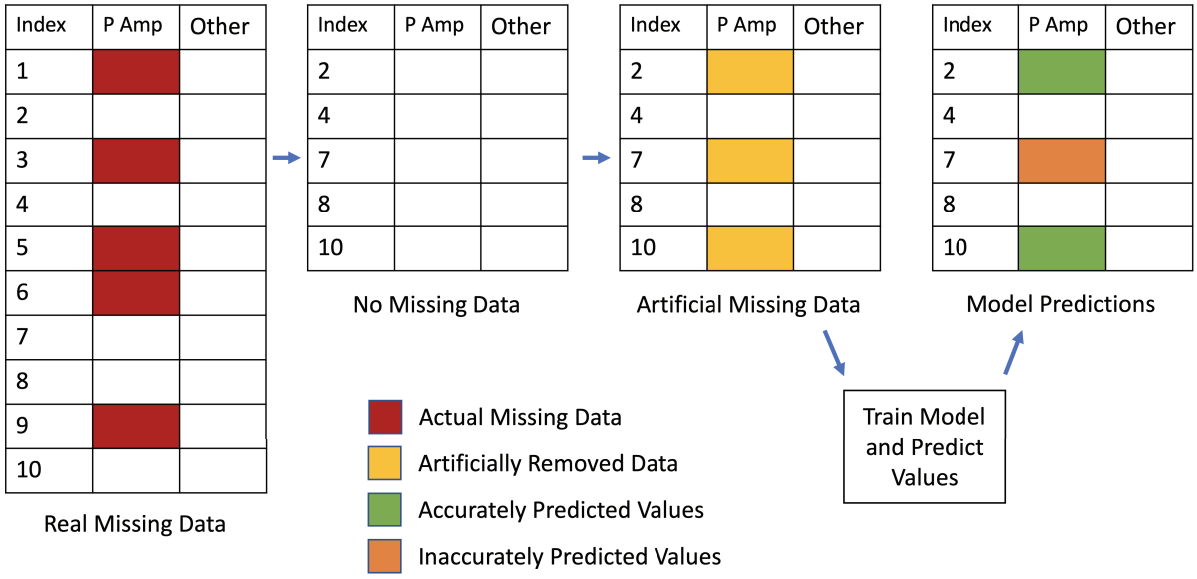

Figure 5.4: Visual description of process for developing synthetic missing data.

matically from the full data set, there is still variability over each stage. As a result, it is necessary to complete an entire stage before computing the median value to be used for imputation. Thus, despite extremely fast computation time, this approach is limited in its ability to be used for real-time decision tasks like early warning or identification of fault reactivation.

Random Forest Imputation

The random forest ensemble learning method is built from a number of decision tree predictors that depend on randomly sampled vectors that are both independent and identically distributed for all trees in the ensemble (Breiman, 2001). It is capable of performing classification, regression, and survival analysis, though in this chapter, the focus is purely regression. In order to optimize performance, an iterative approach was used to tune random forest hyperparameters. One of the advantages of random forest modeling techniques is its resilience in the face of over-fitting. However, there are still key parameters that can improve the overall performance. One important parameter, m_{try}, denotes the number of randomly selected candidate variables that are considered at each split when growing a specific tree (Breiman, 2001). As such, m_{try} was varied in order to determine the best performing model for overall comparison.

The random forest algorithm typically follows the following steps:

1. Perform bootstrap sampling on the original data set.

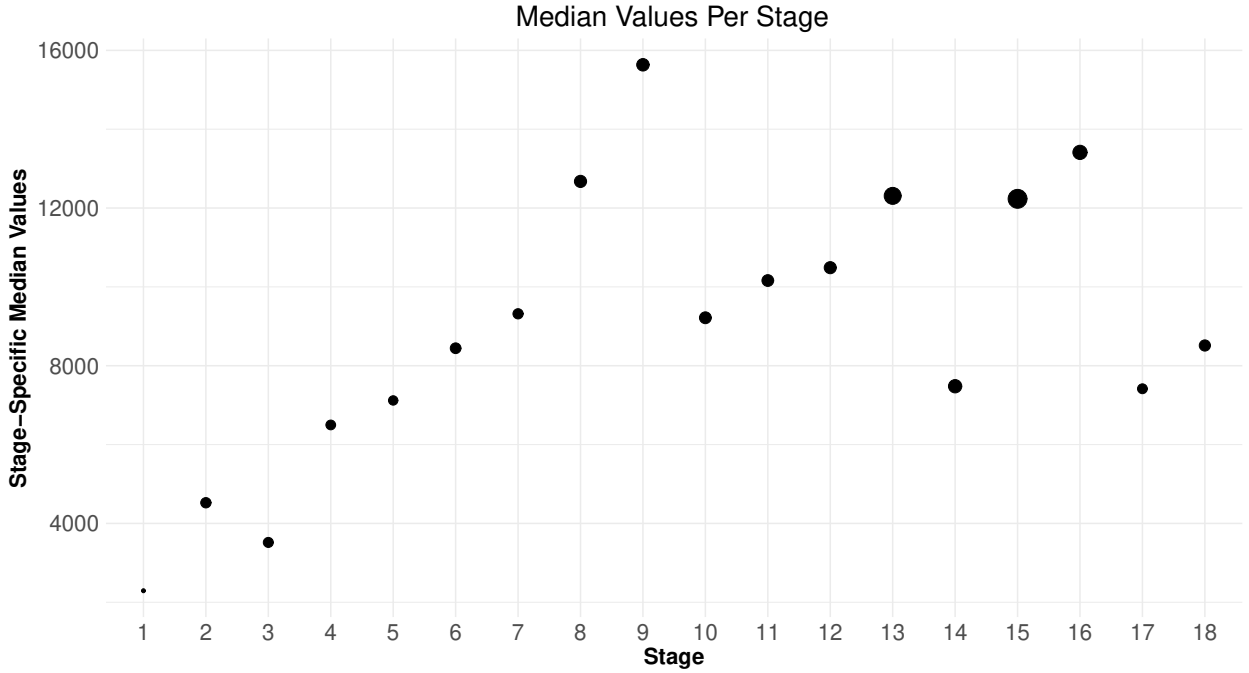

Figure 5.5: Median values per stage, size represents the standard deviation of values per stage.

2. For each sample, a decision tree is grown. Each decision tree is slightly modified where at each node, m_{try} number of predictors are randomly sampled and the best split from those variables is chosen.

3. Aggregation of predictions from all the decision trees leads to new data predictions through majority voting.

4. Performance metrics are calculated and error relates are obtained by aggregating out-of-bag (OOB) predictions.

Multivariate Imputation via Chained Equations (MICE)

The current state-of-the-art approach for imputation of complex data sets relies on a time-consuming implementation of Fully Conditional Specification (FCS), also known as Multivariate Imputation by Chained Equations (MICE). There are many advantages to this technique; however, one of the most valuable attributes of MICE is its ability to perform imputation tasks on both numerical and categorical data with high accuracy when there are missing values in more than one variable of interest. This is accomplished by the use of a modular approach that enables comparison of imputed values at each iteration.

Three main phases of the algorithm are *imputation*, *analysis*, and *pooling*. In the *imputation* phase, a user-configurable number of data sets are generated in parallel and each of

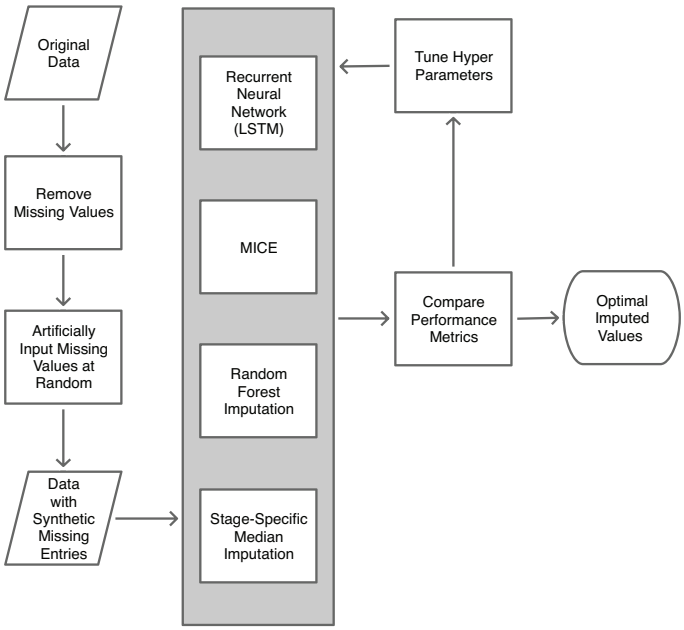

Figure 5.6: Overall workflow for calculating performance metrics for selected imputation methods.

these has different imputed values. It is important to note that although the imputed values differ, the non-missing values remain the same. The next step is the *analysis* phase where a model of imputed values is created for each version of the data set. Finally, in the *pooling* phase, the estimates from all the data sets are pooled and variance is estimated. The result of this approach should generate a complete data set that preserves the relationships present in the data as well as the associated uncertainty (Buuren and Groothuis-Oudshoorn, 2010).

5.5 Model Output and Performance

Scale-dependent measures are particularly useful when comparing different learning methods applied to the same data (Hyndman and Koehler, 2006). Another advantage of scale-dependent measures is that they are typically more interpretable due to the fact that the units of the measure and the units of the predicted variable are equivalent. Three measures used for understanding model performance include Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

The basic premise for understanding model performance begins with quantifying error. Specifically, model error (e_t) is defined as the difference between actual (A_t) and predicted values (P_t) for an observation at time t as seen in (5.2).

$$e_t = A_t - P_t \tag{5.2}$$

Mean Absolute Error is a straightforward approach to quantify model error and is defined by (5.3). MSE (5.4) and RMSE (5.5) are included to better quantify model accuracy.

$$MAE = mean(|e_t|) \tag{5.3}$$

$$MSE = mean(e_t^2) \tag{5.4}$$

$$RMSE = \sqrt{MSE} \tag{5.5}$$

Table 5.1 shows summary statistics for the different wave amplitudes in stage 2 of the hydraulic fracturing project. Table 5.2 describes the measures that were applied to various models to determine overall performance. Additionally, Figure 5.7 shows a visual comparison between model performance based on MAE.

Table 5.1: Summary Statistics

Amplitude	Min	Q1	Median	Mean	Q3	Max
P	1594	3068	4408	9690	7569	134485
S	1734	4119	6562	12487	12114	132919
SH	1302	3431	5686	11202	10482	115685
SV	805	1963	3066	4807	5355	65456

Table 5.2 shows that imputation with the MICE package outperforms all other machine learning and deep learning methods. However, there is also the consideration of computation time and overhead. Random forest and LSTM have slightly greater error rates; however, both of these methods have near equal performance to one another and require significantly less computation time than MICE. Since missing data is a problem that effects all aspects of modeling, a decision should be made regarding the value of imputation before choosing MICE over random forest or LSTM. For example, if the intention is to perform offline analysis and modeling tasks, then the time required for the implementation of MICE is reasonable. If, however, the purpose of imputation is to help inform real-time decisions like the identification of fault reactivation in early warning systems deployed on-site, then random forest and LSTM are the better choice. Due to the fact that early warning systems rely on the ability to quickly identify potentially dangerous phenomena and help inform operational decisions, the time required to implement MICE is prohibitive (Bao and Eaton, 2016). The error rates produced from the implementation of random forest and LSTM are slightly greater than MICE; however, the information gained is an improvement on the traditional methods of rudimentary imputation or the removal of what could be informative data.

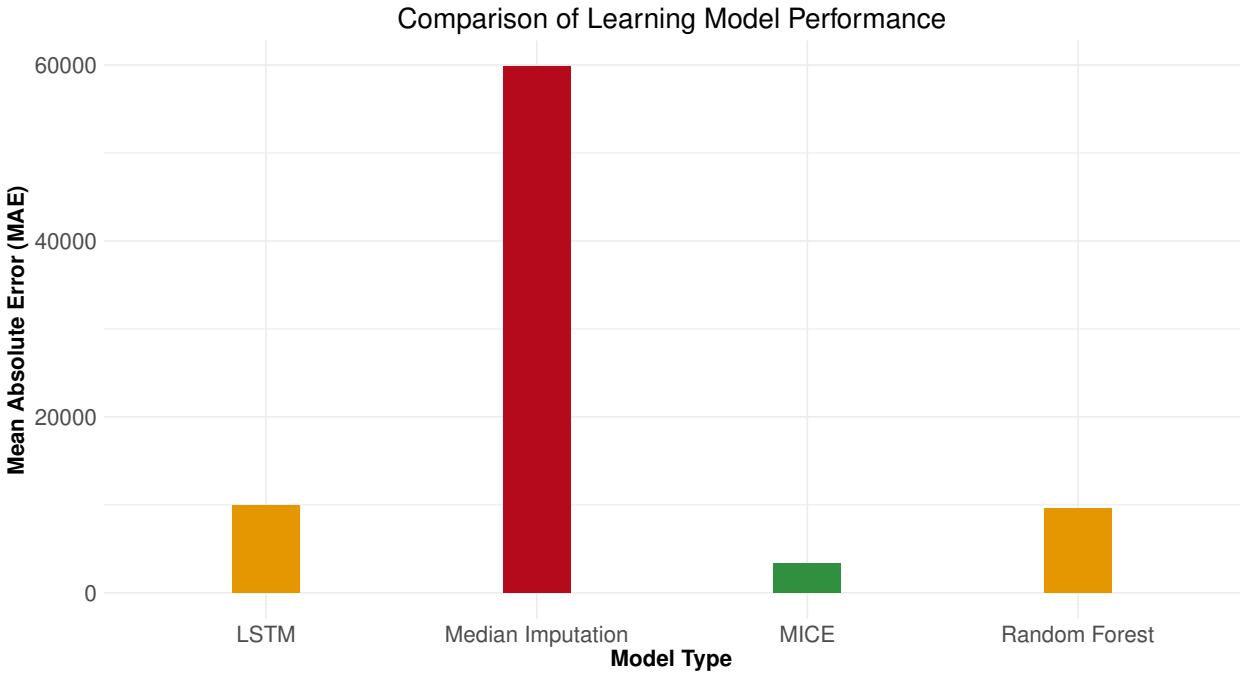

Figure 5.7: Comparison of model performance. Mean Absolute Error (MAE) is the measure of the average of the absolute difference between actual value and predicted value. An advantage of MAE, as well as other scale-dependent metrics is that they work well when comparing performance between different learning models on the same data set. Optimal learning methods minimize absolute error, then, MICE has the best performance.

Table 5.2: Model Performance

Imputation Method	MAE	MSE	RMSE
Median Imputation	59800	2.69e08	16400
Random Forest	9640	4.61e08	21500
MICE	3320	4.83e07	69502
LSTM	9940	4.74e08	21800

5.6 Conclusion

Any real data set will likely contain missing or unreliable information. Microseismic data recorded with downhole sensors from a hydraulic fracturing project in the Marcellus Shale is an example of a data set where a significant portion is missing or contaminated with noise. In an attempt to recover this missing information, a number of machine learning and deep learning methods were explored. As a first step, data cleaning and exploratory data analysis

tasks were performed to reveal that more than 30% of compressional wave amplitudes were missing from the original data set. The creation of synthetic missing values enabled the use of proven machine learning and deep learning techniques like random forest and long short-term memory networks to validate their use for imputation of missing data. Additionally, Multivariate Imputation by Chained Equations (MICE), the current state of the art package for imputation, as well as a stage-specific median imputation technique, were implemented. Standard performance metrics for regression methodologies were calculated for comparison of model performance. While MICE outperforms other learning techniques, the computational time is significantly higher for even a small subset of the data set. As such, the clear choice for imputation tasks depends heavily on the specific goal of the analysis objectives.

Chapter 6

Arrival Time Picking with Ensemble Methods

> *Because learning takes practice, we are more likely to get things right at small stakes than at large stakes. This means critics have to decide which argument they want to apply. If learning is crucial, then as the stakes go up, decision-making quality is likely to go down.*
>
> – Richard Thaler, *Misbehaving: The Making of Behavioral Economics*

6.1 Introduction

In this final chapter, we culminate our work with the application of data science methodologies applied to a time-intensive, manual analysis task. The objective is to offer relief from manual arrival time picking through a data-driven, extensible framework. The value of the work presented in this chapter, or an extension of this work, would help to improve the results of all previous chapters in this dissertation.

The identification of arrival times is a critical component of microseismic and seismic analysis that enables and informs subsequent analysis tasks like source location estimation, focal mechanism, and moment tensor inversion, as well as fracture network reconstruction (Álvarez et al., 2013; Galiana-Merino et al., 2008; Li and Dong, 2014; Xiantai et al., 2011; Yue et al., 2014). There are a number of methodologies that enable the identification of arrival times; however, each has accompanying limitations. For example, manual picking of arrival times requires a significant amount of time to manually identify waveforms and pick arrival times. A confounding effect of this manual work is that the picking accuracy is highly subjective and is influenced by and susceptible to human error. The automation of this time-intensive step is an area of focus in the geophysics community. As such, there are a number of methods that have been created to automate this picking process; however, we focus on Short-Time Average over Long-Time Average for a comparison of performance.

Short-Time Average over Long-Time Average (STA/LTA) is an approach that considers a short window that represents instantaneous energy and a long window that represents temporal amplitude of the seismic noise. When the ratio of these two windows of energy overcome a predefined threshold, the approach is able to identify an arrival. There are a number of parameters that can be used to improve picking accuracy; however, this approach does have limitations that should be considered when using it in various applications. For example, STA/LTA performs well in seismically quiet sites where the dominant source of noise is natural seismic noise. Additionally, it is useful in strong motion seismicity since the presence of a seismic event is typically represented by energy that is much greater than nominal background noise. Moreover, STA/LTA is not as effective in the presence of man-made seismicity (Li et al., 2016; Jones and van der Baan, 2015; Trnkoczy, 1999). As such, tuning of the STA/LTA trigger often requires a tradeoff between detection rate and false triggers. In a hydraulic fracturing project, there is typically a high level of background noise, overlapping microseismic events in time and space, noise due to resonance, and relatively lower amplitude events in general. These limitations are the motivation for leveraging machine learning methods to understand how time series classification can aid in the identification of compressional waves and improve automated arrival picking.

Microseismic arrival picking is an area of research that has long been dominated by traditional signal processing techniques (Capilla, 2006; Gibbons et al., 2012; Gibbons and Ringdal, 2006; Senkaya and Karsli, 2014). However, this chapter explores the application of machine learning methods to help improve arrival time picking accuracy. The arena of artificial intelligence, which includes machine learning and deep learning, is generally split into two main areas: classification of categorical data and forecasting or regression of time series data. This chapter seeks to bring attention to the benefits of applying sophisticated learning methods to time series data with the objective of classifying waveforms.

The method proposed in this chapter considers raw, real data from the Marcellus Shale and relies on a subset of the data to be processed in order to establish known arrival times. From here, the traces are segmented into signal partitions, or chunks, that act as the foundation of the time series classification approach. These chunks aid in identifying where the compressional wave exists within the input signal. Features are created and analyzed for overall importance. Final feature selection is performed based on statistical tests that help inform feature relevance. Next, the known arrival times are used to window the compressional wave, and this step informs the creation of the target variable, which is essential in any supervised learning technique. Lastly, the features and target variable are used to explore the performance of various machine learning and deep learning methods. Standard automated arrival picking methods are employed to understand overall performance gains in arrival time picking.

This chapter begins by presenting the monitoring geometry utilized in the microseismic monitoring project that took place in the Marcellus Shale whose real data are considered here. Next, the methodology is presented that was used to train and test the performance of various learning methods. Then, the final model specification is described with additional information regarding the use of bagging and boosting to improve overall classification perfor-

mance as well as a listing and description of the most important features that were considered for this modeling endeavor. Finally, we summarize the results of our approach and compare them to performance of STA/LTA on the same data and conclude with potentially valuable next steps.

6.2 Survey Geometry

Real data are considered from a hydraulic fracturing project that took place in the Marcellus Shale located in the Susquehanna River Basin in Susquehanna County, Pennsylvania. The Marcellus Shale lies on top of the Onondaga Formation, which is primarily limestones and dolostones. Above the Marcellus Shale is the Mahantango Formation, which is mostly siltstone and shale. The Marcellus Shale is one of the largest shale formations in the world and is thus one of the largest sources of natural gas in the United States. The average thickness of the Marcellus Shale in the local survey area is approximately 46 m with a porosity of 0.08 and permeability of 600 nD (Zhang et al., 2017a).

Although the multiple-well pad held seven parallel horizontal wells, the data considered in this investigation was recorded from one observation well with a lateral distance of approximately 220 m (722 ft) from the treatment well considered (Salehi et al., 2013). The overall project was significant in scale and execution. The average wellbore length in the horizontal direction was 1109 m (3640 ft). Additionally, the wells were located in the lower portion of the Marcellus Shale with a true vertical depth (TVD) of approximately 1981 m (6500 ft). The trajectories of the wells were in the direction normal to the maximum in situ horizontal stress orientation, as is common in horizontal well projects.

The scientific objective of the hydraulic fracturing project was to determine if a change in pump rate would lead to an appreciable change in stimulation efficiency. Namely, to establish a link between rapidly changing pump rates and increased production, reduced water consumption per unit of gas produced, and an overall reduction in environmental impact. In an effort to understand these characteristics, both surface monitoring and downhole monitoring tools were employed. The surface monitoring array spanned an area of approximately 7.8 km^2 (3 mi^2) and monitored 93 stimulation stages while the downhole monitoring array was placed in a single horizontal well and monitored 62 stimulation stages. The outcome of this study was considered successful as an increase in microseismicity was observed with frequent pump rate changes (Ciezobka et al., 2016).

While surface monitoring tools were deployed, this investigation focuses solely on downhole monitoring data. Two horizontal wells are considered, a treatment well and an observation well with horizontal lengths of 1700 m (5577 ft) and 1350 (4430 ft), respectively. There were 18 stimulation stages beginning at the toe and ending at the heel of the treatment well with an interval of 91 m (300 ft), Figure 6.1. Nine of the stimulation stages were executed with constant pump rate, and the remaining nine stages employed a variable pump rate.

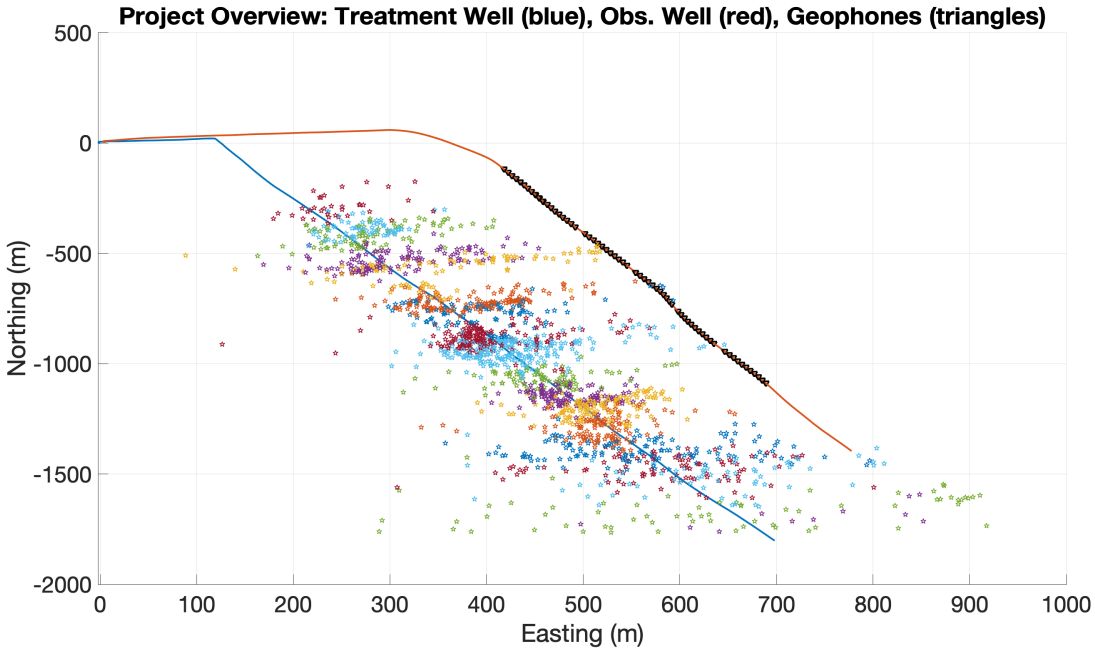

Figure 6.1: Map view of hydraulic fracturing geometry showing fracture stages and geophone locations. Inverted triangles show the different locations of the geophone array in the observation well, shown in red. The locations of microseismic events are shown around the blue treatment well color-coded for each stage.

Downhole monitoring was conducted with an array of three component, 10 Hz geophones. There were 11 geophones in this array with spacing of approximately 11.2 m (50 ft). As is common in downhole monitoring that requires relocation of sensors, the array was towed via tractor for each relocation. Relocation of monitoring sensor arrays is a common practice in downhole monitoring that helps to improve the Signal-to-Noise Ratio (S/N) by minimizing the source-receiver distance.

There were 1842 microseismic events that were identified through standard processing techniques. These events all satisfied some threshold for overall quality; however, there are a number of events that have high noise content, demonstrate the presence of resonant noise in the form of tube waves, or are missing compressional wave arrivals (Nava et al., 2020b).

6.3 Methodology

The method proposed in this chapter follows the workflow in Figure 6.2. Broadly, execution of this method requires the following steps:

1. Manually pick a subset of events
2. Establish initial arrival window and signal partition (chunk) schema

3. Create relevant features for signal chunks
4. Train classification model
5. Implement automatic first arrival technique
6. Tune dynamic parameters and select final model

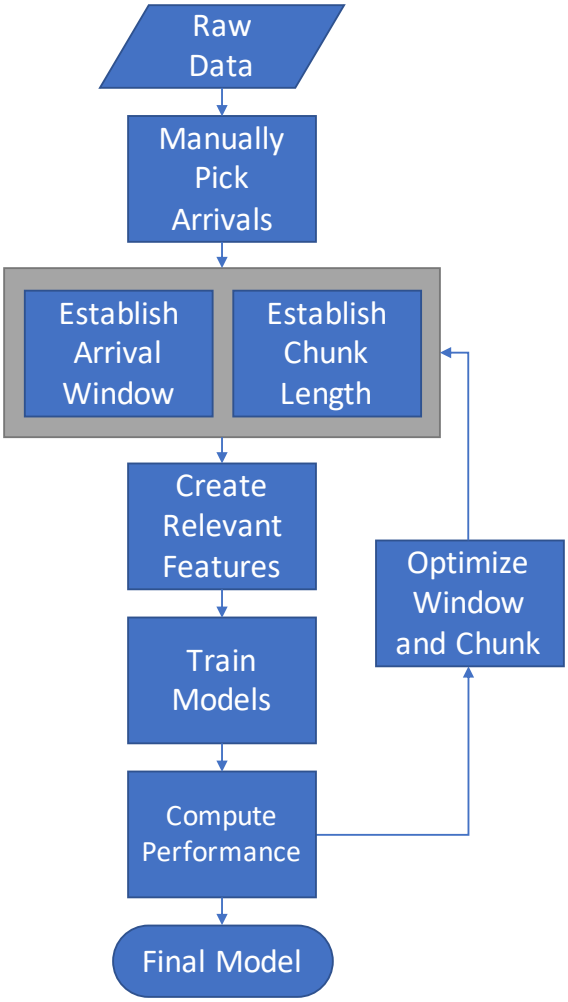

Figure 6.2: The overall workflow of this modeling endeavor begins with raw data. Minimal preprocessing is required with this approach since only a small subset of the events are picked. Manual picking can be employed or automatic picking methods can be utilized if strict quality control steps are taken to ensure the picks are accurate. Next, initial arrival window and chunk lengths are chosen. From here, relevant features are selected and models are trained. An iterative approach is used that incorporates performance of the overall metrics with a feedback loop that varies arrival window and chunk length until optimal performance is achieved.

Preprocessing

Like any other machine learning endeavor, this classification approach requires some amount of preprocessing of input data. Here, the initial effort is establishing ground truth by picking arrival times of a subset of events. This can either be done manually through a subject matter expert physically interpreting some small number of events or through an automatic picking algorithm. It is important to note, however, that it is critical to verify the quality of automatic picks in order to ensure that the subsequent steps do not lead to inaccurate results.

For the purposes of this chapter, we use contractor-provided estimates of arrival times for events with high S/N. The decision to use contractor-provided arrival times was made in order to evaluate the approach's ability to improve on industry standard microseismic processing techniques. Due to the fact that the data set under consideration contained a large number of events that were effected by interference from ringing artifacts, an example of which can be seen in Figure 6.3, the events were sorted by S/N and only events that demonstrated low noise were chosen for model training (Nava et al., 2015). It is important to identify microseismic events with low noise content, specifically, low noise content surrounding the compressional wave to allow the model to accurately identify attributes that are representative of the true signal rather than noise content. If this step were omitted, the predictive power of the model would rapidly diminish due to poor quality training data.

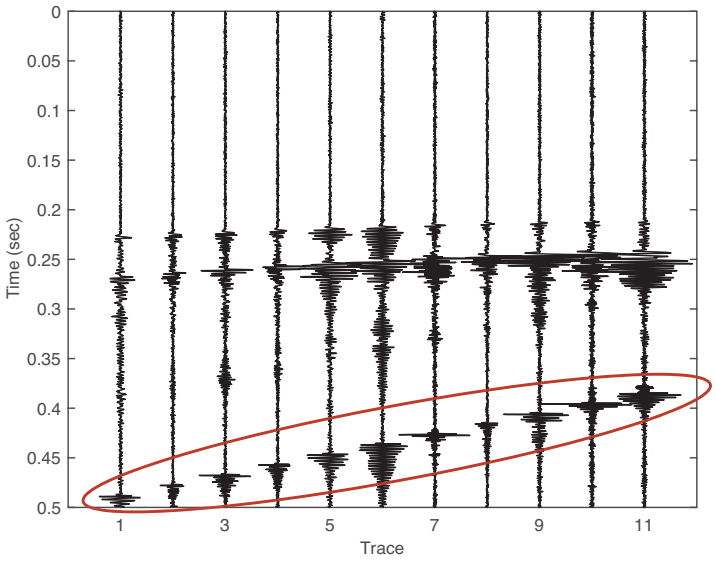

Figure 6.3: Real microseismic event recorded with eleven geophones that shows ringing artifact. Ringing due to resonant tube wave energy propagating down the borehole adds noise throughout the hydraulic fracturing process and is likely caused through insufficient clamping force between geophone and borehole casing. The highlighted artifact is completely removed through traditional processing and is considered noise.

Feature Engineering and Selection

After selecting appropriate events and establishing first arrival picks with high confidence, a set of features that can be used for classification must be created. However, before this step can be effectively executed, it is necessary to window around the first arrival times in order to capture the full compressional wave. Additionally, an appropriate signal partition length, or chunk length, must be chosen. These two steps can be viewed as independent in the sense that the arrival window is bound by the physical properties of the compressional wave, whereas the chunk length is bound by the desired granularity of the estimation approach.

Windowing compressional wave arrivals

There are strict constraints on the arrival window that must be considered in order to minimize error in subsequent estimates. First, if the window length is too small, then there will not be clean distinction between what should be considered a compressional wave and what is simply noise or a non-event signal. Failure to appropriately window the target phenomena in the input signal will likely result in degradation of predictive power. Conversely, an arrival window that is too long will likely carry with it the negative effect of capturing both compressional wave and shear wave attributes (Figure 6.4). This will also result in an inability to effectively classify a compressional wave and distinguish it from the shear wave. Due to the fact that the shear wave energy is typically much greater than the compressional wave energy, this would result in classifying a shear wave arrival rather than a compressional wave arrival (Aki and Richards, 2002). Thus, the error between actual compressional wave arrival times and predicted arrival times would be quite large.

Partitioning signal

Chunk length is the number of samples that are considered when generating aggregate features as part of the feature engineering phase of model development. In order to perform time series classification, relevant features are created for a time series under consideration. Because the intent of this approach is to identify a specific attribute within a time series, it is necessary to appropriately determine the optimal level of granularity. Unlike the arrival window, the chunk length is not bound by a physical constraint. Here, the main constraint in choosing chunk length is the model performance. If the chunk length is prohibitively small, then execution time increases and overall classification accuracy decreases. Conversely, if the chunk length is too large, overall execution time will be much shorter and classification accuracy may be improved; however, this is a misleading result. The objective of this approach is to classify a compressional wave within a given trace with high accuracy. If the chunk length is too large, the overall value of the approach is low due to the fact that there is still high uncertainty regarding the location of the compressional wave arrival. Then, the next step will show only marginal improvement when compared to traditional automated picking techniques. As such, the chunk length should be tuned such that the overall classification

accuracy is high and the chunk length is also small enough to give a sense of where the compressional wave exists within the complete trace.

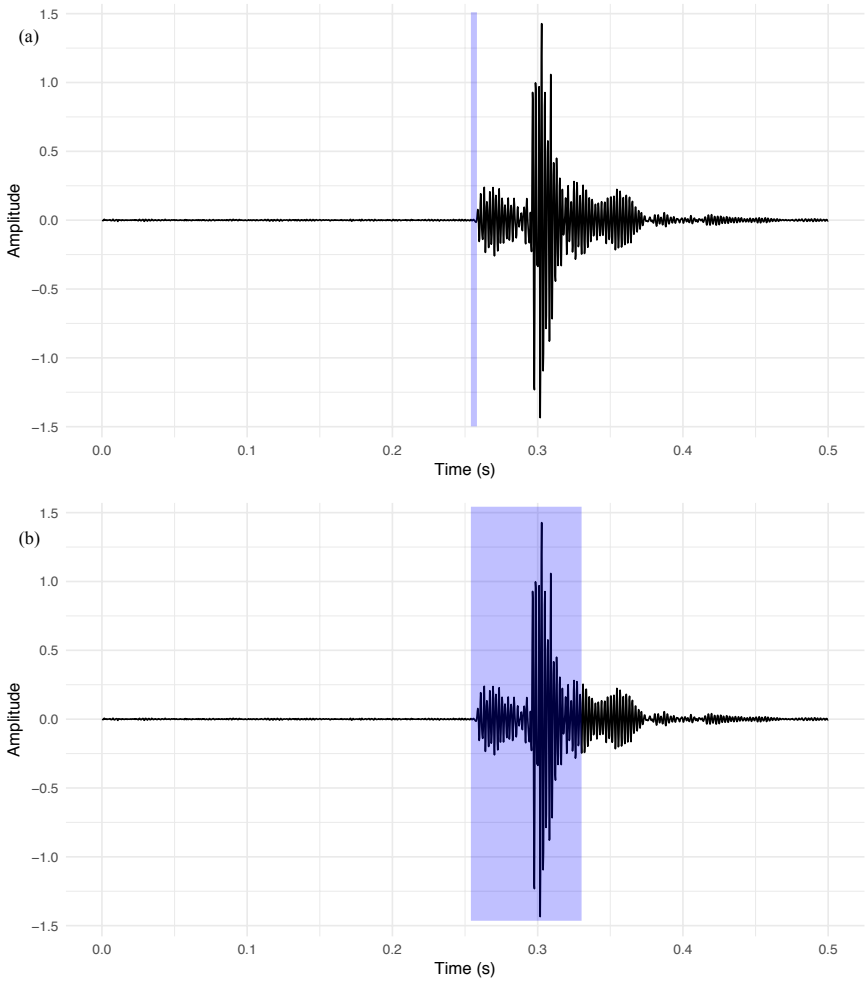

Figure 6.4: Example of non-optimal compressional wave arrival windows. The first window (a) shows an arrival window that is too small and demonstrates an inability to capture compressional wave attributes. The second window (b) shows an arrival window length that is too large that captures compressional wave as well as shear wave energy. Both of these arrival windows lead to sub-optimal predictive performance. Window (a) leads to a significantly higher rate of false negatives and window (b) leads to the model misclassifying the shear wave as the first arrival.

Creating relevant features

Traditionally, feature extraction and engineering is a time-consuming step in the model building process. Recently, there has been significant effort in automating this step (Katz et al., 2016; Severyn and Moschitti, 2013). More specifically, there have been a number of

packages created with the explicit purpose of engineering features from time series data for classification tasks (Cabrera et al., 2017; Mierswa and Morik, 2005; Naul et al., 2016). Due to its fast execution speeds through built-in parallelization, rigorous feature selection and filtering process, and clean integration through standard APIs, we used the Python package Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh).

Tsfresh is quite robust and, by default, uses 63 characterization methods which produce 794 time series features (Christ et al., 2018). It is important to note that not all the features created are relevant to the classification objective. As a result, one of the core components of tsfresh is the identification of *relevant* features through statistical tests. Specifically, each feature is evaluated independently to understand its ability to accurately predict the target variable. In this case, the target is a binary classification of whether or not the compressional wave arrival window is observed. The result of this step is a vector of p-values that represent the significance of each extracted feature. Finally, the computed p-values are used in the Benjamini-Yekutieli procedure, which determines which features are relevant (Benjamini et al., 2001).

Model Training

Through an iterative process of evaluating classification performance, the final model is determined. We examined the classification performance of a number of modeling approaches. Table 6.1 describes the models and overall prediction accuracy. It is important to note that in cases where there are imbalanced classes, overall accuracy must be compared to the No Information Rate (NIR). In the case of a binary classification problem, the NIR is the proportion of the majority class (Kuhn et al., 2008). The NIR associated with this model is 85.4%, and thus any model that provides an accuracy less than the NIR actually performs very poorly.

Table 6.1: Comparison of Classification Performance

Model	Accuracy (%)	Time (min)
XGBoost	94.9	10
Bagged AdaBoost	94.7	14
Random Forest	94.3	7
AdaBoost	93.8	18
Stochastic Gradient Boosting	93.2	1
Bagged CART	88	1
Naive Bayes	87	1
Stacked AutoEncoder Deep NNet	71	3
Self-Organizing Maps	71	5
Linear Discriminant Analysis	68	1
Neural Network	62	1

Group K-Fold Cross-Validation

An important aspect of model development is maximizing predictive power while avoiding the negative effect of overfitting. One widely accepted method to accomplish this is to incorporate cross-validation in the training phase of development (Hastie et al., 2009). There are a number of cross-validation techniques that can be applied to categorical and time series data. Common methods utilized for classification tasks are k-fold, v-fold, and stratified k-fold cross-validation. Additionally, there are cross-validation methodologies that are typically employed for forecasting tasks that rely on time series data. For example, forecast evaluation with rolling origin has been used as a means of improving forecasting performance for some time (Fildes, 1992). Furthermore, nested cross-validation enables cross-validation techniques to be applied to time series data in an effort to preserve temporal associations (Bergmeir and Benítez, 2012; Tashman, 2000; Varma and Simon, 2006). These methods significantly improve model performance and lead to more robust classification and forecasting implementations of machine learning applications. However, these methodologies fail to appropriately handle the task of time series classification as it is used in this chapter.

In this chapter, we attempt to identify where a compressional wave exists within a microseismic trace. Although the input data are time series in nature, this approach is not a forecasting methodology. As such, cross-validation techniques that are applied to forecasting tasks are not applicable. Furthermore, typical cross-validation techniques that are used in more traditional classification endeavors are not applicable here due to the very real risk of data leakage. Data leakage occurs when related data exists in both the training and testing sets used for model development (Kaufman et al., 2012; Kuhn et al., 2008; Nisbet et al., 2009). Standard k-fold cross-validation randomly samples observations and separates the input data into two separate data sets - training and testing/validation sets. If each observation were independent, this would be appropriate; however, since the input data are time series in nature, randomly sampling from all traces would effectively downsample and split each signal between the testing and training sets (Figure 6.5). This would inevitably lead to high performance on training data, but incredibly low performance on unseen data. This behavior is commonly referred to as overfitting.

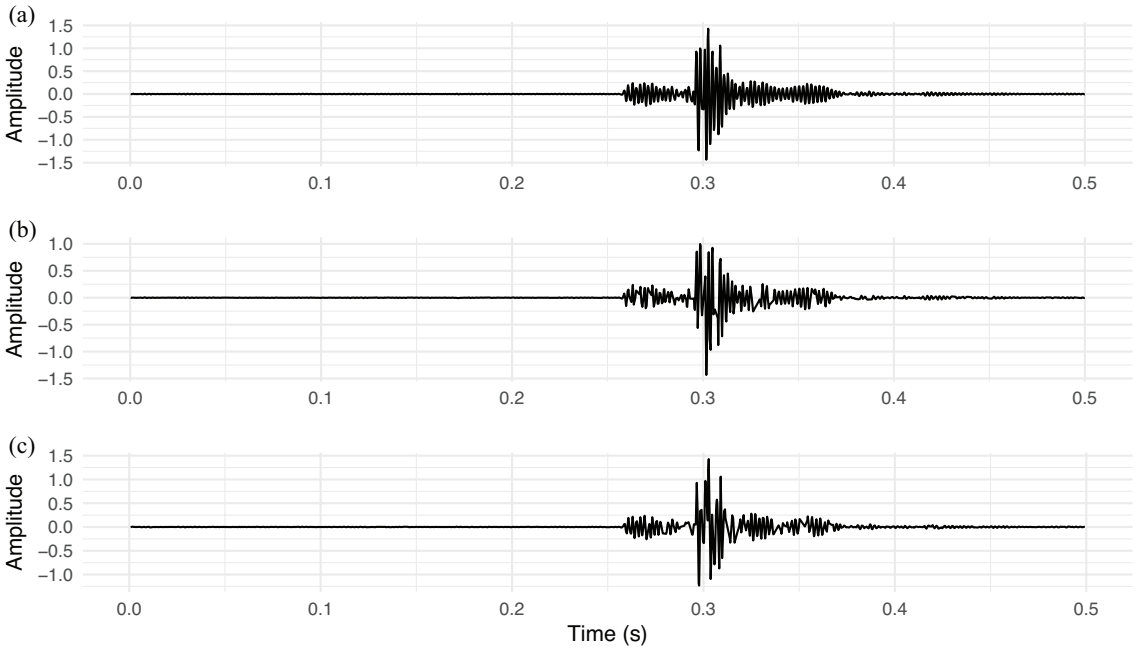

Figure 6.5: Data leakage occurs when information from the same observation is present in both the training and testing data sets. The negative effect is an overly optimistic sense of model performance and the subsequent inability to handle new data. This is commonly known as overfitting. The microseismic event in (a) is the full signal. The two signals in (b) and (c) represent the effect of traditional cross-validation techniques when applied to time series data. The overall effect is random downsampling of the signal. While the signals are not exactly the same, the arrival times remain unchanged and will lead to overfitting.

In order to avoid overfitting while maximizing predictive performance, a more rigorous cross validation schema is required. Below is a general approach to implementing our cross-validation method.

Time Series Group K-Fold Cross-Validation Method:

1. Create list of Event IDs and randomly sample
    a) 67% are assigned to training set
    b) 33% are assigned to test set
2. Retain test set for validation purposes
3. Implement Group K-Fold schema based on Trace ID
    a) This indexes the folds based on Trace ID
    b) Each trace is then viewed as an independent observation

Through the application of group k-fold cross-validation where each event is treated as an independent observation, the disadvantage of data leakage is effectively overcome by preventing features from a single event from being present in both test and train sets. The temporal associations that are present within time series data are preserved and model performance and robustness are improved.

Final Model Specification

Based on the nature of the data from the Marcellus Shale hydraulic fracturing project, the model that had the best classification performance was XGBoost, which is a tree-based ensemble method that incorporates optimized gradient boosting (Chen and Guestrin, 2016).

Bagged AdaBoost, which is a tree-based ensemble method that incorporates bagging from the R package adabag also performed very well (Alfaro et al., 2013). It is important to note that there was only a very small decrease in performance between Bagged AdaBoost, AdaBoost, and Random Forest. This indicates that this problem is well-suited for both adaptive boosting and bagging methods. Furthermore, the top six machine learning methods tested all relied on either bagging or boosting as a core component of implementation. While the poorest performing models have dramatically faster execution speeds, the fact that their accuracies are lower than the NIR indicates that the models actually perform worse that naive estimation.

Bagging

Boosting and bagging are both ensemble methods that generate base classifiers that are both relatively precise and as different as possible (Alfaro et al., 2013). While it is typically disadvantageous to incorporate features that have high variance, with boosting and bagging, this high variance leads to performance gains. In a single decision tree, if the training set contains samples at random, which is common in typical cross-validation approaches, even small differences in the selection of training observations will likely have a significant impact on the classification accuracy. Because of this attribute, an abundance of caution should be exercised when attempting to understand feature importance since this can change significantly with slightly different input data. Boosting and bagging leverage this phenomenon to increase performance while protecting against overfitting. Breiman (1996) comments, "The vital element is the instability of the prediction method. If perturbing the learning set can cause significant changes in the predictor constructed, then bagging can improve accuracy."

Bagging, or more specifically, bootstrap aggregating, is a technique that generates multiple versions of a predictor through random sampling with replacement (bootstrap) and then these predictors are used to determine an aggregate predictor (Breiman, 1996; Quinlan et al., 1996). Base classifiers are created on the bootstrap observations and then voting occurs to determine the final classifier. Specifically, for a given set of N observations, each of which belongs to one of K classes. Here, K can either describe a binary classification task where $K = 2$ or a multi class classification task where $K > 2$. The number of trials, T, indicates

the number of repetitions and may be either a static parameter or determined through the specific cross-validation method employed in model development (Quinlan et al., 1996) . Then, for trial $t = 1, 2, \ldots, T$, bootstrap sampling is performed on the set of N observations. Here, bootstrap sampling describes sampling with replacement independently from the original set. A key attribute of sampling with replacement is that although N remains the same for each trial, each observation may appear multiple times or not at all in any particular replicate set. This attribute is a fundamental aspect that enables the creation of different training sets and its value relies on the use of classification methods that yield highly varying outcomes when the input data are perturbed. Breiman (1996) notes that bagging has the potential to degrade the performance of stable procedures or on highly invariant data sets. As such, this indicates that the data used in this chapter are well-suited for ensemble methods like bagging and boosting. Classifiers for each trial, C_t, are created and then majority voting is used to form $C*$, which is the aggregate, or bagged, classifier that likely leads to performance gains.

Boosting

Bagging and boosting both create a number of base classifiers; however, where bagging creates base classifiers through independent resampling of the original data, boosting maintains a distribution or set of weights over the training set (Freund et al., 1999). An initial weight $w_t(i), i = 1, 2, \ldots, n$ where i is the training observation is created and updated on each successive iteration through T. Then, for $t = 1, 2, \ldots, T$, a classifier is fit using weights $w_t(i)$ on the resampled data set. Next, the error of the classifier e_t is calculated:

$$e_t = \sum_{i=1}^{n} w_t(i)\mathbf{I}(C_t(\mathbf{x}_i) \neq y_i) \tag{6.1}$$

where $\mathbf{I}(\cdot)$ represents an indicator function that generates a 1 if true, and 0 otherwise. Additionally, a constant α_t is calculated:

$$\alpha_t = 1/2 \ln((1 - e_t)/e_t) \tag{6.2}$$

In the next trial, the weight is updated to:

$$w_{t+1}(i) = w_t(i) \exp(\alpha_t \mathbf{I}(C_t(\mathbf{x}_i) \neq y_i)) \tag{6.3}$$

Next, the weights are normalized to sum to 1. Here, the weights associated with misclassified observations are increased and the weights of the correct classified observations are decreased. This results in a classifier in the following iteration that is more focused on the more difficult to classify observations. Then, a final classifier is produced that can be represented by:

$$C_f(\mathbf{x}_i) = \arg\max_{j \in Y} \sum_{t=1}^{T} \alpha_t \mathbf{I}(C_t(\mathbf{x}_i) = j) \tag{6.4}$$

where Y is the set of outcome classes from the classifier (binary classification would result in $Y = 2$). T is the number of trials, and α_t is the constant (6.2).

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a scalable and efficient implementation of the gradient boosting structure proposed by (Friedman et al., 2000). XGBoost relies on a regularized learning objective, can incorporate shrinkage, column subsampling, or dropout to improve overall performance while avoiding overspecialization, and is easily parallelizable. We will present a brief derivation of the regularized learning objective and gradient tree boosting; however, a full accounting can be found in (Chen and Guestrin, 2016) and (Friedman et al., 2000).

A tree ensemble method incorporates K additive functions to predict an outcome.

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{6.5}$$

where $\mathcal{F} = \{f(x) = w_q(x)\}(q : \mathbb{R}^m \mapsto T, w \in \mathbb{R}^T)$ describes the regression tree space. q represents the structure of each tree that maps an observation to the corresponding leaf. T represents the number of leaves and w represents the weights. Each f_k is an independent tree structure. The regularized objective is then:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{6.6}$$

where $\Omega(f) = \gamma T + 1/2\lambda\|w\|^2$ penalizes the complexity of the method. l is a differentiable convex loss function that considers the prediction $\hat{y}_i$ and actual value y_i. It is important to note that XGBoost can use a number of loss functions that may be appropriate for the modeling objective at hand. Chen and Guestrin (2016) further describe the discretization of the regularized objective function to enable its use in Euclidian space.

Another aspect that is incorporated in order to minimize the risk of overfitting is the use of shrinkage (Friedman, 2002). More specifically, the main objective of shrinkage is to prevent over-specialization, which occurs when trees trained earlier in the modeling process have a significantly greater impact on the outcome than later trees. Over-specialization is typically encountered in boosting approaches due to the fact that boosting, unlike bagging, sequentially adds predictors that seek to improve performance. Since subsequent iterations in boosting algorithms are trained on smaller subsets of the data, the overall impact of later predictors decreases. Here, the negative effect of over-specialization is overfitting based on early iterations in the training process. The introduction of shrinkage aims to apply a scaling factor to each tree in a given iteration so that subsequent iterations continue to have a significant impact on predictors.

DART booster

An alternative method that XGBoost can utilize to prevent overfitting via over-specialization is the DART booster (Rashmi and Gilad-Bachrach, 2015; Friedman et al., 2000; Friedman, 2002). Dropouts meet Multiple Additive Regression Trees (DART) incorporates the act of dropping trees in order to combat over-specialization. Shrinkage is useful in a number of modeling endeavors; however, it has been shown that as the size of the ensemble increases, the negative effect of over-specialization tends to return even with shrinkage. As such, in certain cases, DART leads to superior performance. This is the case with the data set considered in this chapter. In a sense, the DART booster creates a version of XGBoost that is more similar to a bagging algorithm like random forests (Breiman, 2001).

DART relies on a parameter that controls the dropout rate between iterations and if this parameter is minimized, no trees are dropped and the boosting algorithm executes normally. However, if the parameter is maximized, then a boosting algorithm like XGBoost more closely resembles random forests. This change occurs because random forests will only consider a random subset of the ensemble at each step. Where DART diverges from a typical random forest approach is that it also performs a normalization step on new trees in order to prevent overtraining. The net result of using DART in combination with XGBoost is that overall classification performance is increased over traditional XGBoost.

Variable importance

Given the fact that the original signal is not directly used in model training, only the features created from the signal are, it is important to understand the relative importance for the features directly. As such, the variable importance is considered here. Figure 6.6 shows the top 10 features in descending importance. It can be seen that there is a variety of feature types that have a high impact on the performance. *Energy ratio by chunks* is shown to be the most important feature. This calculates the sum of squares of a given chunk and is represented as a ratio with the sum of squares over the whole time series. Next, the *sample entropy* represents a measure of complexity by comparing the conditional probability that a short epoch, or template, is repeated during the time series (Richman and Moorman, 2000; Richman et al., 2004). Specifically,

$$SampEn = -\log(\frac{\sum A_i}{\sum B_i}) = -\log(\frac{A}{B}) \tag{6.7}$$

where A_i represents the matches of length $m+1$ with the ith template, and B_i represents the matches of length m with the ith template. Then, it can be seen that this feature is important because it likely aids in the identification of compressional wave energy, or the change in average amplitude, within a given chunk. Similar to energy ratio by chunks, sample entropy enables the model to distinguish between a compressional waveform and general seismic noise.

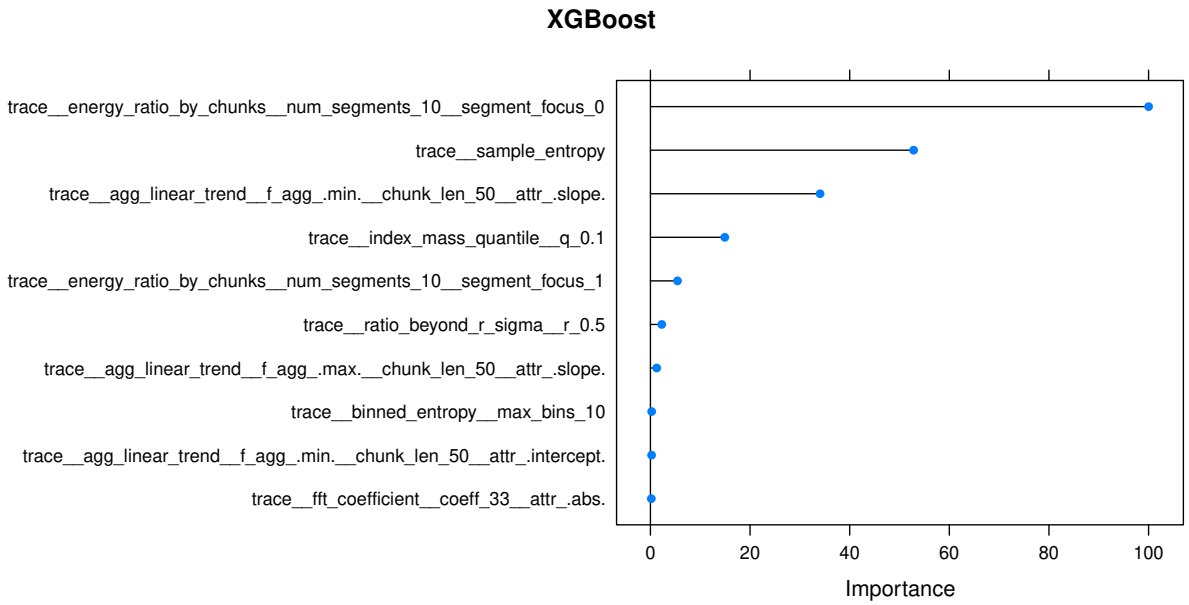

Figure 6.6: Feature, or variable, importance is critical in understanding the impact of model inputs. The top 10 features used in XGBoost are shown. Note that the most important feature incorporates a ratio of energy between a subset of the chunk and the total chunk. Next, the sample entropy indicates overall complexity of the chunk, which likely enables the model to differentiate between a chunk with an arrival versus a chunk that contains pure seismic noise.

Next, the *aggregate linear trend* calculates a linear least-squares regression for values of the input signal that were aggregated over chunks under the assumption that the signal is uniformly sampled. It can consider a number of attributes, for example, p value, r value, slope, standard error and so on. Additionally, the aggregation function can be minimum, maximum, mean or median. Further, the *index mass quantile* represents the relative index where some percent of the mass of the time series resides to the left. Ratio beyond r sigma considers the ratio values that are more than $r * sd(x)$ away from the mean of x, where x is the time series in a given chunk. These features all offer some understanding of local signal information with respect to the full time series, which is intuitively useful given that the objective of this approach is to identify where a specific phenomenon, the compressional wave, exists along the full signal. *Binned entropy* and *Fast Fourier Transform (FFT)* coefficients are also included in the list of most important features.

In contrast, Figure 6.7 shows the top 10 features for Bagged AdaBoost, the second-best performing approach. Here, it can be seen that there are fewer unique features present. In fact, aggregate linear trend accounts for half of the top 10 features. Given that both of these modeling techniques incorporate boosting, it is an indicator that a more diverse feature set leads to a higher performing aggregate predictor.

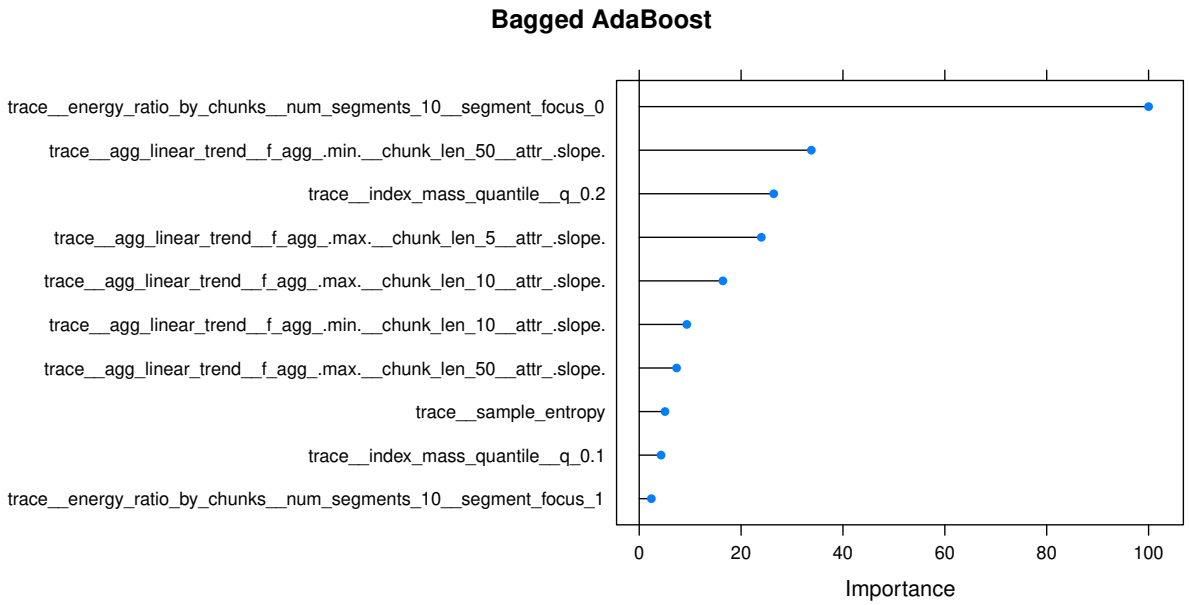

Figure 6.7: Feature importance for the second best performing modeling technique, Bagged AdaBoost. Note that there is less diversity among the most important features than those presented in Figure 6.6.

6.4 Results

Real data were generated from the microseismic monitoring project at the Marcellus Shale and are considered here. A linear array consisting of eleven three-component geophones were deployed. As is common in real microseismic data, there were varying levels of noise, missing information, and a number of events that were negatively impacted due to ringing which was likely caused by insufficient clamping force between the sensor and the borehole casing (Gaiser et al., 1988). There were 1842 microseismic events recorded; however, a number of these events exhibited high noise and the absence of compressional wave arrivals. As such, it is important to incorporate events that have a relatively high compressional wave S/N in the training data. 249 traces were considered for model training and testing from multiple stages throughout the hydraulic fracturing project (Figure 6.8). This is important in order to train the model on a variety of source mechanisms and source-receiver paths.

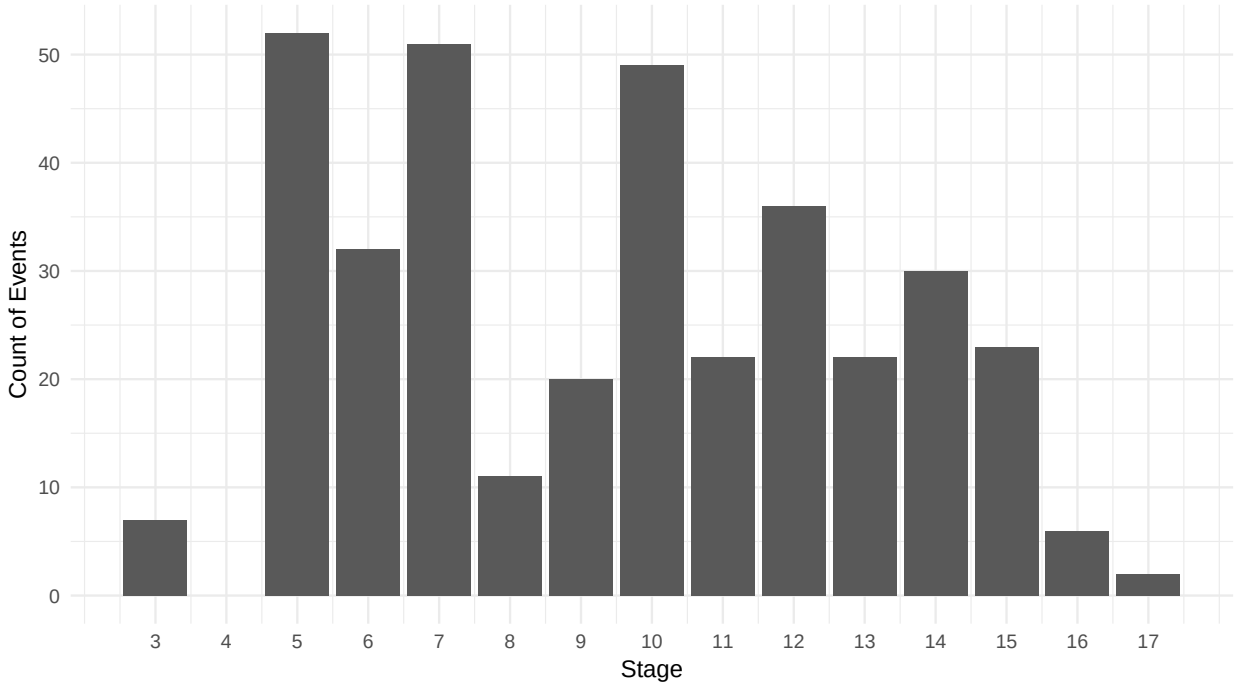

Figure 6.8: While noise content was present throughout the hydraulic fracturing process, it is still important to analyze microseismic events from as many stages as possible. Based on the overall level of noise present, the distribution of events on a stage-basis is shown here. Earlier stages contained more noise, likely due to a significantly greater source-receiver distance and accompanying scattering effects.

Approximately 63% of the data were used for model training and 37% were used for model validation. It is important to note here that this split is done on a trace-specific basis based on event identifier and geophone level in order to avoid data leakage. Group K-Fold cross-validation is also used during the training phase in order to improve performance while minimizing the risk of overfitting from data leakage. The classification performance, as well as accompanying statistical information, is presented via the confusion matrix, which is produced from the test data in Figure 6.9. Here, it is important to note that there is a class imbalance and the majority class is the non-arrival class (85.4%). Downsampling of the majority class was performed; however, sensitivity and specificity are still important classification performance metrics that should be considered along with overall accuracy above the NIR. Figure 6.10 demonstrates this class imbalance and Figure 6.11 demonstrates the overall classification performance with respect to specificity and sensitivity. Here, the Area Under the Curve (AUC) is a measure of the overall predictive power of the model. The maximum value for an ideal learner is 100% and random guessing will produce an AUC of 50%, which is represented by the diagonal line in the chart. It can be seen that the AUC for this method is 90.4%.

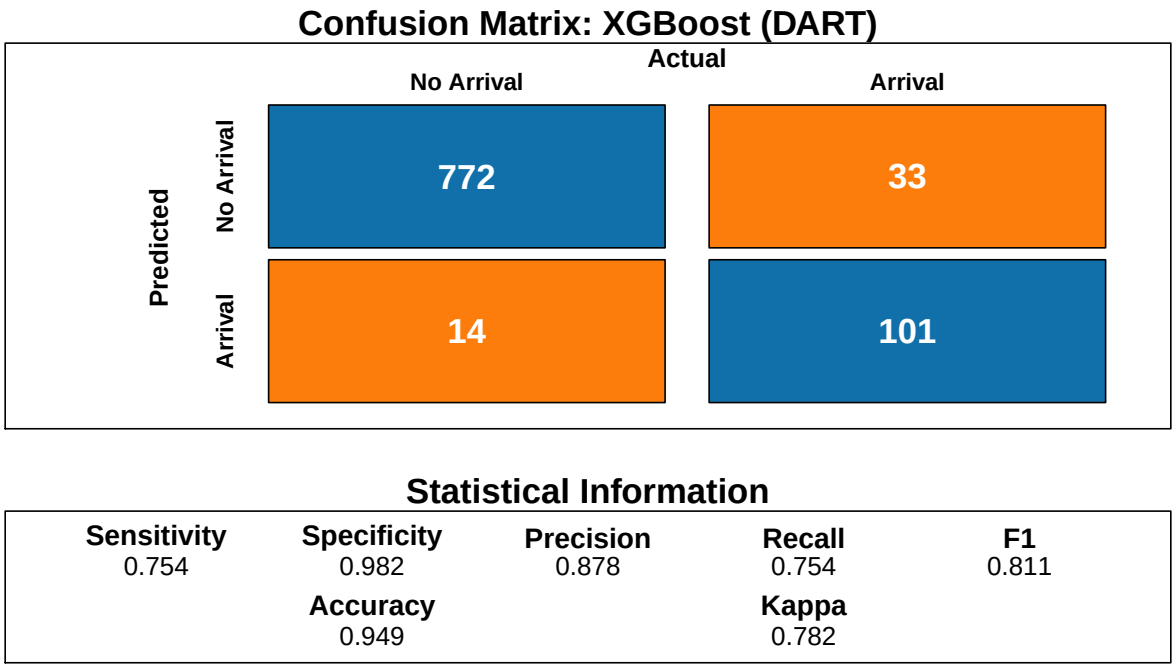

Figure 6.9: Standard classification performance measures are presented in the confusion matrix. Blue rectangles represent optimal predictions (true positive and true negative), while the orange rectangles represent misclassifications (false positive and false negative). In this modeling endeavor, false positives lead to higher overall error given the nature of the subsequent first arrival picking step. Statistical information is also included. Note that the No Information Rate (85.4%) must be considered when evaluating overall accuracy due to a large class imbalance.

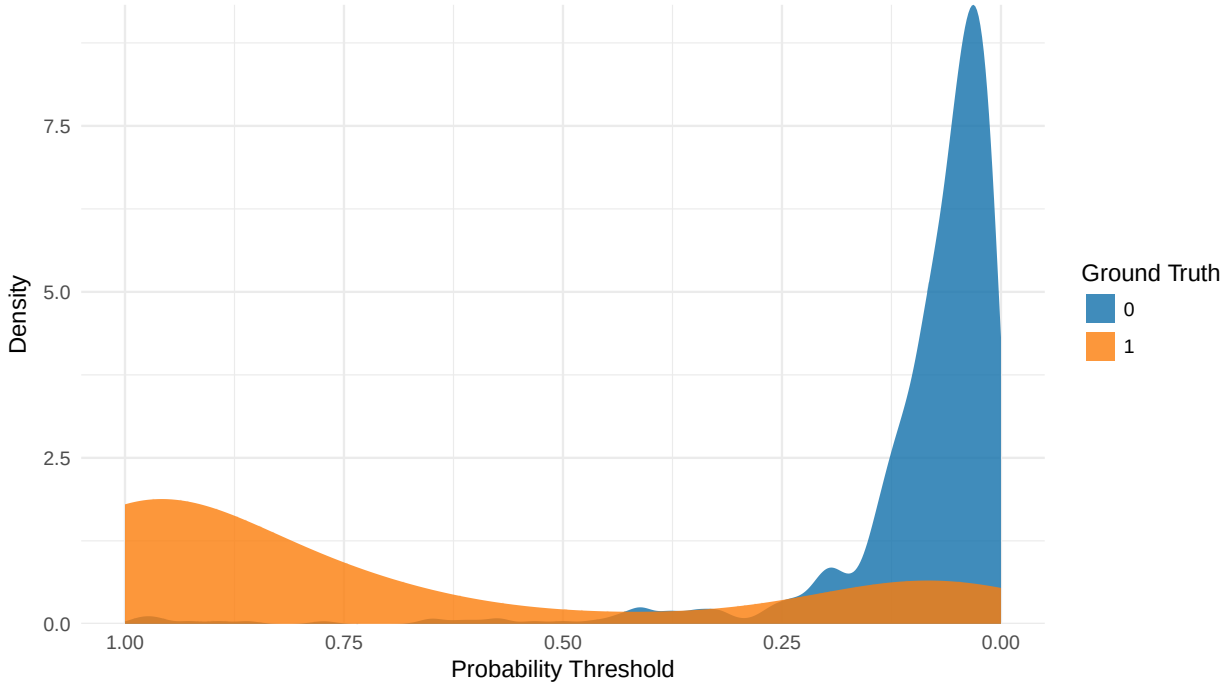

Figure 6.10: A large class imbalance is present in the real microseismic data considered in this chapter. This is illustrated by the significantly larger "no arrival" class that is shown in blue versus the "arrival" class that is shown in orange. It is also important to note that for the arrival class (orange), a bimodal distribution can be inferred by the increase in density between 0.25 and 0.00. This is likely an artifact of the disparity between compressional wave window length and chunk length. This likely impacts the number of misclassifications present in the overall predictions.

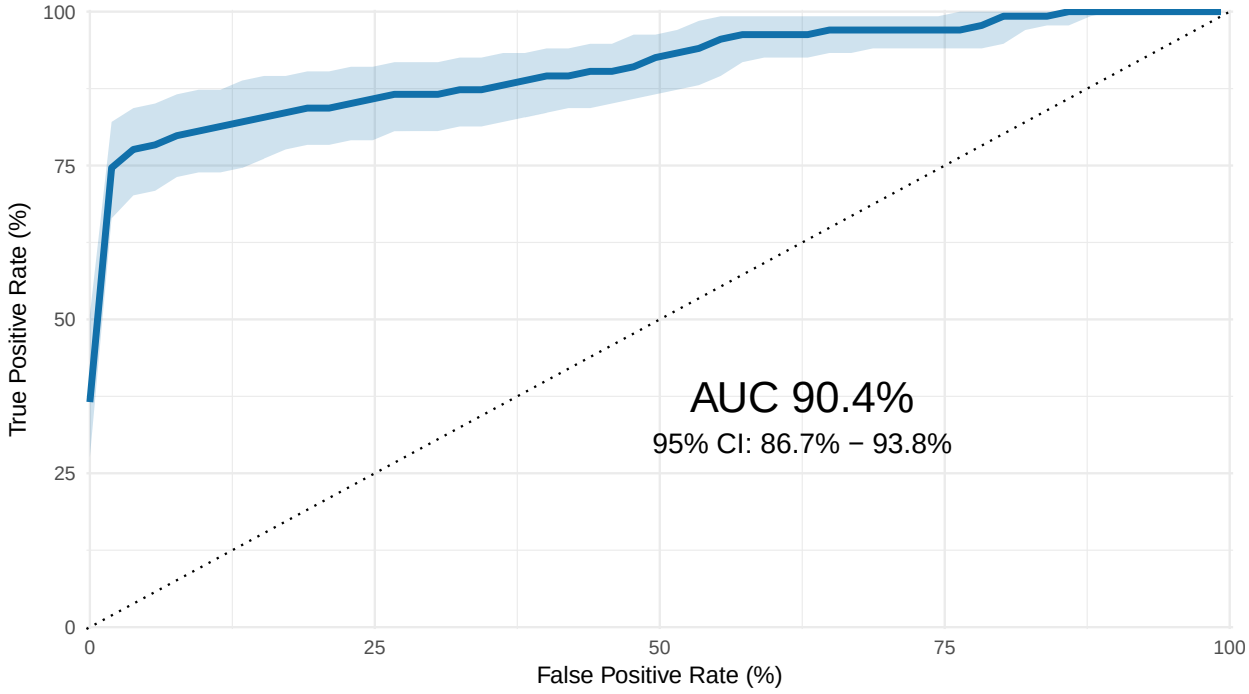

Figure 6.11: ROC plot shows that the overall classification performance is good. Area Under the Curve (AUC) is calculated to be 90.4% which indicates positive results. It is important to note that this plot relied on bootstrap sampling and the confidence bands are shown to represent that fact.

A chunk length of 200 samples was determined to be an acceptable chunk length based on overall performance. Additionally, the final compressional wave window length was also 200 samples. Specifically, the window begins 50 samples before the known pick time and concludes 150 samples following the pick time. This captures the signal immediately before the first arrival and also captures the main content of the compressional wave. In order to determine if an arrival is present within a given signal chunk, we consider the percentage of samples within a chunk that contain the windowed arrival. Iterating through various levels of presence, the optimal threshold for determining whether a chunk contains an arrival was determined to be 60%. These dynamic parameters can be tuned in order to maximize performance based on the traditional classification performance metrics in combination with mean absolute error between the predicted arrival time and the actual arrival time. An example of the chunk lengths and compressional window lengths can be seen in Figure 6.12.

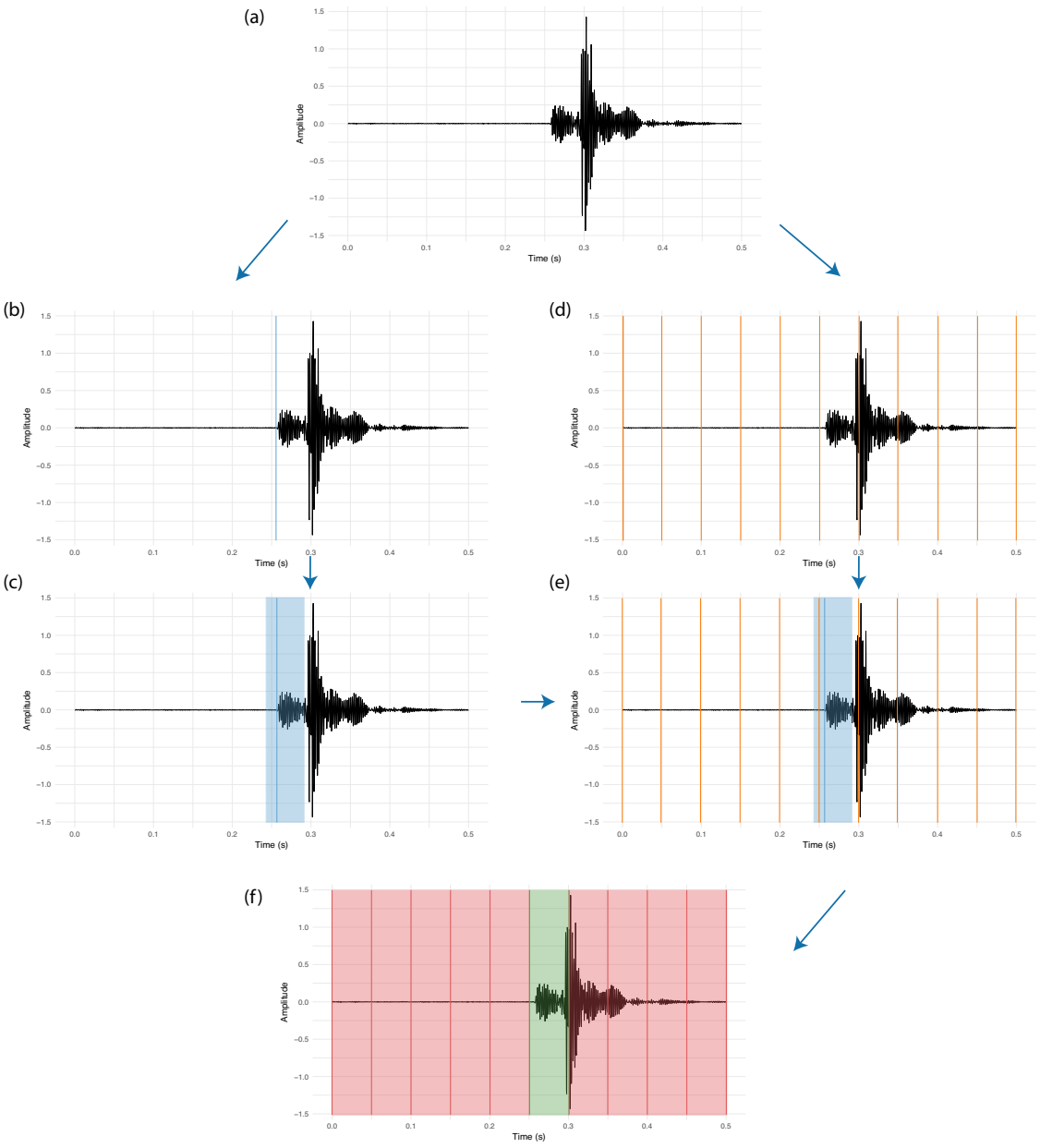

Figure 6.12: An appropriate compressional wave window and chunk length must be determined to create a target variable for the raw signal (a). First, the known first arrival pick time is considered and is shown by the blue line (b). Then, an appropriate compressional window length is determined, which is shown by the blue rectangle (c). In this case, the window begins 50 samples before the arrival and 150 samples after the arrival. Concurrently, a chunk length is determined and is shown by the orange lines (d). In this case, the chunk length is also 200 samples. In order to determine if a given chunk contains an arrival, the compressional wave and chunk must be considered together (e). If the compressional window accounts for 60% of the samples in the chunk, then that chunk is assigned the label of "arrival" for classification via machine learning methods (f). It is important to note that these are dynamic parameters and can be tuned to optimize performance with new data.

STA/LTA has been shown to accurately identify arrival times in seismic and microseismic data. STA/LTA is used to identify arrival times with real data and the performance is compared to our proposed method. The process flow for STA/LTA can be seen in Figure 6.13. Additionally, Figure 6.14 and Figure 6.15 show a comparison of the distribution of errors between predicted arrival time and actual arrival time for our proposed time series classification method and the traditional STA/LTA approach. Moreover, aggregate measures are important in understanding the overall effectiveness of these approaches. As such, the mean absolute error (MAE) for the traditional method is 126 ms whereas the MAE for our proposed method is 23.8 ms.

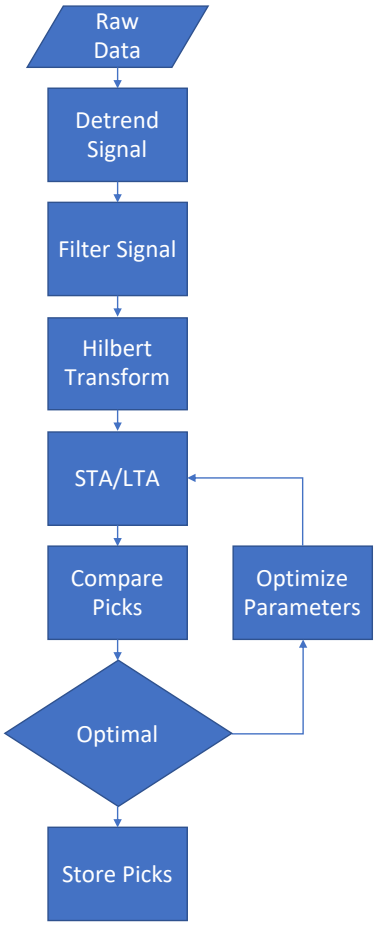

Figure 6.13: Raw data is first detrended, then filtered with a standard band pass filter with cutoff frequencies at 1 Hz and 90 Hz. Then an envelope function is applied through the use of the Hilbert transform. From here, the STA/LTA method is implemented and a list of picks are generated. In the aggregate, the performance of these picks is considered and the STA/LTA parameters are changed to achieve optimal performance. The output is a list of automatically picked first arrival times.

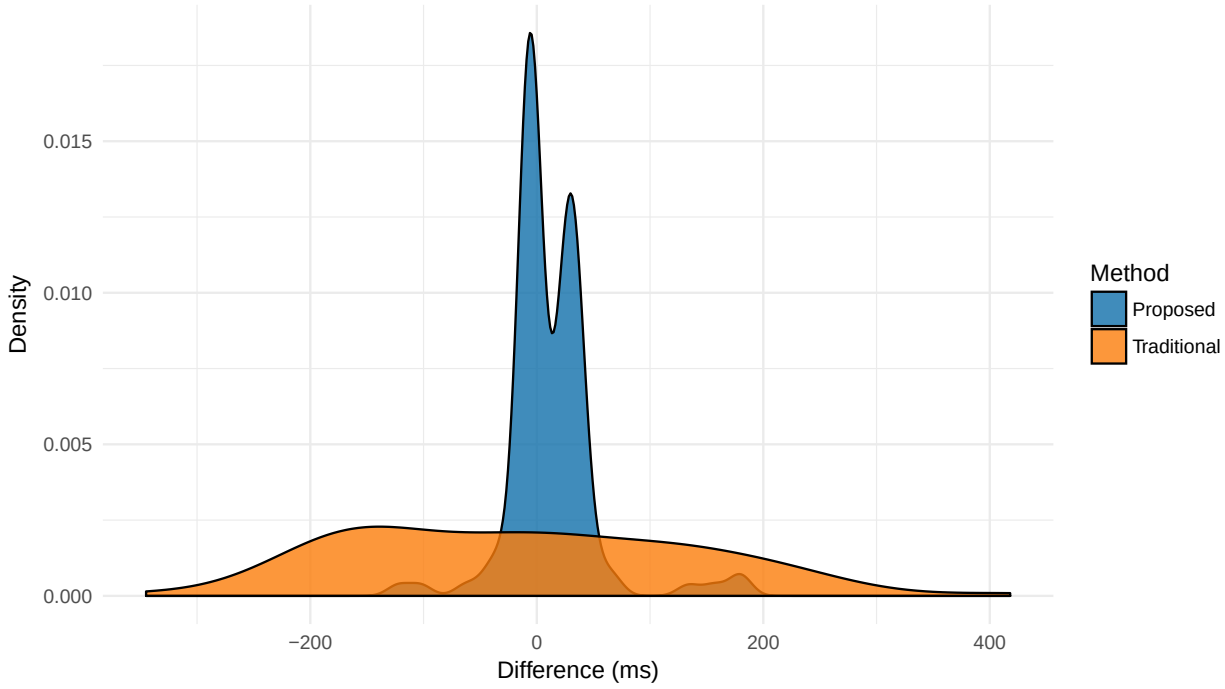

Figure 6.14: Density plots show a comparison of the distribution of errors between predicted arrival time and actual arrival time for our proposed time series classification method (blue) and the traditional STA/LTA approach (orange). There is a larger percentage of the total errors that are centered closer to zero with our proposed method, which indicates that it outperforms the traditional method.

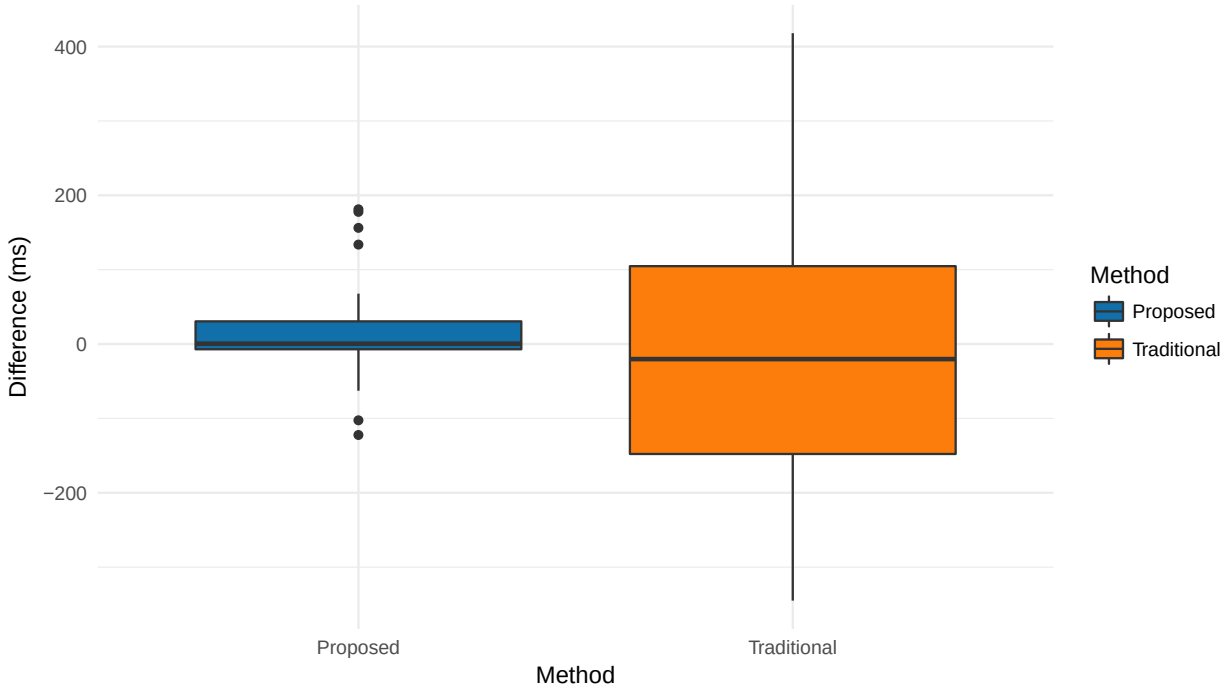

Figure 6.15: Box plots show a comparison of the distribution of errors between predicted arrival time and actual arrival time for our proposed time series classification method (blue) and the traditional STA/LTA approach (orange). A significantly smaller spread is seen in the Interquartile Range (IQR), which demonstrates that our proposed method results in generally smaller error than the traditional method.

Another method of understanding the difference in performance on a chunk-specific basis is to consider the difference in errors between our proposed method and the traditional method. We first calculate the absolute error between predicted time and actual time for each chunk with our proposed method, then the same calculation is performed for each trace with STA/LTA. Next, the difference in absolute errors is calculated. Here, the sign is important because it indicates which approach has the greater error. If the difference in error is positive, then it indicates that the proposed method outperforms the traditional method. Figure 6.16 shows the difference in errors between our proposed method and the traditional method. Note that the majority of the cases have a positive value, which indicates that our proposed method outperforms STA/LTA on the majority of traces considered. As such, it can be seen that the proposed approach outperforms STA/LTA.

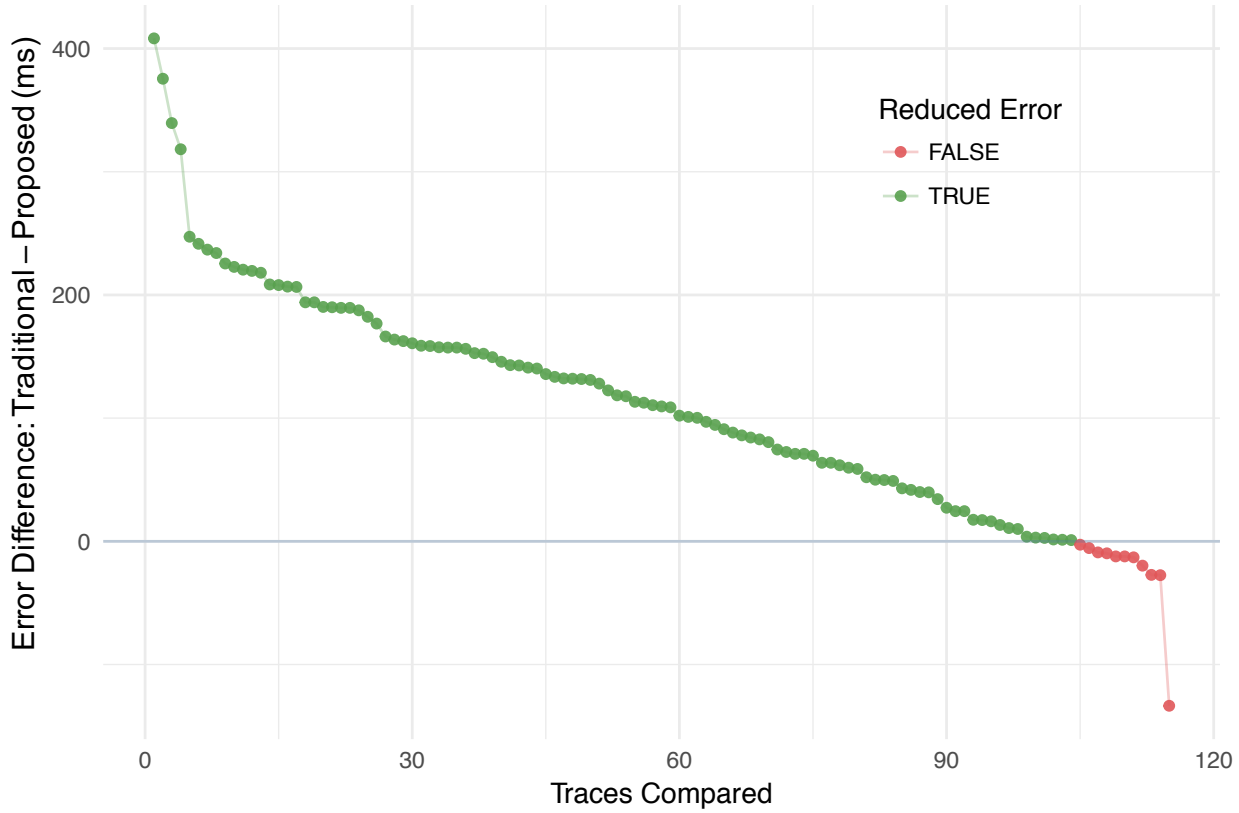

Figure 6.16: Error differences between our proposed method and the traditional method on a chunk-specific basis are shown. Absolute errors are calculated for each method and then the difference between those errors is calculated and presented. Positive values indicate that our proposed method outperforms the traditional method for a given trace, shown in green. Conversely, negative values show the cases where our approach does not outperform the traditional method, shown in red. It is clear that the majority of cases lead to positive values, which indicates superior performance through our proposed method.

Figure 6.17 shows an example of a real microseismic record. The green vertical line represents the contractor-provided pick, the blue vertical line represents the pick from our proposed method, and the orange vertical line represents the pick from STA/LTA. The pick resulting from the STA/LTA approach has the greatest error, while our approach detects the true first motion.

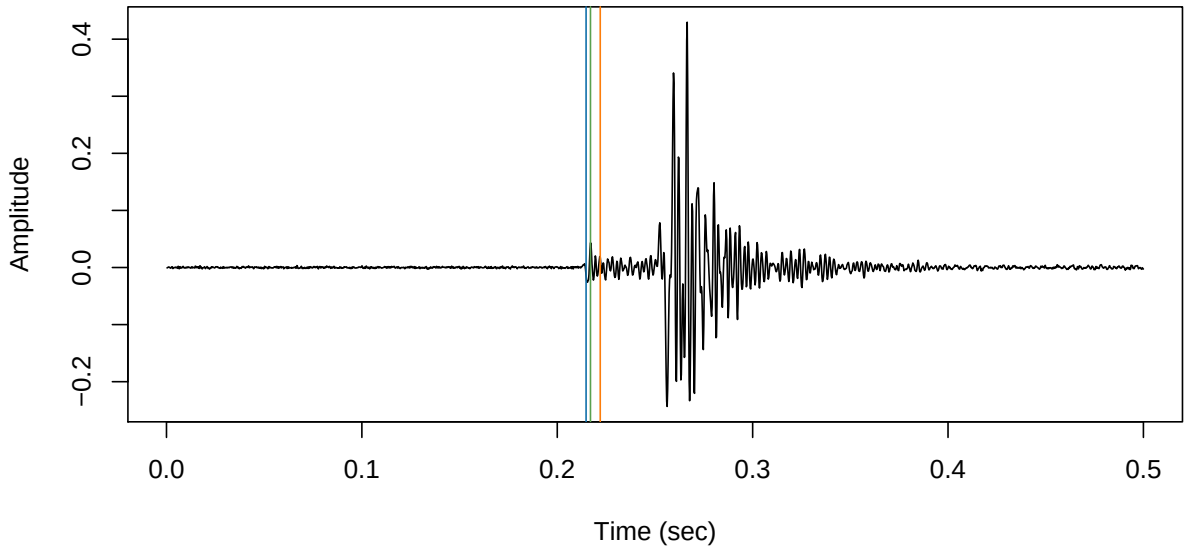

Figure 6.17: Real microseismic record is presented as an example of model performance. The green vertical line represents the contractor-provided pick, the blue vertical line represents the pick from our proposed method, and the orange vertical line represents the pick from STA/LTA. Our proposed method accurately identifies the time where first motion occurs.

6.5 Conclusion

There commonly exists a tradeoff between reduction in human workload through automation of arrival picking and overall accuracy. Current automated picking techniques are limited in the presence of man-made seismicity or high noise environments. The proposed method in this chapter leverages machine learning methods to reduce the amount of time required to manually pick events while improving picking accuracy. Additionally, through the selection of ensemble methods, the impact of noise is reduced and the negative effect of overfitting is greatly mitigated. Furthermore, this approach establishes an extensible framework using widely-available learning packages that can be adapted for a number of microseismic data sets. This method can be used as a stand-alone approach or to improve the outcome of traditional automated picking methods. Finally, an extension of this approach can be used to perform multi-class classification in order to identify first arrivals as well as shear wave arrivals to significantly improve the overall analysis process.

Chapter 7

Conclusion

Out of damp and gloomy days, out of solitude, out of loveless words directed at us, conclusions grow up in us like fungus one morning they are there, we know not how, and they gaze upon us, morose and gray. Woe to the thinker who is not the gardener but only the soil of the plants that grow in him.

– Friedrich Nietzsche

7.1 Summary of Contributions

The unifying thread of the work presented in this dissertation is that of identifying areas where the cultivation of knowledge is limited, attempting to identify and separate those constraints which are immovable from those that exhibit the potential for flexibility, and then applying methodologies in an interdisciplinary approach to overcome significant limitations. Here, we focus primarily on data quality issues that severely limit our understanding of hydraulic fracturing, which are ubiquitous in any engineering problem that utilizes real data.

Whether missing values occur due to transmission and recording issues or the true content of the signal is occluded by environmental noise, there will always be the need for improving the collected data. While this is an arduous task, it is fundamental for any subsequent analysis step. We explore a number of different conditions that would lead to a loss of information in this dissertation and have developed approaches to overcome these limitations.

Engineering Solutions to Data Conditioning Problems

In Chapter 3, we explored the use of analysis in the spectral domain to overcome the limitations that arise from single well downhole monitoring geometries. Here, physical placement of monitoring arrays limit the ability to interpret and understand specific aspects of microseismic events from hydraulic fracturing. While a first-order solution would be to increase

the number of sensors and change their physical placement, economic constraints prevent this from being a viable option. As such, we turn to the spectral domain to create features that enable the characterization of microseismic events.

In Chapter 4, we extend this exploration to incorporate all available information to improve microseismic location estimates. We reduced noise contamination by understanding and removing the spectral content that comes from resonant artifacts in the data due to poor coupling between the geophone and borehole wall. Additionally, through the identification and incorporation of a geophysical phenomenon referred to as head waves, we model a more accurate estimation of microseismic event location. Finally, we recommend an optimal monitoring geometry that balances the ability to reuse the monitoring well, while also reducing S/N and capturing head waves. This monitoring geometry will enable future analysts to better understand the fracturing of rock due to hydraulic fracturing.

Artificial Intelligence and Machine Learning as a Means Recovering Information

In Chapter 5, we turn to machine learning and artificial intelligence to better understand the nature of missing or corrupt data present in the data set under consideration. Most imputation approaches seek only to enable the successful execution of machine learning techniques and to minimize bias. However, these objectives fall short of recovering information that can be used in subsequent analyses. Through the use of machine learning and deep learning methods, we explore the utility of data-driven imputation.

The results illustrate that a single solution does not exist and the optimal approach depends on the amount of available computational time as well as the intended use of the imputation schema.

Improving Automated Analysis with Ensemble Learning

In Chapter 6, we continue to explore the utility of learning methods when applied to problems in geophysics. Here, the objective was to leverage ensemble learning methods to overcome the issue of noisy or corrupt data with the express goal of improving traditional automated arrival picking - a fundamental, albeit time-consuming, first step in the journey of understanding hydraulic fracturing through microseismic analysis. A novel framework was created to leverage dynamic parameterization to provide an extensible computational paradigm.

The results of this time series classification endeavor demonstrate that through the use of ensemble learning methods, dramatic improvements can be achieved when attempting to automatically pick arrival times in the presence of high noise content due to man-made seismicity. Further, the execution time of this computational paradigm has the potential to provide very real savings when compared to the arduous, time-intensive task of manually picking microseismic events. Moreover, there is value in the use of an objective, transparent system of analysis that eliminates the negative impact of subjective interpretation that occurs with current human-in-the-loop analysis workflows.

Finally, given the data-driven nature of this framework, minimal changes are required to change the classification objective of the system. This attribute makes the system reusable for many other analysis tasks in the geophysics community.

7.2 Future Research

The field of data science, which leverages machine learning and artificial intelligence methodologies to create data-driven solutions, is only going to continue to grow. With that in mind, the potential for future research in the arena of hydraulic fracturing when combined with data science is significant. The work presented in this dissertation, specifically, the computational paradigm created to improve automated arrival picking, can be extended to identify and separate waveforms with minimal effort. The implications of this are also significant. In Chapter 4, we discussed the value of identifying head waves in the data in order to improve microseismic event location estimates. With the use of a system capable of classifying common waveforms, head waves can be more easily identified in the presence of noisy or corrupt data. As a result, location estimates can be also improved in other openly available or proprietary hydraulic fracturing data sets. The end result, and the true objective of analysis of hydraulic fracturing data, is to better understand the physical changes occurring deep below the surface in an effort to create more efficient and safer hydraulic fracturing projects. The work presented here endeavors to contribute to that goal.

Bibliography

Aki, K. and Richards, P. G. (2002). *Quantitative seismology*.

Alexander, T., Baihly, J., Boyer, C., Clark, B., Waters, G., Jochen, V., Calvez, J., Lewis, R., Miller, C., Thaeler, J., et al. (2011). Shale gas revolution: Oilfield review autumn. *Schlumberger*, 23:40–55.

Alfaro, E., Gamez, M., Garcia, N., et al. (2013). Adabag: An r package for classification with boosting and bagging. *Journal of Statistical Software*, 54(2):1–35.

Álvarez, I., García, L., Mota, S., Cortés, G., Benítez, C., and De la Torre, Á. (2013). An automatic p-phase picking algorithm based on adaptive multiband processing. *IEEE Geoscience and Remote Sensing Letters*, 10(6):1488–1492.

Artman, B., Podladtchikov, I., and Witten, B. (2010). Source location using time-reverse imaging. *Geophysical Prospecting*, 58(5):861–873.

Artman, B. and Witten, B. (2011). Wave-equation microseismic imaging and event selection in the image domain. In *SEG Technical Program Expanded Abstracts 2011*, pages 1699–1703. Society of Exploration Geophysicists.

Bao, X. and Eaton, D. W. (2016). Fault activation by hydraulic fracturing in western canada. *Science*, 354(6318):1406–1409.

Belayouni, N., Gesret, A., Daniel, G., and Noble, M. (2015). Microseismic event location using the first and reflected arrivalsmicroseismic event location. *Geophysics*, 80(6):WC133–WC143.

Benjamini, Y., Yekutieli, D., et al. (2001). The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4):1165–1188.

Beresnev, I. A. (2001). What we can and cannot learn about earthquake sources from the spectra of seismic waves. *Bulletin of the Seismological Society of America*, 91(2):397–400.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Bracewell, R. N. and Bracewell, R. N. (1986). *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brune, J. N. (1970). Tectonic stress and the spectra of seismic shear waves from earthquakes. *Journal of geophysical research*, 75(26):4997–5009.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68.

Cabrera, D., Sancho, F., Li, C., Cerrada, M., Sánchez, R.-V., Pacheco, F., and de Oliveira, J. V. (2017). Automatic feature extraction of time-series applied to fault severity assessment of helical gearbox in stationary and non-stationary speed operation. *Applied Soft Computing*, 58:53–64.

Capilla, C. (2006). Application of the haar wavelet transform to detect microseismic signal arrivals. *Journal of applied geophysics*, 59(1):36–46.

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subseque nt selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107.

Červenỳ, V. and Ravindra, R. (1971). Theory of seismic head waves, 331.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77.

Ciezobka, J., Maity, D., Salehi, I., et al. (2016). Variable pump rate fracturing leads to improved production in the marcellus shale. In *SPE Hydraulic Fracturing Technology Conference*. Society of Petroleum Engineers.

Cipolla, C., MG, M., SC, M., and RC, D. (2011). A practical guide to interpreting microseismic measurements: Presented at the north american unconventional gas conference and exhibition, spe. Technical report, SPE-144067-MS, doi: https://doi. org/10.2118/144067-MS.

Cladouhos, T. T., Petty, S., Nordin, Y., Moore, M., Grasso, K., Uddenberg, M., Swyer, M., Julian, B., and Foulger, G. (2013). Microseismic monitoring of newberry volcano egs demonstration. In *Proceedings of the 38th Workshop on Geothermal Reservoir Engineering, Stanford, CA*, pages 11–13.

Claerbout, J. F. (1992). *Earth soundings analysis: Processing versus inversion*, volume 6. Blackwell Scientific Publications London.

Coffin, S., Hur, Y., Abel, J. S., Culver, B., Augsten, R., and Westlake, A. (2012). Beyond first arrivals: Improved microseismic event localization using both direct-path and head-wave arrivals. In *SEG Technical Program Expanded Abstracts 2012*, pages 1–5. Society of Exploration Geophysicists.

Dong, W. and Toksöz, M. N. (1995). Borehole seismic-source radiation in layered isotropic and anisotropic media: Real data analysis. *Geophysics*, 60(3):748–757.

Dreger, D., Uhrhammer, R., Pasyanos, M., Franck, J., and Romanowicz, B. (1998). Regional and far-regional earthquake locations and source parameters using sparse broadband networks: A test on the ridgecrest sequence. *Bulletin of the Seismological Society of America*, 88(6):1353–1362.

Du, J., Warpinski, N., Waltman, C., et al. (2013). Anisotropic effects on polarization from highly deviated/horizontal wells in microseismic monitoring of hydraulic fractures. In *2013 SEG Annual Meeting*. Society of Exploration Geophysicists.

Duncan, P. and Eisner, L. (2010). Reservoir characterization using surface microseismic monitoring: Geophysics 75. *75A139–75A146, doi*, 10(1.3467760).

Eaton, D. W. (2011). Q determination, corner frequency and spectral characteristics of microseismicity induced by hydraulic fracturing. In *SEG Technical Program Expanded Abstracts 2011*, pages 1555–1559. Society of Exploration Geophysicists.

Eaton, D. W. (2014). Magnitude, scaling, and spectral signature of tensile microseisms. In *EGU General Assembly Conference Abstracts*, volume 16.

Eisner, L., Duncan, P. M., Heigl, W. M., and Keller, W. R. (2009). Uncertainties in passive seismic monitoring. *The Leading Edge*, 28(6):648–655.

Eisner, L., Le Calvez, J. H., et al. (2007). New analytical techniques to help improve our understanding of hydraulically induced microseismicity and fracture propagation. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, 8(1):81–98.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.

Gaiser, J. E., Fulp, T. J., Petermann, S. G., and Karner, G. M. (1988). Vertical seismic profile sonde coupling. *Geophysics*, 53(2):206–214.

Galiana-Merino, J. J., Rosa-Herranz, J. L., and Parolai, S. (2008). Seismic p phase picking using a kurtosis-based criterion in the stationary wavelet domain. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3815–3826.

Gibbons, S. J. and Ringdal, F. (2006). The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International*, 165(1):149–166.

Gibbons, S. J., Ringdal, F., and Kværna, T. (2012). Ratio-to-moving-average seismograms: a strategy for improving correlation detector performance. *Geophysical Journal International*, 190(1):511–521.

Gill, M. K., Asefa, T., Kaheil, Y., and McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. *Water resources research*, 43(7).

Harris, K. and Bacon, R. (2015). Utilizing source mechanism and microseismic event location to identify faults in real-time using wireless seismic recording systems–an eagle ford case study. *first break*, 33(7).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Haukoos, J. S. and Newgard, C. D. (2007). Advanced statistics: missing data in clinical research—part 1: an introduction and conceptual framework. *Academic Emergency Medicine*, 14(7):662–668.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Jones, J. P. and van der Baan, M. (2015). Adaptive sta–lta with outlier statistics. *Bulletin of the Seismological Society of America*, 105(3):1606–1618.

Katz, G., Shin, E. C. R., and Song, D. (2016). Explorekit: Automatic feature generation and selection. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 979–984. IEEE.

Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):15.

King, G. E. et al. (2012). Hydraulic fracturing 101: What every representative, environmentalist, regulator, reporter, investor, university researcher, neighbor and engineer should know about estimating frac risk and improving frac performance in unconventional gas and oil wells. In *SPE hydraulic fracturing technology conference*. Society of Petroleum Engineers.

Kondrashov, D. and Ghil, M. (2006). Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, 13(2):151–159.

Kuang, W., Zoback, M., and Zhang, J. (2017). Estimating geomechanical parameters from microseismic plane focal mechanisms recorded during multistage hydraulic fracturing. *Geophysics*, 82(1):KS1–KS11.

Kuhn, M. et al. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, 28(5):1–26.

Li, J., Li, C., Morton, S. A., Dohmen, T., Katahara, K., and Nafi Toksöz, M. (2014). Microseismic joint location and anisotropic velocity inversion for hydraulic fracturing in a tight bakken reservoir. *Geophysics*, 79(5):C111–C122.

Li, X. and Dong, L. (2014). An efficient closed-form solution for acoustic emission source location in three-dimensional structures. *AIP Advances*, 4(2):027110.

Li, X., Shang, X., Wang, Z., Dong, L., and Weng, L. (2016). Identifying p-phase arrivals with noise: An improved kurtosis method based on dwt and sta/lta. *Journal of Applied Geophysics*, 133:50–61.

Maxwell, S. (2009). Microseismic location uncertainty. *CSEG Recorder*, 34(4):41–46.

Maxwell, S. (2010). Microseismic: Growth born from success. *The Leading Edge*, 29(3):338–343.

Maxwell, S. (2014). *Microseismic imaging of hydraulic fracturing: Improved engineering of unconventional shale reservoirs*. Society of Exploration Geophysicists.

Maxwell, S. and Cipolla, C. (2011). What does microseismicity tell us about hydraulic fracturing? paper spe 146932 presented at the spe annual technical conference and exhibition, denver, colorado, 30 october–2 november.

Maxwell, S. C., Rutledge, J., Jones, R., and Fehler, M. (2010). Petroleum reservoir characterization using downhole microseismic monitoring. *Geophysics*, 75(5):75A129–75A137.

Mierswa, I. and Morik, K. (2005). Automatic feature extraction for classifying audio data. *Machine learning*, 58(2-3):127–149.

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, 61(1):1–15.

Monner, D. and Reggia, J. A. (2012). A generalized lstm-like training algorithm for second-order recurrent neural networks. *Neural Networks*, 25:70–83.

Naul, B., van der Walt, S., Crellin-Quick, A., Bloom, J. S., and Pérez, F. (2016). Cesium: open-source platform for time-series inference. *arXiv preprint arXiv:1609.04504*.

Nava, M. J., Rector, J. W., and Zhang, Z. (2015). Characterization of microseismic source mechanism in the marcellus shale through analysis in the spectral domain. In *SEG Technical Program Expanded Abstracts 2015*, pages 5069–5073. Society of Exploration Geophysicists.

Nava, M. J., Rector, J. W., and Zhang, Z. (2020a). Automatic first arrival picking of microseismic events via ensemble learning methods. *In Review*.

Nava, M. J., Rector, J. W., and Zhang, Z. (2020b). Recovering compressional wave amplitudes with multiple imputation by chained equations and random forest imputation. *In Review*.

Nisbet, R., Elder, J., and Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

Parra, J., Hackert, C., Xu, P.-C., and Collier, H. A. (2006). Attenuation analysis of acoustic waveforms in a borehole intercepted by a sand-shale sequence reservoir. *The Leading Edge*, 25(2):186–193.

Parra, J. O., Hackert, C. L., Gorody, A. W., and Korneev, V. (2002). Detection of guided waves between gas wells for reservoir characterization. *Geophysics*, 67(1):38–49.

Petersson, N. A. and Sjogreen, B. (2013). User's guide to sw4, version 1.0. *Lawrence Livermore National Laboratory Tech. Rept. LLNL-SM*, 642292:114.

Quinlan, J. R. et al. (1996). Bagging, boosting, and c4. 5. In *AAAI/IAAI, Vol. 1*, pages 725–730.

Quintal, B., Steeb, H., Frehner, M., Schmalholz, S. M., and Saenger, E. H. (2012). Pore fluid effects on s-wave attenuation caused by wave-induced fluid flowpore fluid effects on s-wave attenuation. *Geophysics*, 77(3):L13–L23.

Rashmi, K. V. and Gilad-Bachrach, R. (2015). Dart: Dropouts meet multiple additive regression trees. In *AISTATS*, pages 489–497.

Richards, P. G. and Aki, K. (1980). *Quantitative Seismology: Theory and Methods*. Freeman.

Richman, J. S., Lake, D. E., and Moorman, J. R. (2004). Sample entropy. In *Methods in enzymology*, volume 384, pages 172–184. Elsevier.

Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049.

Rutledge, J. T. and Phillips, W. S. (2003). Hydraulic stimulation of natural fractures as revealed by induced microearthquakes, carthage cotton valley gas field, east texas hydraulic stimulation of natural fractures. *Geophysics*, 68(2):441–452.

Salehi, I. A., Ciezobka, J., et al. (2013). Controlled hydraulic fracturing of naturally fractured shales-a case study in the marcellus shale examining how to identify and exploit natural fractures. In *SPE Unconventional Resources Conference-USA*. Society of Petroleum Engineers.

Schoenberg, M. (1980). Elastic wave behavior across linear slip interfaces. *The Journal of the Acoustical Society of America*, 68(5):1516–1521.

Senkaya, M. and Karsli, H. (2014). A semi-automatic approach to identify first arrival time: the cross-correlation technique (cct). *Earth Sciences Research Journal*, 18(2):107–113.

Severyn, A. and Moschitti, A. (2013). Automatic feature engineering for answer selection and extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 458–467.

Sleefe, G., Warpinski, N., Engler, B., et al. (1995). The use of broadband microseisms for hydraulic fracture mapping. *SPE Formation Evaluation*, 10(04):233–240.

Song, F. and Toksöz, M. N. (2011). Full-waveform based complete moment tensor inversion and source parameter estimation from downhole microseismic data for hydrofracture monitoring. *Geophysics*, 76(6):WC103–WC116.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450.

Trnkoczy, A. (1999). Topic understanding and parameter setting of sta/lta trigger algorithm. *New manual of seismological observatory practice*, 2.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91.

Vavryčuk, V. (2007). On the retrieval of moment tensors from borehole data. *Geophysical Prospecting*, 55(3):381–391.

Warpinski, N. et al. (2009). Microseismic monitoring: Inside and out. *Journal of Petroleum Technology*, 61(11):80–85.

Warpinski, N., Kramm, R. C., Heinze, J. R., Waltman, C. K., et al. (2005). Comparison of single-and dual-array microseismic mapping techniques in the barnett shale. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Warpinski, N. R., Du, J., et al. (2010). Source-mechanism studies on microseismicity induced by hydraulic fracturing. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Wilson, S., Raymer, D., and Jones, R. (2003). The effects of velocity structure on microseismic location estimates: A case study. In *SEG Technical Program Expanded Abstracts 2003*, pages 1565–1568. Society of Exploration Geophysicists.

Xiantai, G., Zhimin, L., Na, Q., and Weidong, J. (2011). Adaptive picking of microseismic event arrival using a power spectrum envelope. *Computers & geosciences*, 37(2):158–164.

Yilmaz, Ö. (2001). *Seismic data analysis: Processing, inversion, and interpretation of seismic data*. Society of exploration geophysicists.

Yuan, D. and Li, A. (2016). Determination of microseismic event back azimuth from s-wave splitting analysis. In *SEG Technical Program Expanded Abstracts 2016*, pages 2667–2671. Society of Exploration Geophysicists.

Yuan, D. and Li, A. (2017). Joint inversion for effective anisotropic velocity model and event locations using s-wave splitting measurements from downhole microseismic data. *Geophysics*, 82(3):C133–C143.

Yue, B., Peng, Z., and Zhang, Q. (2014). Seismic wavelet estimation using covariation approach. *IEEE transactions on geoscience and remote sensing*, 52(12):7495–7503.

Zhang, Z., Nava, M., and Rector, J. (2016). Resonance in downhole microseismic data and its removal. In *SEG Technical Program Expanded Abstracts 2016*, pages 2652–2656. Society of Exploration Geophysicists.

Zhang, Z., Rector, J. W., and Nava, M. J. (2015). Improving microseismic event location accuracy with head wave arrival time: Case study using marcellus shale. In *SEG Technical Program Expanded Abstracts 2015*, pages 2473–2478. Society of Exploration Geophysicists.

Zhang, Z., Rector, J. W., and Nava, M. J. (2017a). Microseismic hydraulic fracture imaging in the marcellus shale using head waves. *Geophysics*, 83(2):KS1–KS10.

Zhang, Z., Rector, J. W., and Nava, M. J. (2017b). Simultaneous inversion of multiple microseismic data for event locations and velocity model with bayesian inference. *Geophysics*, 82(3):KS27–KS39.

Zimmer, U. (2010). Localization of microseismic events using headwaves and direct waves. In *SEG Technical Program Expanded Abstracts 2010*, pages 2196–2200. Society of Exploration Geophysicists.

Zimmer, U. (2011). Microseismic design studies. *Geophysics*, 76(6):WC17–WC25.