Application of Machine Learning Algorithms in Characterisation of Sonic Wave Velocities within Various Geological Formations

Mobarakeh Mohammadpour

(Doctor of Philosophy)

A thesis submitted for the degree of *Doctor of Philosophy* at

Monash University in 2024

(Department of Civil Engineering)

Abstract

This thesis investigates the applicability of Machine Learning (ML) algorithms and hybrid models in characterising compressional (V_P) and shear wave velocities (V_S) to address their limited availability in the mining industry. The primary aim is to evaluate the suitability of ML algorithms for creating predictive models for spatially variable data, specifically V_P and V_S, while exploring alternative or assisting approaches to enhance prediction accuracy.

The study utilizes an extensive dataset from three different mines in Australia, incorporating geophysical logging and laboratory measurements from multiple boreholes. A systematic workflow is developed to assess the impact of spatial variation in training geophysical logs data on the accuracy of ML models in predicting formation properties, such V_P. The workflow involves clustering analysis using the K-mean algorithm to identify homogeneous clusters based on density, gamma ray and depth. Subsequently, 300 Least Squares Support Vector Regression (LSSVR) models are generated based on grouping the dataset into smaller and larger training sets, with evaluation using separate blind boreholes. The results of the case study suggest prioritizing the incorporation of data from various boreholes over collecting more data points from a limited number of boreholes, emphasizing the high lateral variability in the study area. This observation leads to a simpler yet more accurate predictive model, reducing computational costs.

While ML models excel at capturing non-linear correlations among parameters, incorporating the impact of spatial variability remains challenging. To address this, a hybrid approach is proposed, integrating the ML model (LSSVR) into kriging interpolation to estimate sonic wave velocity with high spatial variability over a geological setting. The methodology involves constructing an initial ML model using LSSVR, calculating residuals, and applying kriging interpolation to optimize the model. Results demonstrate the hybrid model's superiority, highlighting its effectiveness in capturing spatial variability in petrophysical data over geological formations.

To further address the unavailability of V_S geophysical logs in Australian mines, a methodology is proposed to convert ultrasonic laboratory measurements to sonic field measurements using a combination of LSSVR and squirt flow models. The developed model predicts the trend of sonic shear wave velocity (SV_S) using downhole geophysical input, and a frequency dispersion physical model converts the trend to actual values. The study confirms

the methodology's accuracy and robustness, showcasing the potential of AI and physics integration for obtaining precise sonic V_S when laboratory measurements are available.

Publications during enrolment

Journal articles:

1) Mohammadpour, M., Roshan, H., Arashpour, M., & Masoumi, H. (2023). Effect of spatial variability of downhole geophysical logs on machine learning exercises. *International Journal of Coal Geology*, 277, 104333.
2) Mohammadpour, M., Roshan, H., Arashpour, M., & Masoumi, H. (2024). Machine learning assisted kriging to capture spatial variability in petrophysical property modelling. *Marin Petroleum and Geology*,167,106967.
3) Mohammadpour, M., Roshan, H., Arashpour, M., & Masoumi, H. (2024). Ultrasonic to sonic shear wave velocity conversion using an integrated machine learning and physics. *International Journal of Rock Mechanics and Mining Sciences (Under Review)*

Conference papers/presentations:

1) Mohammadpour., Arashpour, Roshan, and Masoumi (2021) "Compressional Wave Velocity Estimation Using Gaussian Processes Regression", AESC 2021.
2) Mohammadpour, Arashpour, Roshan, and Masoumi (2022), "The use of ultrasonic data for shear wave sonic log prediction based on the machine learning technique", *International Geomechanics Symposium* 2022.

Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes (*2)* original papers published in peer reviewed journals and (*1)* submitted publications. The core theme of the thesis is (*Thesis by publication*). The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the (*Civil Engineering Department*) under the supervision of (*A/Prof Hossein Masoumi*).

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.)

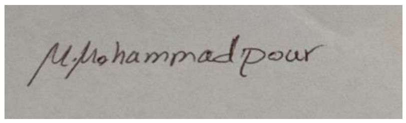
In the case of *chapters 3, 4 and 5,* my contribution to the work involved the following:

Thesis Chapter	**Publication Title**	**Status** *(published, in press, accepted or returned for revision, submitted)*	**Nature and % of student contribution**	**Co-author name(s) Nature and % of Co-author's contribution***	**Co-author(s), Monash student Y/N***
Chapter 3	*Effect of spatial variability of downhole geophysical logs on machine learning exercises*	*Published*	*80%. Concept and collecting data and writing first draft*	*1) Hossein Masoumi, input into manuscript 10%* *2)Hamid Roshan, input into manuscript 5%* *3)Mehrdad Arashpour, input to manuscript 5%*	*No* *No*
Chapter 4	*Machine learning assisted kriging to capture spatial variability in petrophysical property modelling*	*Published*	*80%. Concept and collecting data and writing first draft*	*1)Hossein Masoumi, input into manuscript 10%* *2)Hamid Roshan, input into manuscript 5%* *3)Mehrdad Arashpour, input to manuscript 5%*	*No* *No*
Chapter 5	*Ultrasonic to sonic shear wave velocity conversion using an integrated machine learning and physics*	*Under review*	*80%. Concept and collecting data and writing first draft*	*1) Hossein Masoumi, input into manuscript 10%* *2)Hamid Roshan, input into manuscript 5%* *Mehrdad Arashpour, input to manuscript 5%*	*No* *No*

**If no co-authors, leave fields blank*

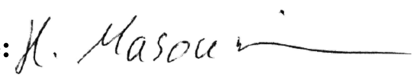
I have renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: Mobarakeh Mohammadpour

Student signature:  **Date: 19 June 2024**

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor name: Hossein Masoumi

Main Supervisor signature:  **Date:** 19/06/2024

Acknowledgements

I sincerely wish to express my profound gratitude to my supervisors A/Prof. Hossein Masoumi, A/Prof. Hamid Roshan and A/Prof. Mehrdad Arashpour for their patience and guidance during my doctoral degree program at the Monash University. I have benefitted from their expertise in the field of rock mechanics, geophysics, and data-driven problem-solving techniques, which have undoubtedly shaped my academic journey.

I would also like to appreciate the Faculty of Engineering of Monash University and the Australian Coal Association Research Program (ACARP) for their financial support which facilitated the successful completion of my research.

My heartfelt appreciation extends to my beloved parents and sisters for their constant emotional support and words of encouragement throughout the program. They have been a source of great inspiration and motivation, contributing significantly to my achievements. Lastly, I want to express my deepest gratitude to my husband. You have been my biggest believer and I cannot thankyou enough for standing by me throughout this challenging journey.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial intelligence
ML	Machine learning
RNN	Recurrent neural network
LSSVM	Least squared support vector machines
LSSVR	Least squared support vector Regression
PSO	Particle swarm optimisation
IQR	Interquartile range
RMSE	Root mean squared error
GA	Genetic algorithm
COA	Cuckoo optimization algorithm
KNN	K-nearest neighbours
MLR	Multivariable linear regression
SVR	Support vector regression
ANFIS	Adaptive neuro-fuzzy inference system
RNN	Recurrent Neural Network
ANN	Artificial Neural Network
ELM	Extreme Learning Machine
UCS	Uniaxial compressive strength
ISRM	International Society of Rock Mechanics

Chapter 1

1. Introduction

This introductory chapter sets out the motivation for the research encompassing aims and objectives. Additionally, the structure of the thesis is outlined, providing readers with a clear and detailed roadmap for navigating through the contents of subsequent chapters.

1.1. Background

Geological formations containing mineral and petroleum resources are often found deep underground, experiencing high pressures and temperatures (Santarelli *et al*., 1989; Song et al., 2013). Effective exploration and management of these resources require predicting the properties of these formations in advance (den Brok *et al*., 1997). However, directly investigating these formations can be challenging due to their inaccessible locations deep underground and the associated costs. Additionally, unconsolidated formations, which are common in some areas, are prone to easily crumbling during testing (Mese and Tutuncu, 1997). Laboratory tests aimed at determining formation properties through the stress-strain relationship are termed static formation properties. To establish this relationship, the formation must undergo deformation beyond its elastic region (Fjær, 2009; Harouaka *et al*., 1995; King and Jing, 2001; Ravazzoli *et al*., 2003). However, such procedures are destructive and therefore expensive, often resulting in the loss of numerous samples during the process.

An alternative to the traditional static method for deriving deformation properties is through dynamic estimation of geomechanical properties. Dynamic formation properties are derived from measurements of sonic velocity or transit time within the formation (Fjær, 2009). It's important to note that static and dynamic estimations of geomechanical properties may yield different results for the same property within the same formation (Fjær, 2009; Holt *et al*., 2012; Mockovčiaková and Pandula, 2003; Svitek and Republic, 2014). This discrepancy is often attributed to factors such as microcracks, porosity, void spaces, and the presence of saturating fluid (Onalo *et al*., 2018b), and it tends to be more pronounced in unconsolidated and weaker formations. Dynamic tests offer advantages over static tests in terms of cost and non-destructiveness. Unlike static tests, dynamic tests do not require the specimen to be destroyed during each evaluation, reducing the need for new samples. This means that fewer formation samples can be used for multiple evaluations. However, obtaining accurate and reliable acoustic measurements for dynamic formation properties can be challenging, as acoustic measurements may not be available in all geological formations during exploration and development (Mullen et al., 2007; Schön, 2015). This raises the critical question: How can we obtain the precise data necessary for assessing dynamic geomechanical properties? Intelligent systems have proven invaluable in various engineering sectors for tackling intricate non-linear challenges, thanks to their remarkable precision and accuracy (Rajabi et al., 2010). Unlike traditional empirical and analytical methods, intelligent systems have excelled in scenarios where others have struggled (Ibrahim and Potter, 2004; Rajabi and Tingay, 2013; Rezaee *et al* et al., 2007; Sbiga and Potter, 2017; Zendehboudi *et al*., 2012). In this study, our aim is to explore and assess suitable techniques for generating reliable data to estimate missing geophysical measurements. These estimates are crucial for predicting the dynamic geomechanical properties of geological formations.

1.2. Problem Statement

V_P and V_S play a crucial role in comprehending subsurface formations, offering insights into lithology, mechanical properties, permeability, and porosity studies (Lv *et al*., 2022; Nabih *et al*., 2022; Riedel *et al*., 2014). These parameters have diverse applications in geophysics, rock mechanics, and fluid flow. By combining V_S and V_P with density, it becomes possible to predict the dynamic elastic properties of rocks and discern lithological characteristics.

The measurement of wave velocities in geological formations is commonly achieved through three main techniques: ultrasonic laboratory testing, borehole sonic logging, and seismic surveys. Sonic and seismic measurements are typically conducted at the field scale, while ultrasonic measurement takes place in a controlled laboratory environment. Ultrasonic wave velocity, a fundamental aspect of rock mechanical testing program, is particularly prevalent in mining projects due to its non-destructive nature and cost-effectiveness (Chawre, 2018). Despite its widespread use, the measurement of sonic shear wave and sonic compressional wave velocities is notably sparse, with the latter not consistently measured in all boreholes within mining operations (Hatherly, 2013; Hatherly *et al*., 2016; Karacan, 2009).

Sonic wave velocities are key physical properties of rocks and are reasonably assumed to be related to other physical properties. This assumption is supported by numerous empirical studies and rock physics research documented in the literature. Various petrophysical properties of rocks are often estimated through laboratory or field geophysical measurements in subsurface areas, and these relationships are explored in this study (Karacan, 2009; Rajabi *et al*., 2010; Asoodeh and Bagheripour, 2012; Onalo *et al*., 2018). By combining petrophysical properties and sonic wave velocities, it is possible to estimate stability parameters, rock mass characteristics, and other factors critical to geotechnical studies and mining engineering.

Exploring correlations between sonic wave velocities and other petrophysical properties could provide a more comprehensive understanding of rock behavior. These insights are crucial for accurately assessing the stability and integrity of rock formations, which are essential for safe and efficient mining operations. The integration of sonic wave velocities with petrophysical data could lead to improved methods for predicting rock mass behavior and stability, enhancing the overall safety and effectiveness of mining engineering practices (Hatherly *et al*., 2016).

Recognizing the significance and practicality of sonic measurements, numerous empirical and non-empirical models, including Machine Learning (ML) and deterministic approaches, have been devised to predict these properties at locations where measurements are unavailable. ML models, in particular, have gained considerable attention in recent years due to their proficiency in capturing both linear and non-linear correlations among various parameters (Rajabi *et al*., 2010; Asoodeh and Bagheripour, 2012; Onalo *et al*., 2018; Anemangely *et al*., 2019; Mehrgini *et al*., 2019; Wang and Peng, 2019; Wang *et al*., 2020; Zhang Y *et al*., 2020; Ebrahimi *et al*., 2022). However, a critical consideration in employing ML algorithms for predicting petrophysical data such as V_P and V_S wave velocities is the need to evaluate assumptions regarding their applicability and usefulness for specific datasets. ML-based models, while powerful, often lack the inherent capacity to capture spatial variabilities and may overlook the autocorrelation between data points (Hashemi ., 2016). This limitation becomes apparent when dealing with petrophysical properties like density, porosity, or wave velocity, where spatial variability is inherent. It is crucial to note that ML techniques, designed with the assumption of random variability, may fall short in adequately capturing the nuanced spatial relationships among petrophysical properties. The typical development of ML algorithms relies on identifying trends through clustering exercises, which might not be suitable for datasets characterized by autocorrelation, especially those exhibiting spatially variable petrophysical properties. This inherent autocorrelation challenges the conventional ML approach, requiring careful consideration and adaptation in the development of accurate predictive models for such geological datasets.

Another critical consideration stems from the severe scarcity of V_S data in Australian mining operations. It is not uncommon to find mines without a single borehole featuring a shear wave velocity log. This poses a significant challenge as the absence of V_S measurement makes it impractical to estimate V_S using logging measurements from other boreholes. This unique predicament, specific to Australian mining, may not have received adequate attention in the literature, particularly since many studies originate from diverse engineering disciplines, such as petroleum engineering, where logging, especially V_S, is more routinely conducted in boreholes. Given the scarcity of Vs in this context, it becomes imperative to explore alternative sources of data to identify potential conversions between related properties. The need for such an investigation arises from the distinct nature of mining operations in Australia, where the absence of V_S logs in boreholes necessitates a workflow to derive valuable insights into subsurface properties. This underscores the importance of adapting methodologies from various disciplines and exploring unconventional sources to address the unique challenges posed by the unavailability of s V_S hear wave velocity data in the Australian mining context.

This research is dedicated to the exploration and evaluation of various techniques and data sources for estimating sonic wave velocities. The primary goal is a comprehensive examination of the data and algorithms utilized in the development of predictive models for V_P and V_S, assessing their effectiveness and accuracy in generating realistic predictions for geological formations. The study critically examines the standard workflow commonly employed in constructing ML-based predictive models for spatially variable data, focusing specifically on V_P and V_S. Special attention is given to the impact of spatial variability, especially in geological settings characterized by high heterogeneity and significant differences in geophysical properties over short distances. Furthermore, the research delves into the limitations of ML algorithms in capturing autocorrelations and spatial variabilities. This is juxtaposed with alternative approaches, including deterministic geostatistical interpolation methods and hybrid techniques like regression kriging. Research aim and objectives

The main aim of this research is to provide valuable insights into the feasibility and effectiveness of employing ML algorithms for predictive modelling of V_P-V_S in the field of petrophysical data. Through a comprehensive investigation of various methodologies, the study seeks to contribute to the understanding of the practical applications and limitations associated with ML-based approaches in the context of sonic wave velocity estimation.

The broad aim of this thesis is to scrutinize and enhance the workflows for development of predictive models of petrophysical properties specifically V_P and V_S, and is supported by the three objectives below:

- *Assessing the development of ML models on large datasets focusing on evaluation of spatial changes*

  While many ML models have been developed in the literature (using different algorithms, datasets and approaches) for estimation of petrophysical properties, they have not examined the training and testing data selection and their impact on capturing the true variability among the variables. This part aims to evaluate the strength of a general ML algorithm in capturing spatial variability, while the number of training data (number of boreholes) changes. This section involves a systematic approach for developing predictive models which include clustering, and then division of the data, generation of 300 datasets and predictive models.

- *Development of a novel hybrid approach for integrating ML and deterministic models to estimate petrophysical properties*

  This part introduces the development of a novel approach to inform ML models about the spatial changes in the data by incorporating kriging interpolation. The new methodology uses kriging interpolation to gradually modify the training data set (input of the ML model) resulting in optimisation of predictive model's accuracy. To evaluate the performance of the proposed hybrid model, a dataset comprising geophysical downhole logging measurements extracted from 122 boreholes was used. The results demonstrate that the hybrid model outperforms the other models, highlighting the effectiveness of employing a hybrid approach in capturing the spatial variability in petrophysical data over geological formations.

- *Establishing a physics informed ML framework to convert laboratory ultrasonic data to field sonic logs*

  This section aims to address the shortage of V_S logging data in mining operations, where no or very limited V_S log can be found. The methodology involves training a predictive model using the Least Squares Support Vector Regression (LSSVR) on data from 178 core samples. The developed model is utilised to predict the trend of sonic shear wave velocity (SV_S) using the same downhole geophysical input as those of laboratory measurements. Once the sonic shear wave velocity trend is obtained, a frequency dispersion physical model (squirt flow) is used to convert the trend to actual values.

Additionally, all the workflows developed to estimate wave velocities are expandable for use in estimating other rock properties or can be adapted for different geographic locations. These workflows aim to highlight the often overlooked role of geophysical measurements in the mining industry. Geophysical measurements can be utilized in various geotechnical and stability studies, providing valuable insights that enhance the accuracy and safety of mining operations. By extending these workflows, researchers and engineers can apply similar methodologies to a broader range of rock properties, facilitating comprehensive geological assessments. This adaptability ensures that the methodologies remain relevant and useful across different geological settings, thereby supporting diverse mining activities. The integration of geophysical measurements into mining studies underscores their potential to significantly contribute to the field, offering more robust and reliable data for making informed decisions in geotechnical and stability assessments.

1.3. Research outputs

The objectives of this research have been progressively achieved primarily through accomplishing extensive works. The scientific contribution of this research is a thesis that includes the following works thus far:

Mohammadpour, M., Roshan, H., Arashpour, M., & Masoumi, H. (2023). Effect of spatial variability of downhole geophysical logs on machine learning exercises. *International Journal of Coal Geology*, 277, 104333.

Mohammadpour, M., Roshan, H., Arashpour, M., & Masoumi, H (2024). Ultrasonic to sonic shear wave velocity conversion using an integrated machine learning and physics. *International Journal of Rock Mechanics and Mining Science (Under Review)*

Mohammadpour, M., Roshan, H., Arashpour, M., & Masoumi, H (2024). Machine learning assisted kriging to capture spatial variability in petrophysical property modelling. Marine Petroleum and geology *(Revised)*

Mohammadpour, Mobarakeh, Arashpour, , Roshan, and Masoumi (2021), , “Compressional Wave Velocity Estimation Using Gaussian Processes Regression”, AESC 2021.

Mohammadpour, Mobarakeh, Arashpour, , Roshan, and Masoumi (2022), , “The use of ultrasonic data for shear wave sonic log prediction based on the machine learning technique”, *International Geomechanics Symposium* 2022.

1.4. Thesis Structure

This research comprises seven chapters with the main content of each chapter summarized below:

Chapter 1 Introduction

This chapter provides a comprehensive overview of the research background, summarizes the problem statement, discusses the motivations and objectives of the study, and outlines the structure of the thesis along with the associated outputs.

Chapter 2 Literature Review

This chapter provides a comprehensive overview of past research focused on constructing predictive models for compressional and shear wave velocities. It primarily delves into the exploration of Machine Learning (ML) and empirical models, which have emerged as the dominant methods in existing literature. Furthermore, the chapter discusses essential geophysical measurements and well-established rock physics principles, along with various ML and hybrid models utilized in prior studies. Moreover, it delves into specific challenges encountered within the context of Australian mining operations, particularly emphasizing limitations associated with restricted sampling and data collection. These constraints pose significant hurdles in developing accurate predictive models tailored to the unique geological characteristics of Australian mining sites

Chapter 3 Effect of Spatial Variability of Downhole Geophysical Logs on Machine Learning exercises

This chapter is dedicated to evaluating the effectiveness of Machine Learning (ML) models utilized in existing literature for estimating crucial geophysical measurements within Australian mines. This chapter aims to address a significant gap identified during the review process. One of the key challenges identified is the suitability of ML models for handling actual field data, which typically involves numerous boreholes across a mining site. This chapter provides a comprehensive guide on utilizing ML for regression and clustering exercises within the mining domain, especially focusing on managing big data sets. It offers detailed insights and strategies for effectively applying ML techniques to navigate the complexities of mining data, ultimately enhancing the accuracy and reliability of predictive models in real-world mining operations. In Chapter 3, a systematic analysis is conducted on a mine site in Bowen Basin, Queensland, Australia, featuring 50 boreholes with downhole logging data, including density, gamma ray, depth, and V_P. Due to the lack of V_P measurements in approximately 50%of boreholes, the development of a predictive model was deemed highly beneficial. Five boreholes were randomly selected for independent testing, with the remaining 45 boreholes used for training and model development. The focus was not solely on achieving the highest prediction accuracy but on optimizing data usage to capture spatial variability for field-scale estimation. The LSSVR algorithm was chosen for model development based on its success for estimating sonic wave velocities in previous literature. The study mainly encountered sandstone/silty sandstone, siltstone, and coal lithologies in the data. Clustering analysis using the K-mean algorithm indicated no distinct clusters in the data. To efficiently capture geological variation while managing computational costs, the data from 45 boreholes were systematically divided into six datasets with varying numbers of randomly selected boreholes. Random selections were repeated five times for each group, resulting in 300 separate datasets with different training set sizes. The performance of each V_P predictive model was assessed using the five boreholes set aside for testing, and the results were plotted in a heatmap to identify the most optimized combinations. This comprehensive analysis aimed to reveal the impact of spatial variability on ML-based predictive models and determine the amount of data needed for an optimized model with low computational costs. The proposed workflow holds potential for enhancing the accuracy and robustness of machine learning models in geological and geophysical applications at a field scale.

Chapter 4 Machine Learning Assisted Kriging to Capture Spatial Variability in Petrophysical Property Modelling

This chapter introduces an innovative workflow that leverages both traditional geostatistical methods and modern Machine Learning (ML) algorithms for estimating geological formation properties. This new model aims to tackle the challenges identified in previous chapters regarding the ability of ML algorithms to capture spatial variability effectively. A general algorithm is proposed to integrate geostatistical models with ML algorithms, offering flexibility in the choice of ML algorithm used in the workflow. However, for the purpose of demonstrating the accuracy of the proposed workflow, the combination of LSSVR with kriging is specifically chosen for a case study. By combining the strengths of geostatistical methods and ML algorithms, this novel workflow offers a promising approach to improve the estimation of geological formation properties, addressing the limitations encountered with ML algorithms in capturing spatial variability. The case study using LSSVR and kriging serves as a practical demonstration of the effectiveness of this approach in achieving accurate results.

In Chapter 4, the initial datasets are expanded, utilized in Chapter 3, to include information from 122 boreholes. From these boreholes, the geophysical downhole logging measurements, encompassing density, gamma ray, depth, and V_P were extracted. Adhering to a standard machine learning approach, 80% percent of the data were designated for training, with the remaining 20% earmarked for testing. The iterative modelling process involved splitting the training data into two sets in each iteration. The first set, constituting 80 percent of the training data, was employed to develop the LSSVR model, capturing correlations with gamma ray, density, depth, coordinates, and the target parameter (sonic V_P). Subsequently, residuals were computed, and kriging interpolation was executed. In the following step, residuals were estimated for the remaining 20 percent of the training data using kriging interpolation. These discrepancies were then utilized to adjust the actual sonic measurements at validation points. The updated validation set and the original training set were amalgamated for a new iteration. This iterative process is crafted to enhance the accuracy of both LSSVR and hybrid models by incorporating considerations for spatial variability in the training data.

Chapter 5 Ultrasonic to Sonic Shear Wave Velocity Conversion using an Integrated Machine Learning and Physics

This chapter serves as a complementary piece to chapters 3 and 4, focusing on an alternative application of ML algorithms for estimating sonic velocities. Unlike the previous chapters, which relied on downhole logs, this chapter utilizes available laboratory measurements. A new methodology is introduced to convert laboratory data to downhole measurements. This methodology involves a combination of ML algorithms and rock physics models. By employing this workflow, ultrasonic data can be effectively converted into logging data. This approach is particularly practical within the Australian mining industry, where ultrasonic data are commonly available while downhole logging data are often scarce. By leveraging laboratory measurements and innovative data conversion techniques, this chapter offers a valuable contribution to the field, providing a viable solution for estimating sonic velocities in mining operations where traditional logging data may be limited. In Chapter 5, a novel methodology that combines Artificial Intelligence (AI) and physical equations is introduced to predict both the trend and magnitude of V_S. This is particularly important because shear log laboratory measurements are often more readily available in mining operations compared to running field shear logs.

The methodology begins by training a predictive model using LSSVR Regression, incorporating data from 178 core samples. The inputs for the model include ultrasonic compressional wave velocity (UV_P), porosity, and density, while the output is ultrasonic shear wave velocity (UV_S). Subsequently, such a developed model is applied to predict the trend of sonic shear wave velocity (SV_S) using the same downhole geophysical inputs as those used in the laboratory measurements. Once the trend of s SV_S is obtained, a frequency dispersion physical model is employed to convert this trend into actual values. The results obtained are then compared with available SV_S data to assess the accuracy and robustness of the proposed methodology. Notably, this methodology demonstrates a good agreement between predicted and measured velocities, showcasing its effectiveness in predicting shear wave sonic velocity in mining operations.

Chapter 6 Estimation of Rock Elastic Properties Using the Predictive Models (application of the predictive models)

This chapter explores one of the applications of the developed models for estimating geomechanical properties. While the models developed in this study (chapters 3, 4, and 5) were primarily designed to directly estimate either V_S or V_P they can also be applied to estimate other crucial geomechanical properties.These properties include Young's modulus, Poisson's ratio, and uniaxial compressive strength, which are essential parameters for understanding the mechanical behavior of geological formations. By leveraging the models developed in this study, it becomes possible to estimate these important geomechanical properties indirectly, utilizing the relationships between V_S, V_P, and other geomechanical parameters established in the literature. This application extends the utility of the developed models, offering valuable insights into the mechanical characteristics of geological formations and enhancing their practical relevance in geomechanical assessments and engineering applications.In this chapter, the exploration takes a brief yet focused approach to one of the practical applications of V_P and V_S wave predictive models: the estimation of profiles for Young's modulus and Poisson's ratio. The predictive models developed in earlier chapters serve as key tools in validating the accuracy of V_P and V_S estimations. Subsequently, these estimations are juxtaposed with laboratory measurements derived from core samples.

Chapter 7 Conclusions and Recommendations

Finally, in the conclusion, the results of this study are summarized and describes some of the broader implications and applications of this work. This chapter also includes suggestions for future research in the same area, aiming to further enhance the understanding and application of the developed models in mining engineering.

1.5. References

Asoodeh M, Bagheripour P (2012) Prediction of compressional, shear, and stoneley wave velocities from conventional well log data using a committee machine with intelligent systems. Rock Mech Rock Eng 45(1):45-63

Anemangely M, Ramezanzadeh A, Amiri H, Hoseinpour S-A (2019) Machine learning technique for the prediction of shear wave velocity using petrophysical logs. J Petrol Sci Eng 174:306-327

Chawre B (2018) Correlations between ultrasonic pulse wave velocities and rock properties of quartz-mica schist. J Rock Mech Geotech Eng 10(3):594-602.

den Brok, S.W.J., David, C., Bernabé, Y (1997). Preparation of synthetic sandstones with variable cementation for studying the physical properties of granular rocks. Comptes Rendus.

Ebrahimi, A., Izadpanahi, A., Ebrahimi, P., & Ranjbar, A. (2022). Estimation of shear wave velocity in an Iranian oil reservoir using machine learning methods. J Petrol Sci Eng, 209, 109841.

Fjær, E (2009). Static and dynamic moduli of a weak sandstone. GEOPHYSICS 74, WA103-WA112.

Harouaka, A., Mtawaa, B., Abdulraheem, A., Klimentos, T (1995). Multistage triaxial testing of actual reservoir cores under simulated reservoir. SCA Conf. 1–10.

Hashemi, M., & Karimi, H. A. (2020). Weighted machine learning for spatial-temporal data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 3066-3082.

Hatherly P, 2013. Overview on the application of geophysics in coal mining. International Journal of Coal Geology, 114, pp.74-84.

Hatherly P, Medhurst T, Zhou B (2016) Geotechnical evaluation of coal deposits based on the Geophysical Strata Rating. Int J Coal Geol 163:72-86

Holt, R. M., Brignoli, M., Kenter, C. J (2000). Core quality: Quantification of coring-induced rock alteration. Int. J. Rock Mech. Min. Sci. 37, 889–907.

Karacan C Ö (2009) Elastic and shear moduli of coal measure rocks derived from basic well logs using fractal statistics and radial basis functions. Rock Mech Rock Eng 46(8):1281-1295

Lv A, Cheng L, Aghighi M A, Masoumi H, Roshan H (2021) A novel workflow based on physics-informed machine learning to determine the permeability profile of fractured coal seams using downhole geophysical logs. Mar Pet Geol 131:10517

Mehrgini B, Izadi H, Memarian H (2019) Shear wave velocity prediction using Elman artificial neural network. Carbonates and Evaporites 34(4):1281-1291

Mese, A.I., Tutuncu, A.N (1997). An investigation of mechanical, acoustic and failure properties in unconsolidated sands. Int. J. Rock Mech. Min. Sci. 34, 314.e1-314.e9.

McMahonSean, ParnellJohn, J.F., B (2016). Evidence for Seismogenic Hydrogen Gas, a Potential Microbial Energy Source on Earth and Mars.

Mullen, M.J., Roundtree, R., Turk, G.A (2007). A Composite Determination of Mechanical Rock Properties for Stimulation Design (What to Do When You Don't Have a Sonic Log). Rocky Mt. Oil Gas Technol. Symp.

Nabih M, Ghoneimi A, Bakry A, Chelloug S.A, Al-Betar M.A, Abd Elaziz M (2022). Rock physics analysis from predicted Poisson's ratio using RVFL based on Wild Geese Algorithm in scarab gas field in WDDM concession, Egypt. Mar Pet Geol 105949.

Onalo, D., Adedigba, S., Khan, F., James, L.A., Butt, S.D (2018a). Data Driven Model for Sonic Well Log Prediction. J. Pet. Sci. Eng.

Onalo, D., Oloruntobi, O., Adedigba, S., Khan, F., James, L., Butt, S (2018b). Static Young's modulus model for drilling operation planning. J. Pet. Sci. Eng.

Rajabi M, Bohloli B, Ahangar E G (2010) Intelligent approaches for prediction of compressional, shear and Stoneley wave velocities from conventional well log data: A case study from the Sarvak carbonate reservoir in the Abadan Plain (Southwestern Iran). Comput Geoci 36(5):647-664

Ravazzoli, C.L., Santos, J.E., Carcione, J.M (2003). Acoustic and mechanical response of reservoir rocks under variable saturation and effective pressure. J. Acoust. Soc. Am. 113,
Riedel M, Goldberg D, Guerin G (2014). Compressional and shear-wave velocities from gas hydrate bearing sediments: Examples from the India and Cascadia margins as well as Arctic permafrost regions. Mar Pet Geol 58:292-320.

Schön, J. (2015). Basic Well Logging and Formation Evaluation Basic Well Logging and Formation Evaluation 4 Contents, 1st ed.

Santarelli, F.J., Detienne, J.L., Zundel, J.P (1989). Determination of the Mechanical Properties of Deep Reservoir Sandstones to Assess the Likelyhood of Sand Production. ISRM Int. Symp.

Song, R., Yue-ming, B., Jing-Peng, Z., De-yi, J., Chun-he, Y (2013). Experimental investigation of the fatigue properties of salt rock. Int. J. Rock Mech. Min. Sci. 64, 68–72.

Svitek, T., Republic, C (2014). Laboratory approach to the study of rock dynamic and static elastic anisotropy under high hydrostatic pressure 1–6.

Zhang Y, Zhong H-R, Wu Z-Y, Zhou H, Ma Q-Y (2020) Improvement of petrophysical workflow for shear wave velocity prediction based on machine learning methods for complex carbonate reservoirs. J Petrol Sci Eng 192:107234

Wang J, Cao J, Yuan S (2020) Shear wave velocity prediction based on adaptive particle swarm optimization optimized recurrent neural network. J Petrol Sci Eng 194:107466

Wang P, Peng S (2019) On a new method of estimating shear wave velocity from conventional well logs. J Petrol Sci Eng 180:105-123

Chapter 2

2. Literature Review

This chapter reviews published literature on predicting models for shear and compressional wave velocities along with their engineering applications associated with the scope of this research.

2.1. Introduction

Ultrasonic laboratory testing, borehole sonic logging, and seismic survey are three main techniques commonly used to measure wave velocities across different scales. While sonic and seismic measurements are performed at the field-scale, ultrasonic measurement is limited to laboratory environment. In downhole sonic logging, a short, high amplitude wave pulse is emitted from the transmitter of a downhole tool (e.g. dipole to measure both compressional and shear wave responses) which travels in drilling fluid and then is refracted to formation which is later received by the receiver/s tool. The wave travel time (slowness) and velocity are then calculated (Serra 1983; Kennedy 2015).

Despite the importance of V_P and V_S, downhole geophysical logging, especially sonic tools are not regularly performed in every exploration borehole of mining operation (Hatherly, 2013; Hatherly ., 2016; Karacan, 2009). As a result, several models have been developed in the literature for estimating these properties. While the models mainly developed empirically or using ML algorithms, their applicability for capturing spatial variation has not been assessed thoroughly. In this chapter, an overview of the primary geophysical logging measurements and the main models developed for estimating both V_P and V_S is provided. This includes a brief discussion of the modelling procedures and data selection criteria, followed by an evaluation of their performance.

2.2. Geophysical logs

2.2.1 Gamma ray log

Gamma-ray logs are utilized to measure the natural radioactivity level of minerals within a formation. Certain minerals emit gamma rays naturally, and these emissions are detected and measured by the gamma-ray log. This data can help identify various lithologies encountered along the well. The measurements are typically expressed in API units. Shale formations often exhibit high radioactivity levels due to the presence of clay minerals such as thorium, potassium, and uranium, whereas sands typically display lower radioactivity values. When combined with resistivity logs, gamma-ray logs can aid in identifying potential hydrocarbon reservoir formations along the well (Rolon *et al*., 2009).

2.2.2 Bulk density log

The formation bulk density log operates by detecting the scattered gamma rays resulting from the interaction between emitted gamma rays and the formation material (Asquith and Gibson, 2004). This is achieved by deploying a gamma-ray source, such as cesium-137 or cobalt-60, into the formation. The electron density of the formation is determined primarily through high-energy Compton scattering, while the lower-energy photoelectric effect contributes to lithology indication. The tool typically has a shallow depth of investigation. Bulk density measurements are expressed in g/cc units. The bulk density of the formation is influenced by several factors, including lithology, formation fluid type (e.g., saltwater, freshwater, oil, gas), and porosity. Notably, bulk density is essentially a porosity log, meaning it can be utilized to calculate density-derived porosity (Rolon *et al*., 2009).

$$\emptyset_{den} = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_{fl}} \tag{2.1}$$

Density logs serve multiple purposes in formation evaluation. They are employed to identify evaporites, gas-bearing formations, and lithology, as well as to determine the density of hydrocarbons (Khandelwal and Singh, 2010). When combined with neutron logs, density logs can also aid in identifying gas-bearing formations. However, it's important to note that shale formations emit radiation that can potentially interfere with density measurements. To mitigate this issue, in formations containing significant shale content, the density-sonic log plot is utilized to determine shale porosity accurately.

2.2.3 Resistivity log

Resistivity logs gauge the electrical conductivity of formation fluids. Specifically, they measure the resistance of formation fluids to the flow of electric currents. This measurement is expressed in ohm.m units. While water typically exhibits low resistance to electric currents, hydrocarbons show higher resistance. Therefore, resistivity logs play a crucial role in distinguishing between oil-bearing and water-bearing formations. When used in conjunction with gamma-ray logs, they aid in identifying potential hydrocarbon reservoirs. During drilling, the presence of drilling mud can interfere with resistivity readings due to mud invasion into the formation. To address this, logging operations often include measurements of shallow, medium, and deep resistivity logs. This is achieved by adjusting the length of the tool's electrode, allowing for corrections to be made for varying depths (Asquith and Gibson, 2004; Bhatt, 2002; Rolon *et al*., 2009; Schön, 2015)

2.2.4 Compressional wave sonic log

The compressional wave sonic transit log records the time taken for a compressional sonic wave to traverse one foot of the formation along the borehole axis (Tixier *et al*., 1975). This measurement, also known as the slowness of the compressional travel wave, is typically expressed in microseconds per foot (the inverse of velocity). The compressional wave sonic travel time can be captured using a borehole compensated (BHC) sonic tool or a dipole sonic tool, both of which include one or more ultrasonic transmitters and two or more receivers.

2.2.5 Shear wave sonic log

The shear wave sonic transit time log records the duration taken for a shear sonic wave to traverse one foot of the formation. This measurement, also termed as the slowness of the shear travel wave, is typically denoted in microseconds per foot. Unlike the compressional wave sonic transit time log, the shear wave sonic travel time cannot be obtained using a borehole compensated sonic tool. Instead, it necessitates the use of a dipole or array sonic tool, comprising both transmitters and receivers. However, it's worth noting that less than 1% of wells in most fields acquire their sonic logs with a dipole sonic tool, resulting in many offset wells lacking shear wave travel time logs (Mullen *et al*., 2007).

2.3. Rock physics and empirical models

Two commonly employed methodologies for estimating sonic velocities are available in the area of geophysics. The first approach involves predictive models grounded in rock physics principles, while the second one employs empirical correlation-based relationships (Wang *et al*., 2019). Predictive models, rooted in rock physics, leverage the physical attributes of rocks to construct petrophysical models for sonic velocity prediction. This entails studying various rock physics models to calculate the effective elastic parameters of rocks. Factors such as porosity, pore shape, fluid inclusion properties, and matrix mineralogy are typically considered in rock physics modelling (Wang *et al*., 2020).

Rock physics modelling has spawned various physics-based models for estimating sonic velocities (Xu and White, 1995; Sun *et al*., 2008; Zhang *et al*., 2012; Guo and Li, 2015; Darvishpour *et al*., 2019; Zhang *et al*., 2020; Ali *et al*., 2021). Theoretically, these model-based methods are not constrained by specific geographical areas, as they address limitations inherent in empirical equations. However, their application often involves intricate estimation processes due to assumptions associated with pore shape, potentially compromising the validity of results. The matrix of elastic parameters, compositions, mixing modes, pore shapes, and fluid constituents must all be considered in these models to achieve accurate predictions of V_S. The complexity of these models, coupled with the need for numerous considerations, limits their practical appeal for real-world drilling and field development applications.

2.3.1 Compressional wave velocity estimation

The velocity of waves moving parallel to the direction of particle motion in geological formations is likely the most extensively studied aspect of acoustic wave energy. Historically, it was the primary focus in formation evaluation (Picket, 1963). One of the earliest studies on compressional wave velocity was conducted by Wyllie in 1956. Using a sonic pulse velocity measurement technique (transmitter), Wyllie examined how compressional wave velocity varied with changes in formation fluid (brine, oil, gas), salinity, saturation, and temperature (Wyllie et al., 1956). Wyllie established a relationship between wave velocity and porosity through equation (2.2) and accounted for density in equation (2.3). φ=porosity, v= velocity, G=shear modulus and ρ= density, while the subscripts ma and fl indicate attributes of the matrix and fluid respectively

$$.\frac{1}{v} = \frac{\varphi}{v_{ma}} + \frac{1-\varphi}{v_{fl}} \tag{2.2}$$

$$\frac{1+G}{v^2} = \left[\frac{\varphi}{v_{ma}^2} + \frac{(1+G)(1-\varphi)}{v_{ma}^2\,\rho_{fl}}\right] * (\varphi\rho_{ma} + (1-\varphi)\rho_{fl} \tag{2.3}$$

Wyllie's research indicates that wave velocity increases as porosity decreases with depth. In porous media, velocity decreases as fluid saturation diminishes, except when a denser fluid replaces the effluent. The theories of Biot (1941) and Gassmann (1951) suggest that elastic constants, particularly Biot’s coefficient, bulk modulus, and shear modulus, also change with the fluids present in interconnected pores (effective porosity) (Biot, 1941; Gassmann, 1951). Birch expanded on this research in 1960 by transmitting wave energy at pressures up to 12 kilobars in metamorphic and igneous rocks. He found that velocity increases with pressure, though he did not consider sedimentary rocks or the effects of porosity (Birch, 1960). Later, Birch demonstrated that at low pressures, velocity depends on porosity, composition, and anisotropy, whereas at higher pressures, it is more influenced by density and atomic weight. In homogenous or isotropic formations with the same atomic weight, velocity has a linear relationship with porosity (Birch, 1961).

Most elastic constants can be estimated if at least two of the formation's elastic constants are known. Birch suggested these estimations are valid only if tests are performed at the same pressure and temperature. However, above 10 kilobars and 600°C, the effects of temperature and pressure can be neglected. Porosity varies across formations during drilling. Raymer *et al*. (1980) described this variation using the "lack of compaction factor" (Cp), which relates to the apparent porosity term in Wyllie's equation through equation (2.4).

$$\varphi_c = \frac{\varphi}{C_p} \tag{2.4}$$

Tosaya (1982) examined how compressional wave velocity is affected by clay content. He found that clay content has a smaller impact on compressional velocity compared to porosity (Tosaya and Nur, 1982). Tosaya proposed the following equation to describe this relationship:

$$v_p = -2.4C - 8.6\,\varphi + 5.8 \tag{2.5}$$

C is clay content.

Equation (2.6) demonstrates a linear relationship between compressional velocity and clay content in consolidated formations. However, the compressional velocity in unconsolidated sandstone formations deviates from this trend, showing lower values than predicted. Besides the porosity of the formation, the micro porosity of clay increases the void space that can be filled by formation fluids, further reducing the velocity (Kowallis et al., 1984).

$$v_p = -5.7C - 9.2\varphi + 5.6 \tag{2.6}$$

2.3.2 Shear wave velocity

In 1943, Birch conducted research on S-waves at elevated temperatures, which he later referenced in 1960 to highlight the long-standing study of secondary wave velocities. Pickett's work with shear waves is considered a foundational contribution to the use of shear waves in formation evaluation. Pickett noted that earlier studies focused only on the first arrival time of wave energy, neglecting the secondary and subsequent wave arrivals. He believed that these later arrivals could be analysed to establish a meaningful trend.

This research led to the well-known method of estimating porosity from sonic well logs based on compressional waves, allowing for the identification of potential lithological intervals along a well. Additionally, it was observed that fractures in the formations affected the travel time of the wave energy, resulting in a reduction in both velocity and signal amplitude. This effect was more pronounced in shear waves than in compressional waves, further demonstrating the utility of shear waves. Although Pickett did not provide an empirical correlation, the graphs and charts from his laboratory experiments served as benchmarks for identifying trends and slopes in clean sandstone, dolomite, limestone, and limy sandstone (Pickett, 1963).

In Pickett's 1963 study, the focus was on borehole fluid, without considering the formation fluid content or saturation. Nur and Simmons (1969) validated Birch's findings, demonstrating that wave velocities increase at higher pressures. Therefore, laboratory investigations of seismic velocity under ambient conditions tend to underestimate the velocities found in reservoir formations. In saturated formations, the reduction in primary wave velocity is minimal and often assumed negligible for shear velocity. However, Pickett's work suggests otherwise, showing that fluid saturation in pores can significantly affect the amplitude of the arrival wave. A formation can possess both matrix pore porosity and porosity due to cracks or fractures. Nur and Simmons (1969) observed that the velocity of saturated granite was higher in cores with cracks compared to those with only pores. This implies that matrix porosity, being part of the grain structure, contributes to a higher bulk modulus than the overall formation. Similar observations are expected with increased confining pressure. The elastic constants, which are similar in both dry and saturated formations, can be expressed as functions of the velocities (Nur and Simmons, 1969). K is bulk modulus, G shear modulus.

$$K = (v_p^2 - \frac{4}{3} v_s^2)\rho \tag{2.7}$$

$$G = v_s^2 \rho \tag{2.8}$$

Similar to porosity, clay content in a formation affects seismic wave velocity by reducing it (Minear, 1982). Minear noted that this reduction in velocity is more significant in formations with laminated and structural clay, but negligible in formations with dispersed clay. This is likely due to the mobility of dispersed clay, which is randomly oriented and does not adhere to the pore surface, unlike laminated clay that has a strong orientation and contributes to anisotropy in the formation. Dispersed clay remains free to move within the pore fluid. Tosaya (1982) updated equation (2.5) to include shear velocity, presenting these relationships in equations (2.9) and (2.10).

$$v_p = -2.4C - 8.6\varphi + 5.8 \tag{2.9}$$

$$v_s = -2.1C - 6.3\varphi + 3.7 \tag{2.10}$$

In geomechanics, effective pressure is defined as the difference between the pore pressure and the total overburden pressure, assuming the Biot coefficient is unity (Zoback, 2010). As effective pressure increases, so does velocity. Eberhart-Phillips *et al*. (1989) described this relationship with equations (2.11) and (2.12).

$$v_p = -1.73\sqrt{C} - 6.94\varphi + 5.77 + 0.466(P_e - e^{-16.7Pe}) \tag{2.11}$$

$$v_{ps} = -1.57\sqrt{C} - 4.94\,\varphi + 3.7 + 0.361(P_e - e^{-16.7Pe} \tag{2.12}$$

More recently, Ramcharitar & Hosein (2016) proposed the following empirical model which included the density of the formation:

$$\Delta t_p = 19C - 0.0066\varphi + 139Z - 18\rho + 153 \tag{2.13}$$

$$\Delta t_s = 33C - 0.0052\varphi + 159Z + 205\rho + 2.98\Delta t_p \tag{2.14}$$

2.4. Machine learning models

With the new advancement,AI has been widely deployed to estimate sonic wave velocities, particularly V_S (Rezaee *et al*. 2007; Kumar *et al*. 2011; Ceryan *et al*. 2012; Ocak and Seker 2012; Rabbani *et al*. 2012; Waqas and Ahmed 2020; Sun *et al*. 2021; Rahman and Sarkar 2021; Saedi and Mohammadi 2021). A number of investigations have demonstrated the superiority of AI algorithms such as Support Vector Regression (SVR), fuzzy logic, gaussian processes and neural network over the proposed empirical correlations for sonic wave velocity estimation (Rajabi et al. 2010; Asoodeh and Bagheripour 2012; Tan *et al*. 2015; Xu *et al*. 2017; Zhang *et al*. 2018; Onalo *et al*. 2018; Anemangely *et al*. 2019; Mehrgini *et al*. 2019; Wang and Peng 2019). Some of these correlations were developed for a specific lithology such as that proposed by Zhang et al. (2020), which includes a workflow for V_S prediction using recurrent neural network method (RNN) in carbonate reservoir. Zhang *et al*. (2020) demonstrated that the developed AI based correlations perform better in comparison with the petrophysical models for complex carbonate reservoir.

In various scientific fields such as petroleum engineering, rock mechanics, mining, and geological engineering, Artificial Neural Networks (ANN) and Least Squares Support Vector Regression (LSSVR) tools are increasingly favored for developing predictive models. These tools utilize complex predictor variables derived from measured core and log data. However, ANN faces challenges such as slow convergence, limited generalization, and overfitting due to excessive hidden neurons. It also struggles with datasets containing a small number of data points, hindering the development of robust models. In contrast, LSSVR, a supervised machine learning technique, can effectively handle limited data points and is quick to train, resulting in reliable and accurate predictive models with better generalization.

Researchers commonly utilize machine learning tools with various inputs based on available core and log data to estimate parameters like V_S and V_P for geological formations. Rock parameters such as formation gamma-ray, deep resistivity, shale content, bulk density, neutron porosity, compressional slowness, rock porosity, cementation factor, formation depth, and permeability are often employed in data-driven models to predict V_S and V_P. Some researchers have critically reviewed the contributions of others, focusing on research methodologies and limitations in obtaining acoustic wave velocities. Additionally, several scholars have analysed the performance of machine learning-based predictive models for sedimentary formations using data samples.

Furthermore, the impact of different input parameters on the accuracy of sonic wave predictive models has been studied by various researchers (Tan *et al*. 2015; Al-Dousari *et al*. 2016; Behnia *et al*. 2017; Xu et al. 2017; Zhang *et al*. 2018; Anemangely *et al*. 2019). Particularly, the dependency of sonic wave velocities on different petrophysical parameters such as density, porosity, permeability, and clay content has been studied. Behnia *et al*. (2017) evaluated the importance of density, porosity and V_P in predicting V_S using an Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm and reported V_P as the most important parameter. Anemangely *et al*. (2019), on the other hand, concluded that an increase in the number of input parameters can lead to better prediction of V_S using non-dominated sorting genetic algorithm. Anemangely *et al*. (2019) also reported that the predictive models did not exhibit any sensitivity to more than five selected features as the input parameters.

In 2007, Rezaee *et al*. constructed a neural network model using the Levenberge Marquardt training algorithm, employing log variables from clastic sedimentary formations. Their model achieved a high correlation coefficient of 0.94, aligning well with other data-driven models. However, they did not analyse the individual predictor variables used in predicting Vs. In 2014, Maleki *et al.* examined a hybrid connectionist model using log data, comparing it with empirical correlations for carbonate reservoirs. Both ANN-GA and SVM-GA based models showed excellent performance, with correlation coefficients of 0.97 and 0.94, respectively. They concluded that compressional slowness, bulk density, and gamma-ray are crucial input variables for predicting Vs in the carbonate reservoirs of the Iranian oil field.

Similarly, Akhundi *et al*. (2014) explored data-driven models, developing correlations using a multi-variate regression approach. They compared their model's performance with that of a neural network-based model, discovering that neutron porosity, rock bulk density, and compressional wave velocity are essential input variables for predicting Vs in the Asamari formation of the Iranian oil field. Later, Aleardi (2015) presented GA based models and compared those with both neural network and multi-varaiate regression-based models. It was observed that gamma-ray has a minimal effect on Vs prediction, while bulk density and true resistivity are the most vital variables for clastic sedimentary formation with sand-shale sequence layers.

Bagheripour *et al*. (2015) studied the SVR based predictive model for Vs using 2879 data samples collected from the gas well field of Iran. According to their research output, it was concluded that sonic compressional slowness has relative significance with least R^2 of about

0.41. Additionally, they examined the data-driven model performance, and also compared with ANN and other empirical correlations for the carbonate rock formation. Furthermore, Al-Dousari *et al*. (2016) developed a data-driven GRNN-based model with 35 data samples from petrographic analysis and petrophysical rock properties of dolomite, limestone, and siliciclastic formations, and compared it with five empirical models of carbonate rock. It was concluded that rock porosity and grain density are the vital input variables, compared with other variables of clay content, permeability, and cementation exponent to estimate Vs.

In 2017, research by Behnia *et al*. explored the performance of a GEP-based model using 516 data points, with 80% allocated for training and 20% for testing. They found that three rock variables V_P, PHI, and RB are crucial for accurately predicting Vs in limestone formations. However, they did not prioritize these parameters based on their significance in predicting sedimentary rock properties. Later, in 2019, Wang and Peng developed a Vs prediction model using an extreme machine learning-based tool with 516 data samples from the Ordos Basin in China. Their analysis revealed that compressional wave velocity, bulk density, and neutron porosity are the primary input variables for predicting Vs in carbonate sedimentary formations. They claimed that the Extreme Learning Machine (ELM) tool outperformed other models with an R2 of 0.97. However, they did not propose a new correlation for estimating Vs using available core and log data.

In another study in 2019, Bukar *et al*. investigated various predictive models using Gaussian process regression, ensembles of trees, and SVM, utilizing log variables from the Browse Basin offshore northwest Australia. They found that the exponential-based Gaussian process regression had the lowest error compared to linear SVM and boosted trees, as evaluated by root mean square error. Additionally, in 2019, Alkinani *et al*. examined the performance of classic ANN and dynamic RNN predictive models using Vp, RB, and NPHI as predictor variables for carbonate formations. They observed that the RNN model outperformed the neural network model.A common approach in developing the predictive models with machine learning was to split available data into training and testing sets. This method ensures that a portion of the data is reserved for evaluating the model's accuracy in predicting the target parameter/s. This split, typically performed randomly, aims to reduce bias and ensure fairness in the evaluation process. Previous studies, including those by Rezaee *et al*. (2007), Behnia *et al.* (2017), Wang and Peng (2019), and Du *et al*. (2019), utilized this approach to develop predictive models for sonic wave velocity estimation. Geophysical logs, such as density, gamma, neutron porosity, and sonic velocity, are extracted from boreholes, and the data is then divided into training and testing sets for model development. The availability of a substantial amount of data from average downhole logging measurements at short increments (e.g., 1, 5, 10 cm) facilitates both model training and testing phases

The primary purpose of using ML algorithms for developing a V_P predictive model, however is to technically estimate the property in locations where no sonic downhole logging is performed. However, testing the model using data extracted from the same borehole provides no insight into how the model behaves at blind locations where data trend and values are not seen by the model. This can be particularly critical in geological settings with high spatial variability, where significant differences in geophysical properties can be observed even over short distances, especially in the areas where multiple rock types with high heterogeneity exist. A limited number of boreholes cannot represent the true variability of the subsurface, and the model may overfit to the training data, leading to poor general performance. In other words, the models constructed on data from one or two boreholes are always shown to perform extremely good because the model performance is assessed on the borehole data where the model was trained with and not on boreholes where the data was not seen by the model at all. The effect of such spatial variability on ML based models' performance has not been assessed to date leaving a practical gap to conclude that the ML models are in fact more accurate than simple one-dimensional empirical models. The other important point to consider is the computational efficiency of ML based predictive models. If a significant number of boreholes exist in a site, how much of data and how many boreholes are needed to train the best representative model with the lowest computational need.

When employing machine learning algorithms to construct predictive models for petrophysical data like V_P and V_S, it is crucial to assess the assumptions regarding their applicability and usefulness within the specific datasets. ML-based models, given their typical lack of consideration for spatial variabilities, often overlook the autocorrelation between data, as highlighted by Hashemi *et al*. in 2016. Also, spatial variability among petrophysical properties such as density, porosity, or wave velocity cannot be adequately captured by ML techniques considering the data as whole. Most machine learning algorithms in fact are developed based on the assumption of random variability and trends can be only identified through clustering exercises. This approach, however, becomes problematic and inaccurate for trending data with spatially variable petrophysical properties due to their autocorrelation.

2.5. Kriging and geostatistical approaches

Other than ML algorithms, deterministic geostatistical interpolation methods, such as simple kriging, have been widely recognized as suitable approaches for modelling spatial autocorrelations and spatial variable estimation in land resource inventories, remote sensing technologies, soil surveyor mapping, and other fields (Mitas and Mitasova, 1999; Hengl *et al*. 2003).

Kriging algorithms are known as the best linear unbiased estimator and generate robust estimates at unsampled locations (Goovaerts, 1997). However, these methods are unable to extract information from all available measurements, thus cannot establish links between different source measurements (Chiles and Delfiner, 2012). Hybrid interpolation techniques, such as regression kriging, are able to model spatial variability while using supplementary information (Odeh *et al*., 1995; Hengl *et al*., 2007).

Regression kriging uses linear regression to capture the correlation between the target variable and any available auxiliary variable and then employs simple kriging to interpolate the residual from the developed linear regression model. Employing regression kriging and auxiliary data has shown improvement in prediction accuracy compared to sole linear regression or kriging (Odeh *et al*., 1994, 1995; Goovaerts, 1999b; Bishop and McBratney, 2001). For example, using elevation data as an auxiliary variable in temperature mapping with kriging has shown to improve the accuracy of temperature prediction (Hudson and Wackernagel, 1994). Or adding a slope map as an explanatory variable has been found to enhance horizon thickness estimation of formations (Bourennane *et al*. (1996, 2000)). While regression kriging can add extra value to a sole kriging method, it is mostly desirable when a linear correlation between auxiliary and target variable exists. However, in cases of strong non-linear correlation between the input and target variables, these models may struggle to capture and represent the correlation adequately.

2.6. Hybrid approaches

The utilization of ML algorithms has gained significant traction in recent years, prompting numerous studies to explore the integration of diverse ML techniques aimed at enhancing prediction accuracy in various domains (Wang *et* al. 2020; Karimpouli and Fattahi 2018; Du *et al*. 2019). These endeavours have resulted in the development of hybrid models strategically designed to harness the strengths of multiple algorithms, thereby minimizing prediction errors.

An exemplary instance of this approach is illustrated by Anemangely *et al*. (2019), who delved into the enhancement of least square support vector machines (LSSVM) through the implementation of distinct optimization algorithms. Through the incorporation of Particle Swarm Optimization (PSO), Genetic Aalgorithm (GA), and Cuckoo optimization algorithm (COA), Anemangely *et al*. (2019) constructed three distinct models for V_S estimation. Notably, the LSSVM-COA model emerged as the most efficient algorithm, showcasing the potential of hybridization in refining predictive capabilities.

The hybridization of algorithms extends beyond machine learning, with certain members of the kriging family, such as regression kriging, being considered hybrid models. Regression kriging, for instance, amalgamates linear regression and simple kriging techniques. In literature, machine learning algorithms, including Artificial Neural Networks (ANN) and support vector regression, have been employed to mitigate the limitations associated with the linear assumption in regression kriging. This has been evident in applications related to climate data estimation (Demyanov *et al*., 1998), wind speed prediction (Chung *et al*., 2012), precipitation forecasting (Seo *et al*., 2015), and shale volume estimation (Li *et al*., 2020).

Hybrid models that blend machine learning and kriging algorithms aim to use the unique strengths of both approaches. In contrast to regression kriging, which assumes linearity between auxiliary variables and the target parameter, these hybrid models incorporate the robust capabilities of ML to capture non-linear relationships, resulting in heightened prediction accuracy. Typically, the construction of hybrid models involves two primary steps: (1) the training and development of a regression machine learning model while calculating residuals, and (2) the application of residual kriging interpolation. While the second step is somehow contingent on the first, they operate independently, with ML treating the data without explicitly considering spatial variability. This two-step process exemplifies the intricate synergy between ML and kriging techniques in hybrid modelling approaches.

2.7. Conclusion

In this chapter, a comprehensive review of predictive models was conducted, primarily focusing on empirical and ML approaches. Previous studies predominantly emphasized the algorithmic choices and the number of features employed in modelling, with a primary goal of achieving optimal accuracy. However, it is crucial to highlight that the accuracy metrics were often based on a restricted set of boreholes, frequently excluding blind boreholes and incorporating segments of the training data. This practice may introduce inaccuracies, as ML models might struggle to adequately capture spatial variability. The meticulous optimization of the training, validation, and test data is imperative in enhancing the modelling procedure.

Further exploration is warranted to establish an effective workflow for ML modelling. While geostatistical methods have not gained the same popularity as ML algorithms in estimating sonic velocities, it was hypothesised that incorporating measures of similarity, akin to the semivariogram function employed in methods like kriging, could serve as a valuable aid in ML model training.

It is noteworthy that the utilization of petrophysical logs is not widely acknowledged in Australian mining operations, and shear wave velocity measurements are often lacking in many mine sites. Although some laboratory measurements of ultrasonic properties may exist for a limited number of core samples at each site, these data are not extensively documented, and their application as an alternative to sonic measurements is hindered by inherent differences in frequency and tools. Despite these challenges, ultrasonic shear wave velocity data, where available, could serve as a valuable resource in sites where sonic measurements are unavailable. Although ultrasonic data may differ intrinsically from sonic measurements, they offer a potential link to estimating shear wave velocity, providing a practical avenue for overcoming data limitations in certain mining contexts particularly in Australia.

2.8. References

Ali M et al (2021) Machine learning-A novel approach of well logs similarity based on synchronization measures to predict shear sonic logs. J Petrol Sci Eng 203:108602.

Al-Dousari M, Garrouch A A, Al-Omair O (2016) Investigating the dependence of shear wave velocity on petrophysical parameters. J Petrol Sci Eng 146:286-296

Asoodeh M, Bagheripour P (2012) Prediction of compressional, shear, and stoneley wave velocities from conventional well log data using a committee machine with intelligent systems. Rock Mech Rock Eng 45(1):45-63

Anemangely M, Ramezanzadeh A, Amiri H, Hoseinpour S-A (2019) Machine learning technique for the prediction of shear wave velocity using petrophysical logs. J Petrol Sci Eng 174:306-327

Behnia D, Ahangari K, Moeinossadat S R (2017) Modeling of shear wave velocity in limestone by soft computing methods. Int J Min Sci Technol 27(3):423-430

Bishop, T. F. A., and A. B. McBratney. "A comparison of prediction methods for the creation of field-extent soil property maps." *Geoderma* 103, no. 1-2 (2001): 149-160.

Bourennane, Hocine, Dominique King, and Alain Couturier. "Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities." *Geoderma* 97, no. 3-4 (2000): 255-271.

Chung, J. W., & Rogers, J. D. (2012). Interpolations of groundwater table elevation in dissected uplands. Groundwater, 50(4), 598-607.

Chiles, J., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty,Wiley, New York.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, New York.

Ceryan N, Okkan U, Kesimal A (2012) Application of generalized regression neural networks in predicting the unconfined compressive strength of carbonate rocks. Rock Mech Rock Eng 45(6):1055-1072

Darvishpour A et al (2019) Wellbore stability analysis to determine the safe mud weight window for sandstone layers. Pet Explor Dev 46:1031–1038.

Demyanov, V., Kanevsky, M., Chernov, S., Savelieva, E., & Timonin, V. (1998). Neural network residual kriging application for climatic data. *Journal of Geographic Information and Decision Analysis*, *2*(2), 215-232.

Du Q, Yasin Q, Ismail A, Sohail G M (2019) Combining classification and regression for improving shear wave velocity estimation from well logs data. J Petrol Sci Eng.182:106260

Goovaerts, P., 1999a. Geostatistics in soil science: state-of-the-art and perspectives, Geoderma 89 (1– 2), 1 – 45.

Guo Z, Li X-Y (2015) Rock physics model-based prediction of shear wave velocity in the Barnett Shale formation. J Geophys Eng 12:527–534.

Hatherly P, 2013. Overview on the application of geophysics in coal mining. International Journal of Coal Geology, 114, pp.74-84.

Hatherly P, Medhurst T, Zhou B (2016) Geotechnical evaluation of coal deposits based on the Geophysical Strata Rating. Int J Coal Geol 163:72-86

Hudson, Gordon, and Hans Wackernagel. "Mapping temperature using kriging with external drift: theory and an example from Scotland." *International journal of Climatology* 14, no. 1 (1994): 77-91.

Hengl, Tomislav, Gerard BM Heuvelink, and David G. Rossiter. "About regression-kriging: From equations to case studies." Computers & geosciences 33, no. 10 (2007): 1301-1315.

Hengl, Tomislav, David G. Rossiter, and Alfred Stein. "Soil sampling strategies for spatial prediction by correlation with auxiliary maps." Soil Research 41, no. 8 (2003): 1403-1422.

Karacan C Ö (2009) Elastic and shear moduli of coal measure rocks derived from basic well logs using fractal statistics and radial basis functions. Rock Mech Rock Eng 46(8):1281-1295

Kennedy M, 2015. Practical petrophysics, Elsevier.

Kumar B R, Vardhan H, Govindaraj M (2011) Prediction of uniaxial compressive strength, tensile strength and porosity of sedimentary rocks using sound level produced during rotary drilling. Rock Mech Rock Eng 44(5):613-620

Karimpouli S, Fattahi H (2018) Estimation of P-and S-wave impedances using Bayesian inversion and adaptive neuro-fuzzy inference system from a carbonate reservoir in Iran. Neural Comput Appl 29(11):1059-1072

Li, X., Ao, Y., Guo, S., & Zhu, L. (2020). Combining regression kriging with machine learning mapping for spatial variable estimation. IEEE Geoscience and Remote Sensing Letters, 17(1), 27-31.

Mehrgini B, Izadi H, Memarian H (2019) Shear wave velocity prediction using Elman artificial neural network. Carbonates and Evaporites 34(4):1281-1291

Mitas, Lubos, and Helena Mitasova. "Spatial interpolation." Geographical information systems: principles, techniques, management and applications 1, no. 2 (1999).

Onalo D, Adedigba S, Khan F, James L.A, Butt S (2018) Data driven model for sonic well log prediction. J Petrol Sci Eng 170:1022-1037

Odeh, I., McBratney, A., Chittleborough, D., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63 (3– 4), 197– 214.

Odeh, Inakwu OA, A. B. McBratney, and D. J. Chittleborough. "Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging." Geoderma 67, no. 3-4 (1995): 215-226.

Ocak I, Seker S E (2012) Estimation of elastic modulus of intact rocks by artificial neural network. Rock Mech Rock Eng 45(6):1047-1054

Rajabi M, Bohloli B, Ahangar E G (2010) Intelligent approaches for prediction of compressional, shear and Stoneley wave velocities from conventional well log data: A case study from the Sarvak carbonate reservoir in the Abadan Plain (Southwestern Iran). Comput Geoci 36(5):647-664

Rezaee M R, Ilkhchi A K, Barabadi A (2007) Prediction of shear wave velocity from petrophysical data utilizing intelligent systems: An example from a sandstone reservoir of Carnarvon Basin, Australia. J Petrol Sci Eng 55(3-4):201-212

Rahman T, Sarkar K (2021) Lithological Control on the Estimation of Uniaxial Compressive Strength by the P-Wave Velocity Using Supervised and Unsupervised Learning. Rock Mech Rock Eng:1-17

Rabbani E, Sharif F, Koolivand Salooki M,Moradzadeh A (2012) Application of neural network technique for prediction of uniaxial compressive strength using reservoir formation properties. International journal of rock mechanics and mining sciences (1997) 56:100-111

Saedi B,Mohammadi S D (2021) Prediction of Uniaxial Compressive Strength and Elastic Modulus of Migmatites by Microstructural Characteristics Using Artificial Neural Networks. Rock Mech Rock Eng:1-21

Sun, F.-l., et al. (2008). An S-Wave Velocity Predicted Method [J]. Progress in Geophysics 2.

Sun H, Du W, Liu C (2021) Uniaxial compressive strength determination of rocks using X-ray computed tomography and convolutional neural networks. Rock Mech Rock Eng:1-13
Serra O (1983) Fundamentals of well-log interpretation.

Tan M, Peng X, Cao H, Wang S, Yuan Y (2015) Estimation of shear wave velocity from wireline logs in gas-bearing shale. J Petrol Sci Eng 133:352-366

Wang, S.-J., et al. (2014). Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. Neural processing letters 39, 25–43.

Wang, J., et al. (2019). An effective method for shear-wave velocity prediction in sandstones. Marine Geophysical Research 40,655–664

Wang P, Peng S (2019) On a new method of estimating shear wave velocity from conventional well logs. J Petrol Sci Eng 180:105-123

Waqas U, Ahmed M F (2020) Prediction Modeling for the Estimation of Dynamic Elastic Young's Modulus of Thermally Treated Sedimentary Rocks Using Linear–Nonlinear Regression Analysis, Regularization, and ANFIS. Rock Mech Rock Eng 53(12):5411-5428

Wang J, Cao J, Yuan S (2020) Shear wave velocity prediction based on adaptive particle swarm optimization optimized recurrent neural network. J Petrol Sci Eng 194:107466

Xu S, White RE (1995) A new velocity model for clay-sand mixtures 1. Geophys Prospect 43:91–118.
Xu J, Tan M, Wang X, Wu C (2017) Predicting acoustic-wave velocities and fluid sensitivity to elastic properties in fractured carbonate formation. Interpretation 5(1):SB69-SB80

Zhang, G.Z., et al. (2012). A shear velocity estimation method for carbonate rocks based on the improved Xu-White model: Oil Geophysical Prospecting.
Zhang D, Lin J, Peng Q, Wang D, Yang T, Sorooshian S, Liu X,Zhuang J (2018) Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. Journal of Hydrology 565:720-736

Zhang Y, Zhong H-R, Wu Z-Y, Zhou H, Ma Q-Y (2020) Improvement of petrophysical workflow for shear wave velocity prediction based on machine learning methods for complex carbonate reservoirs. J Petrol Sci Eng 192:10723

Chapter 3

3. Effect of spatial variability of downhole geophysical logs on machine learning exercises

3.1. Summary

In Chapter 3 an innovative workflow that assesses the impact of spatial variation in training geophysical logs data on the accuracy of ML models in predicting formation properties e.g. compressional wave velocity was constructed. To build this systematic workflow, a large dataset containing geophysical logs from 50 boreholes in a coal mine in Bowen Basin, QLD, Australia was selected for the analysis. Boreholes mainly intersected geological formations from Rangal coal measure rocks. The workflow consists of i) a clustering analysis using the K-mean algorithm to determine if the data could be partitioned into smaller, more homogeneous clusters based on density, gamma ray and measured depth, ii) putting aside five boreholes for validation (these boreholes are not seen by the models, named test boreholes) and iii) selectin training datasets randomly from the remaining boreholes grouping them into 1, 5, 10, 20, 30 and 40 boreholes.

The content of this chapter has been published in the "International Journal of Coal Geology". The article is available as an open-access publication at:

https://doi.org/10.1016/j.coal.2023.104333

3.2. Introduction

Compressional wave velocity is an important parameter for understanding the properties of subsurface formations such as lithology, mechanical properties, permeability and porosity with numerous applications across geophysics, rock mechanics and fluid flow. Sonic tools such as Dipole (i.e. full wave form sonic) are used in boreholes to measure the formation acoustic properties including compressional and shear sonic velocities. While combination of V_P and V_S alongside with density logs can be used for estimating rock dynamic mechanical properties, downhole geophysical logging especially V_P is not regularly performed in every exploration borehole in mining operations (Hatherly, 2013; Hatherly *et al*., 2016; Karacan, 2009). As a result, numerous empirical correlations have been proposed for sonic waves (V_S and V_P) estimation using porosity, density and clay content of different rock types (Eberhart-Phillips *et al*., 1989; Castagna and Backus, 1993; Khandelwal and Singh, 2009, Abbas *et al*., 2018). While empirical models are simple to use, they do not yield a good accuracy in many instances. Hence, Artificial intelligence (AI) has been recently employed to develop sonic wave velocity predictive models with varying features to make the predictions more accurate (Rezaee *et al*., 2007; Cranganu and Breaban, 2013). A number of investigations have demonstrated the superiority of machine learning algorithms such as SVR), fuzzy logic, gaussian processes and ANN over the empirical correlations for sonic wave velocity estimation (Rajabi et al., 2010; Asoodeh and Bagheripour, 2012; Onalo *et al*., 2018; Anemangely *et al*., 2019; Mehrgini *et al*., 2019; Wang and Peng, 2019; Wang *et al*., 2020; Zhang Y *et al*., 2020; Ebrahimi *et al*., 2022).

The impact of different input parameters on the accuracy of sonic wave predictive models has been investigated by various researchers (Tan *et al*. 2015; Al-Dousari *et al*. 2016; Behnia *et al*. 2017; Xu *et al*. 2017; Zhang *et al*. 2018; Anemangely *et al*. 2019) especially density, porosity, permeability and clay content. For example, Behnia *et al* . (2017) evaluated the importance of density, porosity and V_P in predicting V_S using an adaptive neuro-fuzzy inference system (ANFIS) algorithm and reported V_P as the most important parameter controlling V_S. Anemangely et al. (2019), on the other hand, concluded that an increase in the number of input parameters can lead to better prediction of V_S using non-dominated sorting genetic algorithm. Anemangely *et al*. (2019) also reported that the predictive models did not exhibit any sensitivity to more than five selected features as the input parameters. A common methodology for developing predictive models using machine learning techniques is to split the available data into training and testing sets. The purpose of splitting is to reserve a portion of the data for evaluating the model's accuracy in predicting the target parameter/s. This approach helps to ensure that the model has adequately learnt from the data to predict the output as accurate as possible with new input data. The split between training and testing data is typically performed randomly to reduce bias and to ensure fairness in the evaluation. In previous studies, this approach has been used to develop predictive models for sonic wave velocity estimation (Rezaee *et al*., 2007, Behnia *et al*., 2017; Wang and Peng, 2019; Du *et al*., 2019;). Typically, a set of geophysical logs, such as density, gamma, neutron porosity and sonic velocity are extracted from a borehole (Rajabi *et al*., 2010; Asoodeh and Bagheripour, 2012; Al-Dousari *et al*., 2016; Anemangely *et al*., 2019, Karimpouli and Fattahi, 2018). The data is then split into training and testing sets for model development. Since average downhole logging measurement is performed at short increments (e.g. 1, 5, 10 cm), adequate amount of data is available for both model training and testing.

The primary purpose of developing a V_P predictive model, however, is to technically estimate the property in locations where no sonic downhole logging is performed. However, testing the model using data extracted from the same borehole provides no insight into how the model behaves at blind locations where data trend and values are not seen by the model. This can be particularly critical in geological settings with high spatial variability, where significant differences in geophysical properties can be observed even over short distances, especially in the areas where multiple rock types with high heterogeneity exist. A limited number of boreholes cannot represent the true variability of the subsurface, and the model may overfit to the training data, leading to poor general performance. In other words, the models constructed on data from one or two boreholes are always shown to perform extremely good because the model performance is assessed on the borehole data where the model was trained with and not on boreholes where the data was not seen by the model at all. The effect of such spatial variability on ML based models' performance has not been assessed to date leaving a practical gap to conclude that ML models are in fact more accurate than simple one-dimensional empirical model. The other important point to consider is the computational efficiency of ML based predictive models. If a significant number of boreholes exist in a site, how much of data and how many boreholes are needed to train the best representative model with the lowest computational need.

To answer above questions, a systematic approach was utilized on a site containing a high number of boreholes (50) with a set of downhole logging measurements including density, gamma ray, depth and compressional wave velocity in Bowen Basin, Queensland, Australia. In the study mine, a significant proportion of boreholes (around 50 percent) lacked V_P measurements in their logging dataset thus developing a predictive model is highly beneficial. Five boreholes were randomly selected, and all the data of these five boreholes were set aside for model evaluation. It is crucial that only this independent testing set was considered for model evaluation as the models do not see any of the datapoints of these boreholes throughout the study. The remaining 45 boreholes were considered for training and model development purposes. As mentioned previously, unlike earlier studies, this study is not after an algorithm that shows the highest accuracy in prediction, but rather optimising the data usage that can capture the spatial variability in the study area and can be used for field-scale estimation. The LSSVR was nominated among ML algorithms for model development as it has been identified to be one of the most successful models for estimating sonic wave velocities in the literature (Cranganu and Breaban, 2013; Bagheripour *et al.*, 2015; Olayiwola and Sanuade, 2020). The study data included a relatively large number of boreholes mainly intersecting sandstone/silty sandstone, siltstone and coal lithologies with minor tuff and pebbles mostly within the Rangal coal measure sequence. For the modelling, a clustering analysis using the K-mean algorithm was first performed to determine if there were any clusters present in the data. The analysis aimed to confirm whether the data could be partitioned into smaller, more homogeneous clusters to improve the accuracy of predictive models for estimating V_P. The results of clustering analysis suggested that there were no distinct clusters present in the data. To develop a training set that captures the geological variation of the study area while being computationally efficient, the data from 45 was systematically divided boreholes into several smaller datasets. Six datasets were created consisting of 1, 5, 10, 20, 30 and 40 randomly selected boreholes from the original 45 boreholes. To account for uncertainty in the selection process, the random selection five times was performed for each group of boreholes. To further assess the variation in each set of data, we considered different portions of the data for training. Splits of 10, 20, 30, ..., 90 and 100 % were used to create a total of 300 separate datasets. Each dataset was then used as a separate training set to develop a V_P predictive model. Finally, the accuracy of all models developed were assessed using the five boreholes that were originally set aside for testing and plotted them in a heat map to find the most optimised combination. The above analysis can identify i) the effect of spatial variability on the performance of ML based predictive models and ii) the amount of data necessary to develop an optimised predictive model while keeping the computational costs low. This workflow can contribute to the development of more accurate and robust machine learning models for geological and geophysical applications at field-scale.

3.3. Data and methods

The data used in this study comprises 50 boreholes with downhole geophysical logs, including gamma ray, density, and compressional wave velocity available from exploration boreholes. All data were obtained from a coal mine located in the Bowen Basin (see Figures 3.1) of Queensland, Australia. The Permian-Triassic Bowen Basin is home to Australia's largest coal reserves, with 44 active coal mines (Mutton, 2003; Maxwell *et al.*, 2019). The basin comprises up to 10km of continental and shallow-marine, clastic sediments and deposits of black coal. The four dominant economically viable coal seams identified in this basin are classified into four domains based on their geological features. These domains include (I) the Reids Dome beds located in the southwestern part of the basin, which can reach up to 30 m thickness, (II) the Collinsville and Rugby coal measures as well as a group of deposits in the Clermont area, (III) coal measures deposited in the Collinsville Shelf and (IV) the Rangal and Baralaba coal measures plus the Bandanna formation (Mutton, 2003). The study mine is situated within the fourth group of coal seams.

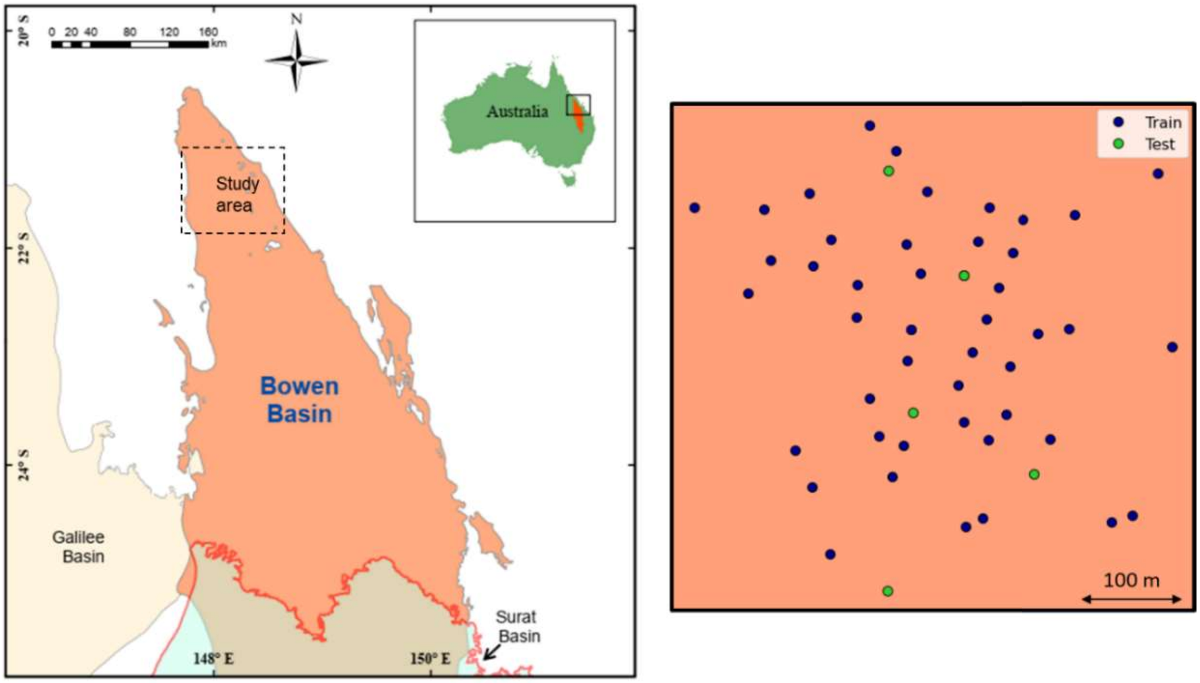


Figure 3.1 Approximate location of the study mine and the boreholes. Five test boreholes are coloured green.

During the Early Permian extensional phase in the Bowen Basin, sediment deposition occurred with the presence of fluvial and lacustrine sediments as well as volcanics in a series of half-grabens in the eastern region. Subsequent to rifting, a thermal subsidence (sag) phase extended from the Early to Late Permian, leading to a basin-wide transgression. This transgression allowed for the deposition of deltaic and shallow marine predominantly clastic sediments, alongside extensive coal measures. As the Late Permian progressed, foreland loading from east to west led to accelerated subsidence, facilitating the deposition of thick successions of Late Permian marine and fluvial clastics, including coal and Early to Middle Triassic fluvial and lacustrine clastics. The sedimentation process in the basin came to an end during the Middle to Late Triassic period. The stratigraphy of the Bowen Basin is presented in Figure 3.2, where the Late Permian Blackwater Group encompasses the Moranbah Coal Measures (MCM), the Fort Cooper Coal Measures (FCCM) and the Rangal Coal Measures (RCM) in chronological order (Mutton, 2003).

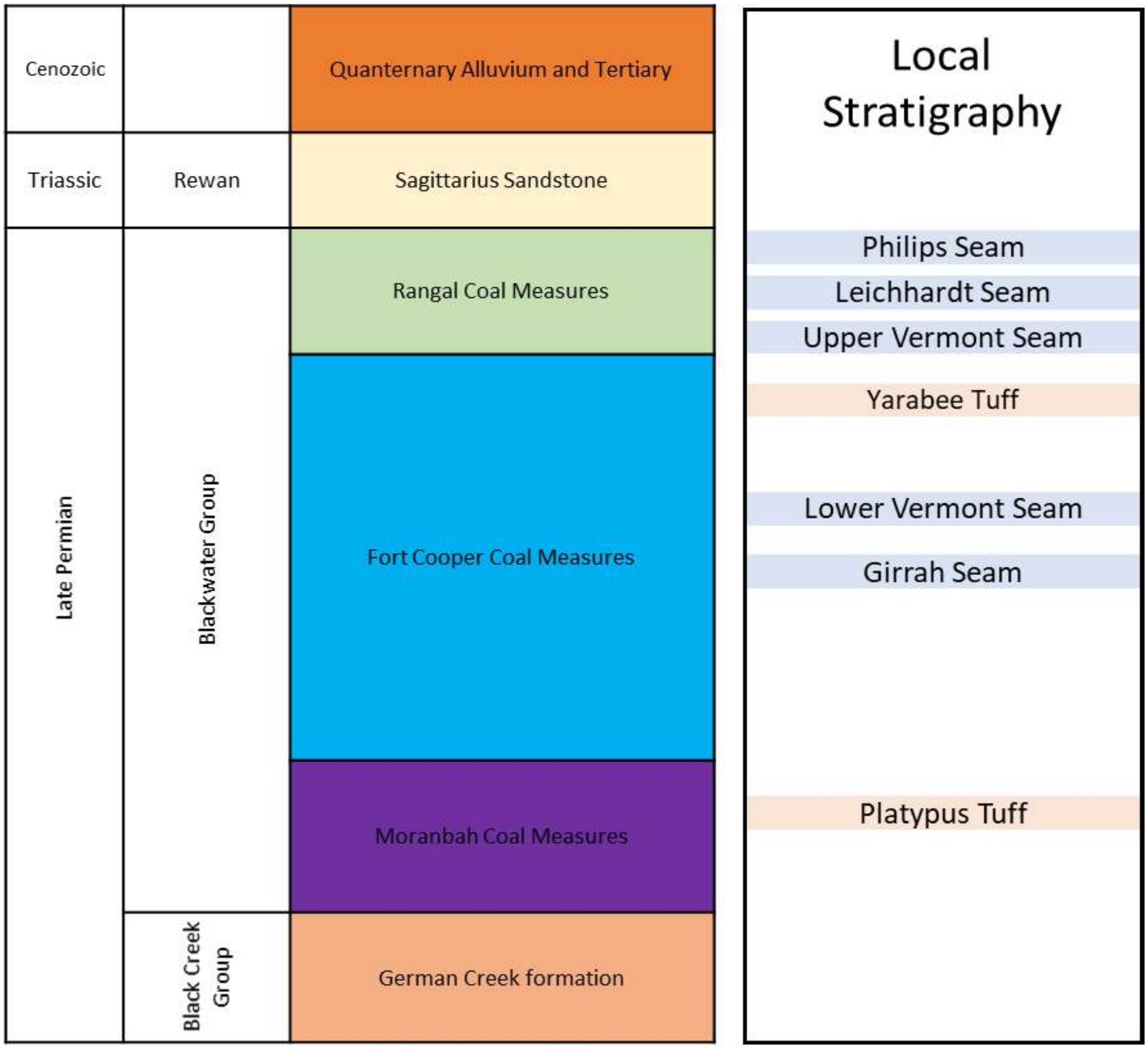


Figure 3.2. Stratigraphy of the study area

In this study, data from a comprehensive set of 50 boreholes having downhole geophysical logs, including gamma ray, density, and compressional wave velocity (as shown in Figure 3.3) were utilized. These exploration boreholes provided valuable insights into the geological composition of the region. The study boreholes predominantly intersect geological formations from Rangal Coal Measures (RCM) and Fort Cooper Coal Measures (FCCM), encompassing three primary lithology types: sandstone/silty sandstone, siltstone, and coal. These rock types are prevalent within the study area, and exhibit mostly horizontal units without any significant dipping.

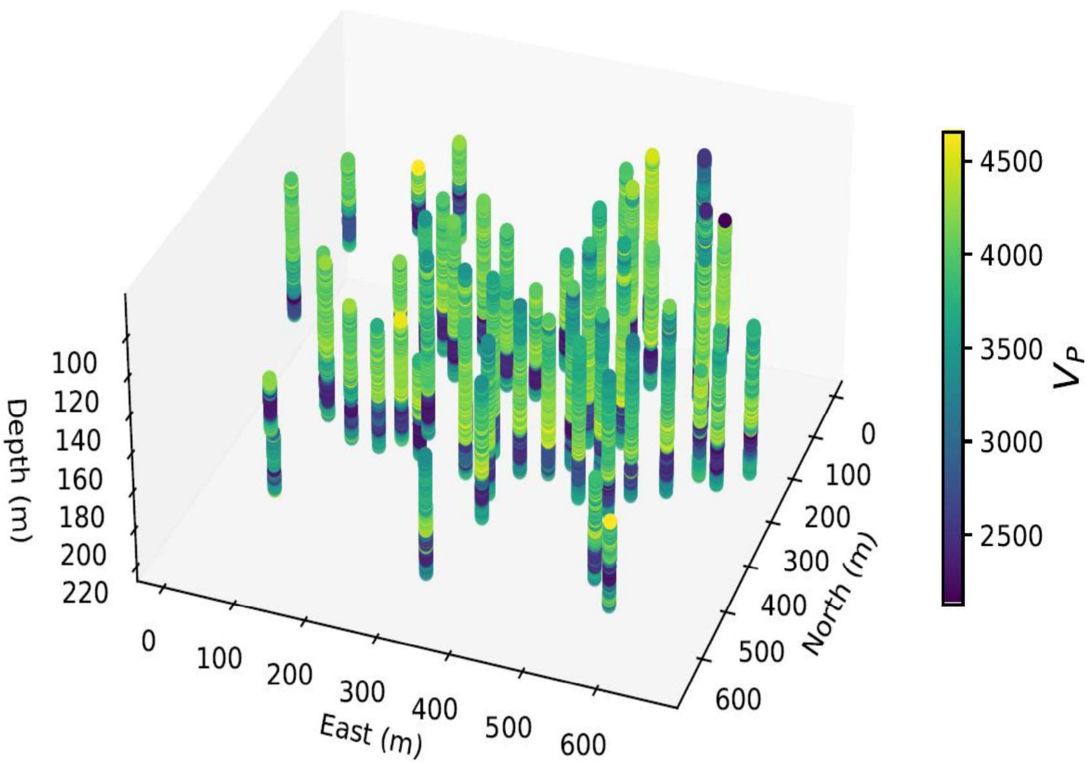


Figure 3.3. 3D view of study boreholes and variation of V_P with depth

3.3.1 Methodology

The primary objective of this research is to gain a deeper understanding of the impact of spatial variability in data on the accuracy of V_P predictive models using available downhole density, gamma ray and depth measurements. Therefore, the identification of the most appropriate ML algorithm or any enhancements to a particular algorithm is not within the scope of this study. Various ML algorithms have been employed in different studies for predicting V_P and V_S, and their advantages and disadvantages have been extensively discussed (Asoodeh and Bagheripour, 2012; Anemangely et al., 2019; Mehrgini et al., 2019; Wang and Peng, 2019; Lv et al., 2021, Ibrahim, 2022).

SVR is a commonly used machine learning algorithm used for similar applications (Cranganu and Breaban, 2013; Bagheripour et al., 2015; Olayiwola and Sanuade, 2020). However, SVR has a significant disadvantage of high computational requirement for optimization. To address this limitation, LSSVR method was developed (Suykens et al., 2002), which is an improved version of SVR. The LSSVR was therefore used for predictive modelling. In order to develop accurate and effective predictive models, it is crucial to first gain an understanding of the data being analysed and identify any underlying patterns or clusters. Clustering analysis is a technique commonly used to achieve this goal. The process of clustering analysis involves grouping data points based on their shared characteristics or features. By identifying such groupings, it becomes possible to gain insights into the underlying structure of the data. This can inform the development of predictive models that are tailored to the specific characteristics of the data being analysed. One commonly used clustering technique is the K-mean algorithm, which was employed in this study.

3.3.1.1. LSSVR

LSSVR is an improved version of SVR algorithm whose objective and restriction functions are slightly different to that of SVR (Suykens et al., 2002). LSSVR solves linear equation instead of quadratic hence reduces the computational cost and complexity of regression analysis (Wang and Hu, 2005). Its objective function is defined as:

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2 \qquad \text{Subject to } y_i = w^T.F(x) + b + e_i \tag{3.1}$$

where F is a non-linear function for mapping input into a higher dimensional feature space, w is the weight factor, e_i is the error in each data point and γ is the regularisation parameter. The Lagrange multiplier is used to find the optimum values:

$$L(\omega, b, e; \alpha) = J(w, e) - \sum \alpha_k\{w^T.F(x_k) + b + e_k - y_k\} \tag{3.2}$$

where the derivatives of Lagrange multiplier are determined as:

$$\begin{aligned}
&\frac{\partial L}{\partial w} = 0 \to w = \sum_{k=1}^{N} \alpha_k\{w^T.F(x_k) \\
&\frac{\partial L}{\partial b} = 0 \to \sum_{k=1}^{N} \alpha_k = 0 \\
&\frac{\partial L}{\partial e_k} = 0 \to \alpha_k = \gamma e_{k,} \qquad\qquad k = 1, \dots, N \\
&\frac{\partial L}{\partial \alpha_k} = 0 \to w^T.F(x_k) + b + e_k - y_k = 0, \quad k = 1, \dots, N
\end{aligned} \tag{3.3}$$

Solving the above equations and applying Mercer's condition results in an estimation function resulting in LSSVR to be used for nonlinear function:

$$y_i = \sum_{i=1}^{l} \alpha_i K(x.x_i) + b \tag{3.4}$$

where α_i and K are the support vectors and kernel functions, respectively.

3.3.1.2. K-mean

K-means is a popular clustering algorithm used in unsupervised ML. It aims to partition a given dataset into K clusters, where K is a user-specified parameter representing the number of clusters to be formed. The algorithm works by iteratively assigning each data point to the nearest cluster centre (centroid) and then updating the cluster centres based on the mean of the data points assigned to each cluster. Euclidean distance metric is a common technique used in K-mean and was thus employed here. Considering x as the feature of new sample, x_i as the feature vector of training samples and j as the number of training samples, the Euclidean distance is defined according to:

$$\text{Distance} = \sqrt{\sum_{j=1}^{m} \left| x_j - x_{ij} \right|^2} \tag{3.5}$$

3.4. Data processing and preparation

Gamma ray, density, and compressional wave velocity logs with depth were available in all 50 boreholes in the study area. Before performing modelling, the data from younger depositions (approximately 0-40 m of logging depth) were removed. The statistical information on the log data samples of all 50 boreholes are presented in Table 3.1. In the first step for data preparation, the outliers should be removed. Tukey's method (Tukey, 1975) is a well-established statistical technique for removing data outlier (Madhubabu et al., 2016; Anemangely et al., 2019; Mehrad et al., 2020). In this method, first, the interquartile range (IQR) is computed as a distance between the first and the third quartiles of data. Then, the inner and the outer boundaries are defined at 1.5 IQR distance before the first and after the third quartiles, and finally all the data outside these boundaries are considered outliers and are removed from the analysis. The outlier removal has been performed for all downhole log set in each borehole. To further enhance the outlier detection process and preserve potentially valuable information, a refined approach was employed. In this approach, instead of directly removing potential outlier data points, the data from each borehole was divided into subsections of 5-meter intervals. Subsequently, the Tukey method was applied on each of these subsections independently. This localized outlier detection technique allowed identifying the outliers within smaller intervals, ensuring that useful information was retained and not eliminated during the outlier removal process. This approach also effectively balanced the need for outlier detection with the goal of retaining valuable data, leading to a more accurate and reliable dataset for subsequent analyses.

Table 3.1 Summary of the study log data

Log parameters	Max	Min	Mean	STD
Density (g/cm^3)	3.18	0.72	2.3439	0.431
Gamma Ray (API)	218.45	0.01	75.8	34.35
logging depth (m)	180.65	40.01	100.161	28.52
V_P (m/s)	4913.42	833.33	3501.9	732.169

It is noteworthy that, since different geophysical logs have various ranges of values (e.g. a typical value for density is 2.21 g/cm^3 while common measurement of V_P is around 2500 m/s), the data should be normalised for each log (or parameter) to have equal contribution to the model training. Eq. (3.6) was used for data normalisation which requires the standard deviation and mean of input and output parameters in each borehole (Mendenhall and Sincich, 2016):

$$z = \frac{x - \mu}{SD} \tag{3.6}$$

where x is the log value, μ and SD are the mean and standard deviation of data, respectively and z represents the normalised value. It should be noted that in this study, the term "normalisation" specifically refers to standardizing the data (Eq. (3.6)) and is different to the calibration or normalization, which are commonly used in downhole logging.

Prior to applying the regression algorithms, it is important to select the key or most influential features so that the models can perform effectively (Miah et al. 2020; Du et al. 2019). To do so, multivariate regression analysis was deployed and initially, all four-log data including density, gamma ray, resistivity and depth were considered as input parameters. Hence, a set of simple predictive models was developed for V_P using the linear regressions and the sensitively of each parameter was assessed by excluding them one by one. Finally, it was found that the resistivity log had a negligible effect on the modelling process and thus was excluded (see Table 3.2). Therefore, later analyses and the modelling were performed using density, gamma ray and depth as the input parameters for V_P prediction.

Table 3.2 Sensitivity assessment of different logs for features selection

Input parameters	R^2
Density, Gamma Ray, Depth, Resistivity	0.83
Density, Gamma Ray, Depth	0.83
Density, Gamma Ray	0.80
Density	0.75

3.4.1 Clustering analysis

To ensure the accuracy and reliability of predictive models for estimating V_P, an initial analysis of our data obtained from 50 boreholes with varying geological formations was conducted. The potential utility of clustering to improve the accuracy of our models and employed the K-mean clustering algorithm to partition our data into smaller, more homogeneous clusters was recognized. The aim was to investigate whether performing regression analysis on these smaller clusters would lead to better results compared to using all the data for regression analysis (to predict V_P). To identify the clusters, several key parameters such as density, gamma, depth, and location of each borehole (not rock type) were considered. By analysing these factors, the aim was to uncover potential clusters within the dataset. A set of trial-and-error analyses were conducted to determine the optimal number of clusters used in the clustering. An analysis was conducted by varying the number of clusters from 1 to 10 and evaluated the clustering performance using the Silhouette score and Within-Cluster Sum of Squares (WCSS). The results were visualized in Figure 3.4. While WCSS and Silhouette score are commonly used measures to identify the appropriate number of clusters, the cross plots in Figure 3.4 did not provide clear-cut indications for selecting the optimal number of clusters.

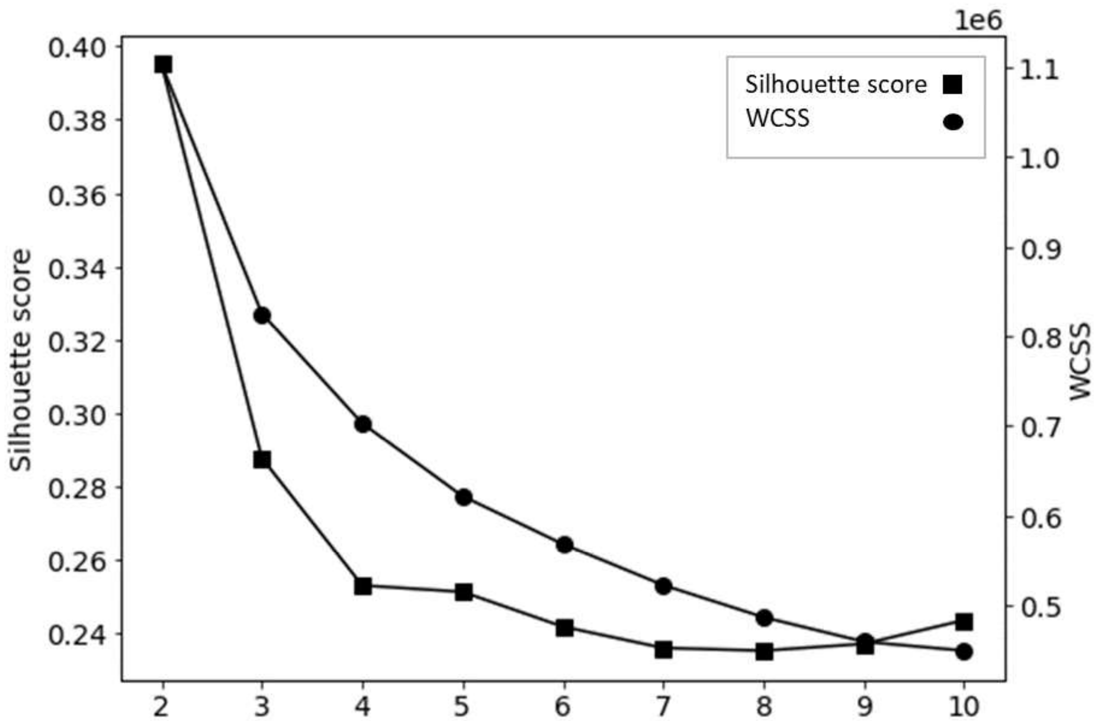


Figure 3.4. Variation of silhouette score by changing the number of clusters

However, it is noteworthy that the Silhouette score was minimized when the number of clusters was set to 2 or 3. The Silhouette score measures how well each data point fits into its assigned cluster, indicating that 2 and 3 clusters displayed the best fit for our data. Consequently, based on this analysis, it has been decided to focus on considering two and three clusters for further modelling. It is important to mention that although the cross plots did not provide a definitive criterion for selecting the number of clusters, the Silhouette score's behaviour and the WCSS values guided our decision-making process. By narrowing down our options to two and three clusters, we aim to create a more interpretable and meaningful clustering solution for our data.

To visualize the data and the identified clusters, two main Principal Components (PCs) (Jackson, 2005) was used that capture the most variation in the data. This allowed us to reduce the dimensionality of the data while retaining as much information as possible. The scatter plot of the first two PCs, with each point coloured according to its assigned cluster, is presented in Figure 3.5. it was observed that the clusters are not well-separated or distinct from each other suggesting no clustering of the data.

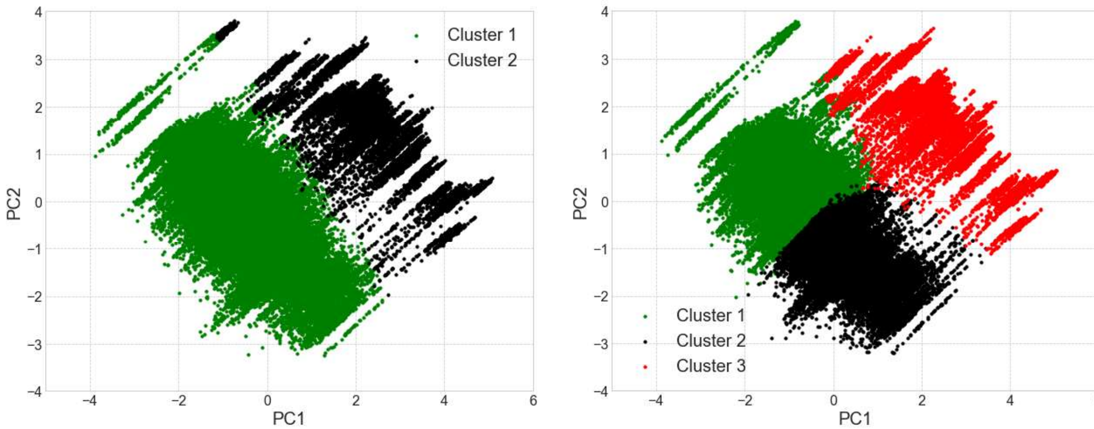


Figure 3.5. Identified clusters using K-mean for two and three clusters.

In the subsequent stage, it was proceeded to train machine learning (ML) models to estimate V_P (for each identified cluster) with the aim of investigating whether partitioning the data into smaller datasets can lead to improved prediction performance. The models were built using gamma, density, and depth as input parameters, and V_P as the output variable. For the regression modelling, it was employed the LSSVR (Least Squares Support Vector Regression) algorithm. To develop the models, to a standard methodology of splitting the dataset into 80% for training and 20% for testing, ensuring a robust evaluation of the model's predictive capabilities was adhered.

Based on the five clusters identified from the first and second set of clustering, individual models were developed for each cluster. Additionally, a model using all the data without any clustering was created. This allowed us to compare the performance of cluster-specific models against the model with no clustering. In Tables 3 and 4, A comprehensive comparison between the predicted V_P values and the actual V_P values for each model, providing insights into their respective accuracies and effectiveness in estimating V_P for the different clusters is presented.

Table 3.3 The accuracy of developed models with and without clustering

Model	R^2	RMSE
Cluster 1	0.58	0.97
Cluster 2	0.65	0.82
Without clustering	0.88	0.42

Table 3.4 The accuracy of developed models with and without clustering for first set of clustering (only 3 clusters assumed)

Model	R^2	RMSE
Cluster 1	0.39	1.48
Cluster 2	0.42	1.21
Cluster 3	0.65	0.82
Without clustering	0.88	0.42

The results revealed that the model trained on all data without clustering has the highest prediction accuracy, with the lowest error between the predicted and actual V_P values. In contrast, the models trained on separate clusters showed lower prediction performance, with a higher error between the predicted and actual V_P values. The modelling results again confirm that the clustering did not improve the prediction performance. It is possible that the differences between the subpopulations in the data were not significant enough to justify separate modelling or that the LSSVR algorithm was more effective in capturing the underlying relationships between the features and the target variable for the entire dataset. Therefore, it was decided not to consider clusters in development of V_P predictive models.

3.4.2 Design of datasets

The study area consisted of 50 boreholes, out of which 5 were selected randomly for validation purposes. The remaining 45 boreholes were used for training the developed models. To ensure that the training sets captured the randomness and variability within the data, several training sets were created. Each training set was selected using a random number generator, without any human intervention. The number of boreholes considered for each training set varied between 1, 5, 10, 20, and 30, 40 and each set was tested five times to account for the randomness factor. This process produced 30 distinct sets of training data. These datasets were divided into different portions, ranging from 10% to 100%, to determine the optimal amount of data needed to extract from each borehole to capture the vertical variation in the data. This would result in achieving the optimal balance between accuracy and computational cost. This resulted in 300 final training datasets, each of which was used to train the developed models. Overall, this rigorous approach in selecting and creating training datasets helps to ensure that the developed models are robust and can perform well on a variety of data inputs. Figure 3.6 shows the selection procedure of data training.

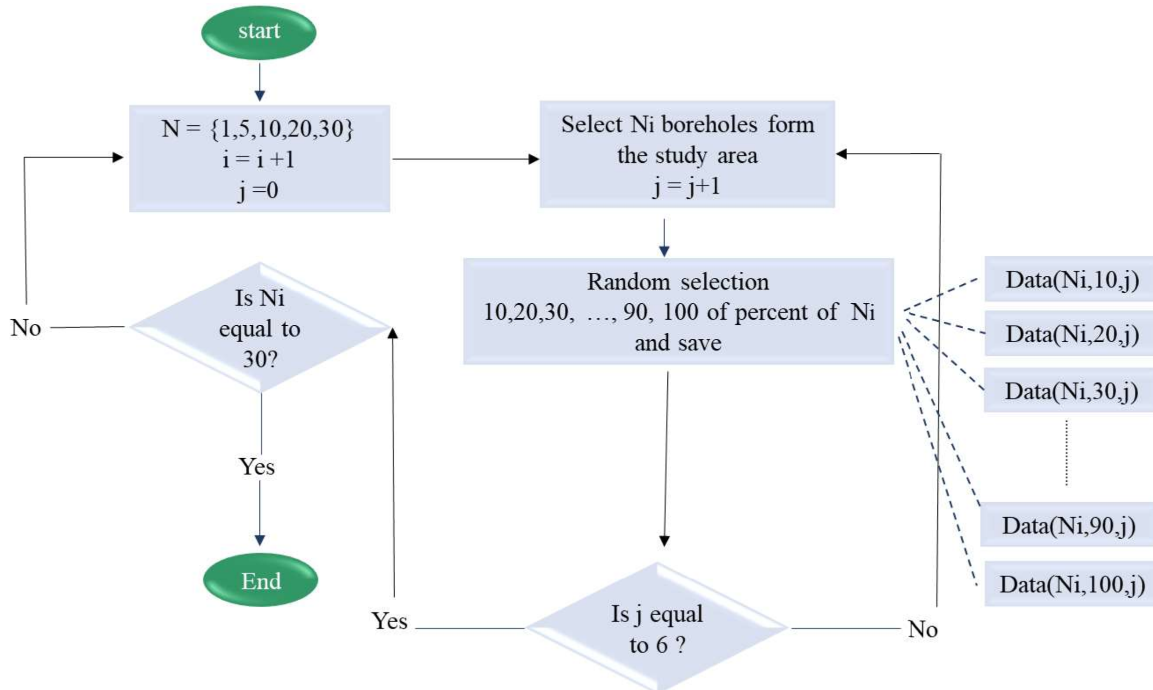


Figure 3.6. Flowchart of data selection

3.5. Model development

To develop robust and accurate predictive models for estimating compressional wave velocity (V_P), a dataset of 300 records, each comprising three downhole logs - density, gamma ray, and depth - as input parameters and V_P as output was utilized. These logs provide information on the subsurface characteristics of the geological formations, which are essential for predicting V_P accurately. The data from each borehole was pre-processed by removing the detected outliers and normalizing the data to ensure consistency and to avoid biases in the results. the data from various lithologies have not been segregated since the boreholes intersect predominantly horizontal layers, each comprising a limited number of lithologies. Consequently, it was chosen to adopt a generalized modelling approach. It is important to note that while this method proves effective for the current geological context, it may not be suitable for more intricate geological settings. it was then used the LSSVR method to develop 300 predictive models for V_P, one for each dataset.

To map the data into the non-linear space, the radial basis function (RBF) was implemented as the kernel function for all the developed models. RBF is a widely used kernel function that is known to perform well for various types of data (Keerthi and Lin, 2003). This kernel function allowed to capture the non-linear relationships between the input parameters and the output (V_P) more effectively. To optimize the hyperparameters of LSSVR, including C and γ, a random search approach were employed for each dataset separately. The hyperparameters were optimized by exploring a range of possible values using a random search algorithm (You et al., 2010). This approach allowed us to find the best combination of hyperparameters that resulted in the most accurate predictions for each dataset. The coefficient of determination (R^2) and root-mean square error (RMSE) were calculated as the statistical indicators to assess the predictability of the resulting models. While R^2 is a well-known indicator in engineering applications, the RMSE is estimated through the following function:

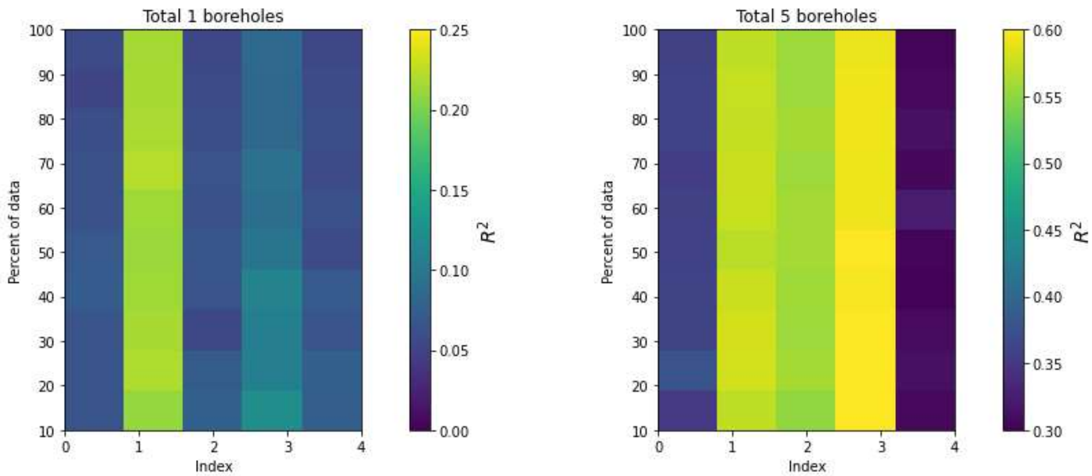
$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2} \tag{3.7}$$

where n is the total number of samples, y_i denotes the measured variable and f_i is the predicted value. The predictability of the resulting models was examined based on the low or high values of statistical indicators in which the best fit happened at the maximum and minimum of R^2 and RMSE, respectively.

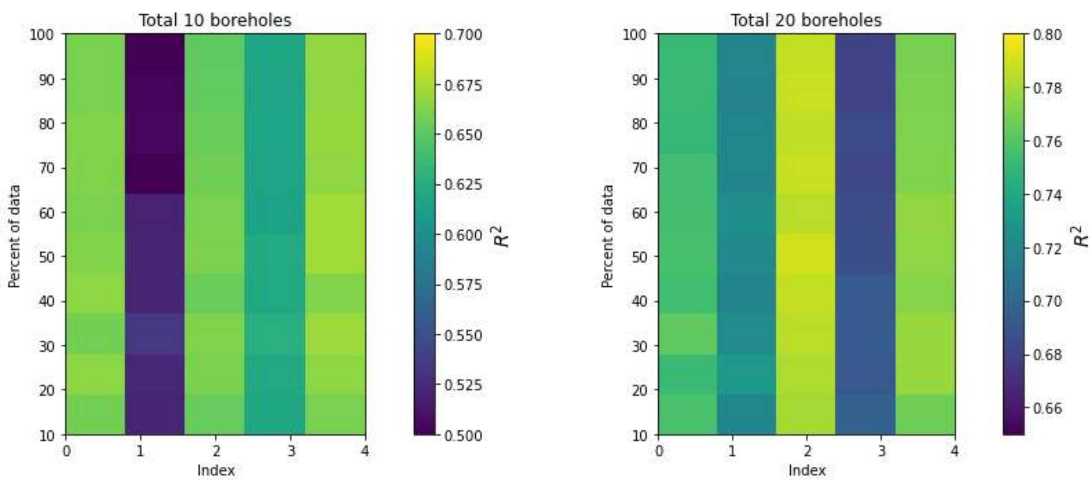
3.6. Model assessment

To evaluate the effectiveness of the developed models in real field predictions and to estimate V_P at blind locations, the data from five random boreholes were used in the study area as validation data. Previously, 300 datasets were designed, each containing different portions of data extracted from a different number of boreholes. These datasets were used to develop 300 models, each of which was used to estimate V_P for the five test boreholes. Then the R^2 and RMSE were calculated for all the models to assess their performance. To better visualize the performance of the models, six heat maps were created, each illustrating the performance of 50 models, as shown in Figure 3.7. The heat maps only include R^2 values, as the RMSE showed the exact same behaviour. The heat maps helped to identify the most effective models that could be used for real field predictions and to estimate V_P at blind locations. They can also assist in observing the trend of variation in model performance based on the number of boreholes and portion of data used in the training set. Each heat map in the figure displays all the models developed using the same number of boreholes but with different portions of data and varying combinations of random selections.

(a) (b)



(c) (d)



(e) (f)

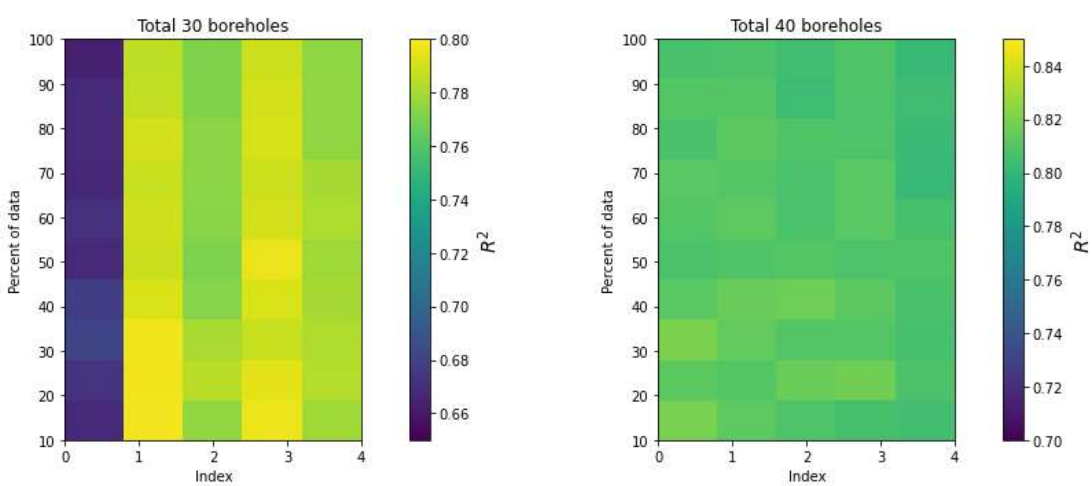


Figure 3.7. The heatmaps showing the estimated R^2 obtained from all 300 developed models.

Figure 3.7 presents subfigures, where used "Percent of data" and "index" were used to identify the training dataset that was used for developing the predictive models, while keeping the number of boreholes constant. The index axis contains five values (0, 1, 2, 3, 4) which refer to how many times the random selection of boreholes was conducted representing the randomness effect in the analysis. This factor was used to extract different combinations of data from the study area, as described in section 3.2. The amount of data that was extracted from each borehole varied from 10% to 100% and is presented on the y-axis. To illustrate the performance of the developed models, the R^2 values were used as the colour bar. The R^2 value represents the degree of correlation between the predicted and actual V_P values, with higher values indicating better predictions. The colour bar allowed us to easily identify the R^2 values for each model and to compare their performance across different datasets. Overall, the subfigures in Figure 3.7 provide a comprehensive overview of the performance of the developed models for different combinations of training datasets based on considered number of boreholes, portion of the data and randomness in borehole selection.

A sharp difference can be observed among each iteration, in subfigures (a) and (b) (1 and 5 boreholes). The sudden change of colour in different selections (index) suggests that the data from one borehole can be significantly different to the data from another borehole in the study area. For example, when selecting five boreholes, the data extracted from the second selected borehole (index 1) could provide better predictions for the test data than in the first iteration (index 0). This observation suggests that the spatial variability in the horizontal direction is likely to be higher than the vertical direction (all rock types considered together). This finding is further supported by the comparison of the average standard deviation of V_P in the horizontal direction, which is approximately 560.76, while the vertical standard deviation across different rock types ranges between 270 and 460.10. The high spatial variability in the horizontal direction indicates that local geological factors could play a crucial role in controlling the properties of the subsurface. It is important to note that as we increased the number of boreholes and selected samples from a total of 45 boreholes, the change in colour became less distinct. Additionally, increasing the number of boreholes in the sampling process also raises the likelihood of including the same boreholes in each iteration. For example, in subfigure e, in each iteration, we selected 40 boreholes from 45 boreholes resulting in a very similar dataset with overlapping training set (boreholes).

Upon closer examination of each subfigure in Figure 3.7, where the number of boreholes used in training remain constant, increasing the portion of training data had little to no impact on the resulting R^2. This trend is evident in all subfigures, as there are no noticeable changes when following the increase in data portion for each index. For instance, in all 10 models developed using one borehole in the first iteration (0 index), increasing the data portion from 10 %to 100% did not provide any additional information to the model and thus, did not improve the R^2. This observation interestingly suggests that the vertical variation in this dataset is either negligible or even 10% of the data selected randomly can capture the vertical variability. Therefore, increasing the portion of training data does not improve the accuracy of the predictive models. This finding is consistent across all the predictive models developed based on different number of boreholes e.g. models developed based on data from 40 boreholes follow the same trend.

After analysing the R^2 range in each subfigure and examining the trend from subfigure a to f, it becomes evident that increasing the number of boreholes leads to a noticeable improvement in R^2. This trend in R^2 can be compared with the behaviour observed when the percentage of data was varied, which suggests that the variation in the horizontal plane is more critical than the vertical direction in model development. In other words, the model should prioritize incorporating data from different boreholes instead of gathering more data points from a limited number of boreholes.

The other important observation is related to the number of boreholes random selection (index) from 0 to 4 times. Interestingly, the variation across index 0 to 4 with increase in number of boreholes is intuitive as more boreholes can reduce the compositional selectivity i.e. the more similar selections occur with increase in boreholes.

To better visualize the changes in R^2 value resulting from increasing the number of boreholes, all the models with the same index (i.e., the random repeated models) were averaged and presented the results in Figure 3.8. As expected, the increase in the number of boreholes led to models with higher accuracies. However, the rate of improvement in R^2 became less steep after 20 boreholes (Figure d) and eventually stabilized. This suggests that selecting 20 boreholes can capture the required information for a representative prediction. Additionally, this figure confirms the previously observed negligible variation in R^2 resulting from increasing the percentage of training data, as all the models from 10% to 100% align on a single point. Thus, one expects that a model trained with only 10% of the data from 20 boreholes is the most effective model having low computational cost yet good prediction accuracy.

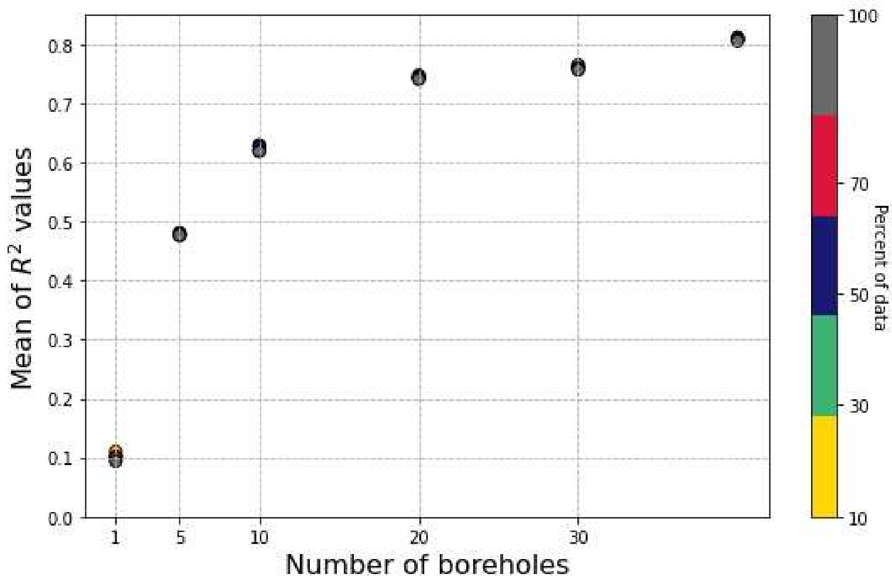


Figure 3.8. Averaged R^2 excluding the randomness factor from the developed models.

To further investigate the impact of spatial variability and its influence on the development of rock physical properties, hybrid approaches integrating machine learning with geostatistical simulation is required. While such integration is beyond the scope of this study, it will be a valuable addition to current state of the art in stochastic simulations.

3.7. Conclusion

This study presents a systematic workflow that evaluates the impact of spatial variation in geophysical logs on the accuracy of ML models in predicting formation properties. The analysis uses a large dataset of geophysical logs from 50 boreholes in a coal mine in Bowen Basin, QLD, Australia. To perform this evaluation, a clustering analysis was initially performed using the K-mean algorithm to identify any smaller, homogeneous clusters in the data, where no distinct clusters were found. Out of 50 boreholes, 45 were used for training the predictive models, while the remaining 5 were used for validation purposes. To account for data variability and randomness, multiple training sets were created using a random number generator. The number of boreholes included in each training set varied, ranging from 1, 5, 10, 20, and 30 to 40, and each set underwent five repetitions to consider the element of randomness. This process yielded 30 distinct training datasets. Furthermore, these datasets were divided into different portions, spanning from 10% to 100%, to determine the optimal amount of data required from each borehole, effectively capturing the vertical variation within the data. This approach aimed to strike a balance between accuracy and computational cost, resulting in 300 final training datasets, each employed to train the developed models. The meticulous selection and creation of training datasets through this rigorous procedure ensure the robustness of the models and their ability to perform well across diverse data inputs. These sets were then used to train the developed models using least squares support vector regression and evaluated using test data.

The results of the analyses showed that the number of boreholes had a greater impact on model performance than the number of data points. A larger number of boreholes with a smaller portion of data can lead to more accurate and computationally efficient models. The study also found that vertical variation in the dataset was negligible, while spatial variability in the horizontal plane was much higher, indicating that local geological factors play a crucial role in controlling subsurface properties. Overall, this study provides valuable insights into ML-based predictive models for estimating petrophysical properties from downhole geophysical logs for field-scale applications. The rigorous approach to selecting and creating training datasets ensures that the developed models are robust and can perform well on a variety of different data inputs. These findings can be beneficial for geoscientists who rely on predictive models to estimate subsurface properties.

3.8. References

Abbas A K, Flori R E, Alsaba M (2018) Estimating rock mechanical properties of the Zubair shale formation using a sonic wireline log and core analysis. J Nat Gas Sci Eng 53:359- 369

Al-Dousari M, Garrouch A A, Al-Omair O (2016) Investigating the dependence of shear wave velocity on petrophysical parameters. J Petrol Sci Eng 146:286-296

Anemangely M, Ramezanzadeh A, Amiri H, Hoseinpour S-A (2019) Machine learning technique for the prediction of shear wave velocity using petrophysical logs. J Petrol Sci Eng 174:306-327

Asoodeh M, Bagheripour P (2012) Prediction of compressional, shear, and stoneley wave velocities from conventional well log data using a committee machine with intelligent systems. Rock Mech Rock Eng 45(1):45-63

Asquith, G., Gibson, C (2004). Basic well log analysis for geologists, American Association of Petroleum Geologists Methods in Exploration.

Bhatt, A (2002). Reservoir properties from well logs using neural networks.

Bagheripour P, Gholami A, Asoodeh M, Vaezzadeh-Asadi M (2015) Support vector regression based determination of shear wave velocity. J Petrol Sci Eng 125:95.

Behnia D, Ahangari K, Moeinossadat S R (2017) Modeling of shear wave velocity in limestone by soft computing methods. Int J Min Sci Technol 27(3):423-430

Biot, M.A (1941). General theory of three-dimensional consolidation. J. Appl. Phys. 12, 155–164.

Birch, F (1961). The velocity of compressional waves in rocks to 10 kilobars: 2. J. Geophys. Res. 66, 2199–2224.

Birch, F (1960). The velocity of compressional waves in rocks to 10 kilobars: 1. J. Geophys. Res. 65, 1083–1102.

Castagna J P, Backus M M (1993). Offset-dependent reflectivity—Theory and practice of AVO analysis, Society of Exploration Geophysicists.

Cranganu C, Breaban M (2013) Using support vector regression to estimate sonic log distributions: a case study from the Anadarko Basin, Oklahoma. J Petrol Sci Eng 103:1-1

Du Q, Yasin Q, Ismail A, Sohail G M (2019) Combining classification and regression for improving shear wave velocity estimation from well logs data. J Petrol Sci Eng 182:10626

Eberhart-Phillips D, Han D-H, Zoback M (1989) Empirical relationships among seismic velocity, effective pressure, porosity, and clay content in sandstone. Geophysics 54(1):82-89

Ebrahimi, A., Izadpanahi, A., Ebrahimi, P., & Ranjbar, A. (2022). Estimation of shear wave velocity in an Iranian oil reservoir using machine learning methods. J Petrol Sci Eng, 209, 109841.

Gassmann, F (1951). Elastic waves through a packing of spheres. Geophysics 16, 673–685.

Hatherly P, 2013. Overview on the application of geophysics in coal mining. International Journal of Coal Geology, 114, pp.74-84.

Hatherly P, Medhurst T, Zhou B (2016) Geotechnical evaluation of coal deposits based on the Geophysical Strata Rating. Int J Coal Geol 163:72-86

Ibrahim, A. F. (2022). Application of various machine learning techniques in predicting coal wettability for CO2 sequestration purpose. International Journal of Coal Geology, 252, 103951.

Jackson, J. E. (2005). A user's guide to principal components. John Wiley & Son

Karacan C Ö (2009) Elastic and shear moduli of coal measure rocks derived from basic well logs using fractal statistics and radial basis functions. Rock Mech Rock Eng 46(8):1281-129

Khandelwal, M., Singh, T.N (2010). Artificial neural networks as a valuable tool for well log interpretation. Pet. Sci. Technol. 28, 1381–1393.

Karimpouli S, Fattahi H (2018) Estimation of P-and S-wave impedances using Bayesian inversion and adaptive neuro-fuzzy inference system from a carbonate reservoir in Iran. Neural Comput Appl 29(11):1059-1072

Keerthi S S, Lin C-J (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput 15(7):1667-1689

Kowallis, B.J., Jones, L.E.A., Wang, H.F (1984). Velocity-porosity-clay content systematics of poorly consolidated sandstones. J. Geophys. Res. Solid Earth 89, 10355–10364.

Khandelwal M, & Singh, T. N. (2009). Correlating static properties of coal measures rocks with P-wave velocity. International Journal of Coal Geology, 79(1-2), 55-60.

Lv A, Cheng L, Aghighi M A, Masoumi H, Roshan H (2021) A novel workflow based on physics-informed machine learning to determine the permeability profile of fractured coal seams using downhole geophysical logs. Mar Pet Geol 131:105171

Madhubabu N, Singh P, Kainthola A, Mahanta B, Tripathy A, Singh T (2016) Prediction of compressive strength and elastic modulus of carbonate rocks. Measurement 88:202-2

Maxwell K, Rajabi M, Esterle J (2019) Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. Int J Coal Geol 214:103284

Mehrad M, Bajolvand M, Ramezanzadeh A, Neycharan J G (2020) Developing a new rigorous drilling rate prediction model using a machine learning technique. J Petrol SciEng 192:107338

Mehrgini B, Izadi H, Memarian H (2019) Shear wave velocity prediction using Elman artificial neural network. Carbonates and Evaporites 34(4):1281-1291

Mendenhall W M, Sincich T L, (2016). Statistics for Engineering and the Sciences, CRC Press.

Minear, J.W (1982). Clay Models and Acoustic Velocities, in: SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Miah M I, Ahmed S, Zendehboudi S, Butt S (2020) Machine Learning Approach to Model Rock Strength: Prediction and Variable Selection with Aid of Log Data. Rock Mech Rock Eng 53(10):4691-4715

Mutton A J (2003) Queensland coals. Physical and chemical properties, colliery and company information.

Mullen, M.J., Roundtree, R., Turk, G.A (2007). A Composite Determination of Mechanical Rock Properties for Stimulation Design (What to Do When You Don't Have a Sonic Log). Rocky Mt. Oil Gas Technol. Symp.

Nur, A., Simmons, G (1969). The effect of saturation on velocity in low porosity rocks. Earth Planet. Sci. Lett. 7, 183–193.

Olayiwola T, Sanuade O A (2020) A data-driven approach to predict compressional and shear wave velocities in reservoir rocks. Petroleum

Onalo D, Adedigba S, Khan F, James L A, Butt S (2018) Data driven model for sonic well log prediction. J Petrol Sci Eng 170:1022-1037

Pickett, G.R (1963). Acoustic Character Logs and Their Applications in Formation Evaluation. J. Pet. Technol. 15, 659–667.

Ramcharitar, K., Hosein, R (2016). Rock Mechanical Properties of Shallow Unconsolidated Sandstone. SPE Trinidad Tobago Sect. Energy Resour. Conf

Rajabi M, Bohloli B, Ahangar E G (2010) Intelligent approaches for prediction of compressional, shear and Stoneley wave velocities from conventional well log data: A case study from the Sarvak carbonate reservoir in the Abadan Plain (Southwestern Iran). Comput Geoci 36(5):647-664

Raymer, L.L.L., Hunt, E.R.R., Gardner, J.S (1980). An Improved Sonic Transit Time to Porosity-to-Porosity transform. 21st SPWLA Logging Symp. Trans. 1–13.

Rezaee M R, Ilkhchi A K, Barabadi A (2007) Prediction of shear wave velocity from petrophysical data utilizing intelligent systems: An example from a sandstone reservoir of Carnarvon Basin, Australia. J Petrol Sci Eng 55(3-4):201-212

Rolon, L., Mohaghegh, S.D., Ameri, S., Gaskari, R., McDaniel, B (2009). Using artificial neural networks to generate synthetic well logs. J. Nat. Gas Sci. Eng. 1, 118–133.

Russell S, Norvig P (2002) Artificial intelligence: a modern approach.

Schön, J (2015). Basic Well Logging and Formation Evaluation Basic Well Logging and Formation Evaluation 4 Contents, 1st ed.

Suykens J.A, De Brabanter J, Lukas L, Vandewalle J (2002) Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48(1-4):85-105

Tosaya, C., Nur, A (1982). Effects of diagenesis and clays on compressional velocities in rocks. Geophys. Res. Lett. 9, 5–8.

Tosaya, C.A (1982). Acoustical properties of clay-bearing rock. Stanford University.

Tixier, M.P., Loveless, G.W., Anderson, R.A (1975). Estimation of Formation Strength From the Mechanical-Properties Log. SPE J. Pet. Technol. 27, 283–293.

Tan M, Peng X, Cao H, Wang S, Yuan Y (2015) Estimation of shear wave velocity from wireline logs in gas-bearing shale. J Petrol Sci Eng 133:352-366

Tukey J W (1975), Conference. Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians, Vancouver.

Wang J, Cao J, Yuan S (2020) Shear wave velocity prediction based on adaptive particle swarm optimization optimized recurrent neural network. J Petrol Sci Eng 194:107466

Wang P, Peng S (2019) On a new method of estimating shear wave velocity from conventional well logs. J Petrol Sci Eng 180:105-123

Wang H, Hu D (2005) Comparison of SVM and LS-SVM for regression. 2005 International conference on neural networks and brain, IEEE.

Xu J, Tan M, Wang X, Wu C (2017) Predicting acoustic-wave velocities and fluid sensitivity to elastic properties in fractured carbonate formation. Interpretation 5(1):SB69-SB80

Zhang D, Lin J, Peng Q, Wang D, Yang T, Sorooshian S, Liu X,Zhuang J (2018) Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. Journal of Hydrology 565:720-736

Zhang Y, Zhong H-R, Wu Z-Y, Zhou H, Ma Q-Y (2020) Improvement of petrophysical workflow for shear wave velocity prediction based on machine learning methods for complex carbonate reservoirs. J Petrol Sci Eng 192:107234

Zoback, M.D (2010). Reservoir Geomechanics. Cambridge University Press

Chapter 4

4. Machine learning assisted kriging to capture spatial variability in petrophysical property modelling

4.1. Summary

In this chapter, a ML LSSVR model is integrated into kriging interpolation to estimate the sonic compressional wave velocity with high spatial variability over a geological setting. This new methodology consists of two steps: (1) constructing an initial machine learning model using (LSSVR and calculating the residuals, and (2) applying kriging interpolation and gradually modifying the training data using kriging estimation to optimise the machine learning model. The proposed model effectively captures the non-linear relation between auxiliary and target variables along with the spatial autocorrelation of the data.

The content of this chapter has been published in the "Marine Petroleum Geology". The article is available as an open-access publication at:

https://doi.org/10.1016/j.marpetgeo.2024.106967.

4.2. Introduction

The use of machine learning (ML) techniques to develop models for predicting geophysical and petrophysical properties has gained a significant attention in recent years. These techniques have been used to predict a range of petrophysical properties such as permeability, lithofacies, and stratigraphy as well as other petrophysical properties from geophysical logging data (Lv et al 2021; Al Khalifah et al. 2020; Farouk et al. 2021; Zhou et al. 2021; Handhal et al. 2020; Zhang et al. 2021). Artificial neural network, support vector regression, fuzzy logic and Gaussian processes have been used to capture linear and non-linear correlations between petrophysical properties and geophysical logging data (Rajabi et al. 2010; Asoodeh and Bagheripour 2012; Onalo et al. 2018; Anemangely et al. 2019; Mehrgini et al. 2019; Wang and Peng 2019; Al-Dousari et al. 2016; Behnia et al. 2017; Abbas et al. 2018; He et al. 2019, Saporetti et al. 2022). Previous ML-based studies primarily focused on improving the ability of models' prediction through identifying the most effective algorithms (Asoodeh and Bagheripour 2012; Wang and Peng 2019); optimizing input parameters using feature selection techniques (Tan et al. 2015; Al-Dousari et al. 2016; Behnia et al. 2017; Xu et al. 2017; Zhang et al. 2018; Anemangely et al. 2019) and integrating rock physics models to mitigate limitations inherent in machine learning algorithms (Lv et al, 2021; Saad et al, 2018; Zhao et al, 2024).

These ML-based models are usually not designed to capture spatial variabilities thus do not consider the spatial autocorrelation between data. The spatial autocorrelation defines as “data from locations near one another in space are more likely to be similar than data from locations farther from each other” (Hashemi et al 2016; Cressie 2015; Santibanez et al. 2015). Current ML models treat spatial data similar to regular data without autocorrelation. Taking sole location and time as features in model training is not the best way of incorporating result of autocorrelation and it leaves prediction residual out of the analysis (Shekhar et al. 2010).

When the examined variable is spatially autocorrelated, the assumption of independency among observations in the cross-validated data is violated (i.e. the fitted model uses almost identical data for training and testing). Recent studies on ML modelling have revealed that when using the conventional non-spatial random k-fold cross-validation, the predictive performance tends to be overly optimistic in the presence of spatial autocorrelation within the data. The extent of spatial overfitting varies depending on factors such as the level of autocorrelation among training points, environmental similarities across regions and specific ML technique employed (Roberts et al., 2017; Meyer et al., 2018; Ploton et al., 2020). To mitigate the impact of autocorrelation, various cross-validation strategies were proposed. These include incorporating buffer distances between training locations, employing successive distances, leaving out one training location at a time and utilizing spatial block training. Spatial blocks can be defined through techniques such as geographical coordinates clustering, latitudinal-blocking, systematic assignment and clustering based on environmental similarity. Moreover, autocorrelation also affects ML feature selection methods and hyperparameter optimization leading to suboptimal variable selection and model parameterization (Meyer et al, 2019; Schratz et al, 2019; Gazis and Greinert 2021).

Deterministic geostatistical interpolation methods such as ordinary and universal Kriging have been widely recognized as suitable approaches for modelling spatial autocorrelations and spatial variable estimation in land resource inventories, remote sensing technologies, soil surveyor mapping and mapping annual participation (Amini et al. 2019; Da Silva et al. 2019; Zhang et al. 2018; Mitas and Mitasova, 1999; Hengl et al. 2003). Spatial autocorrelation, a local statistical behaviour, can be replicated by the nugget effect in simple or ordinary Kriging. However, these methods for non-stationary data may not be as effective as stationarity owing to its key assumption of ordinary Kriging. Spatial stationarity implies that the mean or variance of data remains relatively constant across space (Bohling, 2007). These methods are also unable to extract information from all available measurements, thus cannot establish links between different source measurements (Chiles and Delfiner, 2012; Kalambukattu et al. 2018). One the other hand, hybrid interpolation techniques, such as regression Kriging are able to model non-stationary spatial variable data while using supplementary information (Odeh et al., 1995; Hengl et al., 2007; Han et al., 2020). Regression Kriging uses linear regression to capture the correlation between the target variable and any available auxiliary variable and then employs simple Kriging to interpolate the residual from the developed linear regression model. Employing regression Kriging and auxiliary data has shown improvement in prediction accuracy compared to sole linear regression or Kriging (Odeh et al., 1994, 1995; Goovaerts, 1999b; Bishop and McBratney, 2001). For example, using elevation data as an auxiliary variable in temperature mapping with Kriging has revealed to improve the accuracy of temperature predictions (Hudson and Wackernagel, 1994); or adding a slope map as an explanatory variable has been found to enhance horizon thickness estimation of formations (Bourennane et al. 2000, 2010). While regression Kriging can add extra value to a sole Kriging method, it is mostly desirable when a linear correlation between auxiliary and target variable exists.

In the literature, machine learning algorithms such as artificial neural networks and support vector regression have been employed to overcome the linear assumption in regression Kriging when estimating climate data (Demyanov et al., 1998), wind speed (Chung et al., 2012), precipitation (Seo et al., 2015), soil properties (Tarasov, et al.2017), soil heavy metals (Sergeev et al, 2019) and shale volume (Li et al., 2019). Hybrid models that combine machine learning and Kriging algorithms claim to leverage the advantages of both approaches. Compared to regression Kriging, which assumes linearity between auxiliary variables and the target parameter, these hybrid models utilize a more robust tool (machine learning) to capture non-linearity, resulting in higher prediction accuracy. Typically, hybrid models involve two main steps: (1) training and constructing a regression machine learning model while calculating the residuals, and (2) employing residual Kriging interpolation. Although the second step is somewhat dependent on the first step, they operate independently and machine learning treats the data without considering spatial variability.

To consider the spatial variability in the integrated model, we propose a new methodology that enhances the performance of machine learning training by incorporating Kriging residual estimation. The methodology consists of two steps: (1) constructing an initial machine learning model using Least Squared Support Vector Regression (LSSVR) and calculating the residuals, and (2) applying Kriging interpolation and gradually modifying the training data using Kriging estimation to optimise the machine learning model. This iterative process aims to improve the accuracy of LSSVR and hybrid models simultaneously. We collected a set of geophysical downhole logging measurements including density, gamma ray, depth and sonic velocity from 122 boreholes in a mine site located at the Bowen basin in Queensland, Australia. Similar to conventional machine learning model development, 85% of data were allocated for training and the remaining 15% used for model testing (Rajabi et al., 2010; Asoodeh and Bagheripour, 2012; Al-Dousari et al., 2016; Anemangely et al., 2019, Karimpouli and Fattahi, 2018). The modelling process involved several iterations, where each iteration divided the training data into two sets. The first set, consisting of 80% of training data, was used to develop the LSSVR model, capturing the correlation between gamma ray, density, depth, coordinates and the target parameter (sonic compressional wave velocity). The residuals were then calculated at each point to model the semivariogram and perform Kriging interpolation. In the subsequent step, the residuals were estimated for the remaining 20% of training data by applying Kriging interpolation to the validation points. These residuals (or errors) were then used to modify the actual sonic measurements at the validation points. At this stage, the updated validation set and the original training set were combined, initiating a new iteration.

4.3. Data Availability and preparation

All the data were collected from a coal mine located at the Bowen basin in Queensland, Australia. Such a basin that spans the Permian-Triassic period holds Australia's largest coal reserves and houses 44 operational coal mines (Mutton, 2003; Maxwell et al., 2019). The boreholes examined in this study primarily intersect geological formations from the Rangal Coal Measures and Fort Cooper Coal Measures. These formations comprise three main lithology types: sandstone/silty sandstone, siltstone and coal. In many areas, the sandstone layer lies above coal seam, from which in-situ stresses are estimated. This underscores the significance of characterizing the sandstone formation above coal seam.

Our dataset consisted of 122 boreholes, located in an area approximately 6 km x 2 km. Each borehole contains measurements of density, gamma ray, sonic P-wave velocity (V_P) with depth along with their x and y coordinates. As shown in Figure 4.1, the boreholes were primarily distributed from the northwest to the southeast of sampled area. Measurements for gamma ray, bulk density and VP were taken over an approximately 40-meter depth interval in each borehole, encompassing several sandstone and coal layers. However, for the purpose of this investigation, we selected and extracted data associated with the immediate sandstone layer above the coal layer. To ensure accuracy, we visually inspected density-velocity cross plots, considering that coal typically exhibits lower density compared to sandstones. This data was subsequently averaged over the measured depth in each borehole, as the spatial variability was only modelled in 2D. The averaging technique was applied under the assumption that the data within the selected interval was rather homogeneous. As a result, our dataset consisted of 122 points for each measurement and all of them were used for modelling efforts. Figure 4.1 illustrates the distribution of each feature across the sampled area highlighting that depth (see Figure 4.1 is a non-stationary autocorrelated (low nugget effect) variable. Bulk density and gamma-ray present similar statistical behaviour, although some high values appear along the south-west edge due to some heterogeneity in this area.

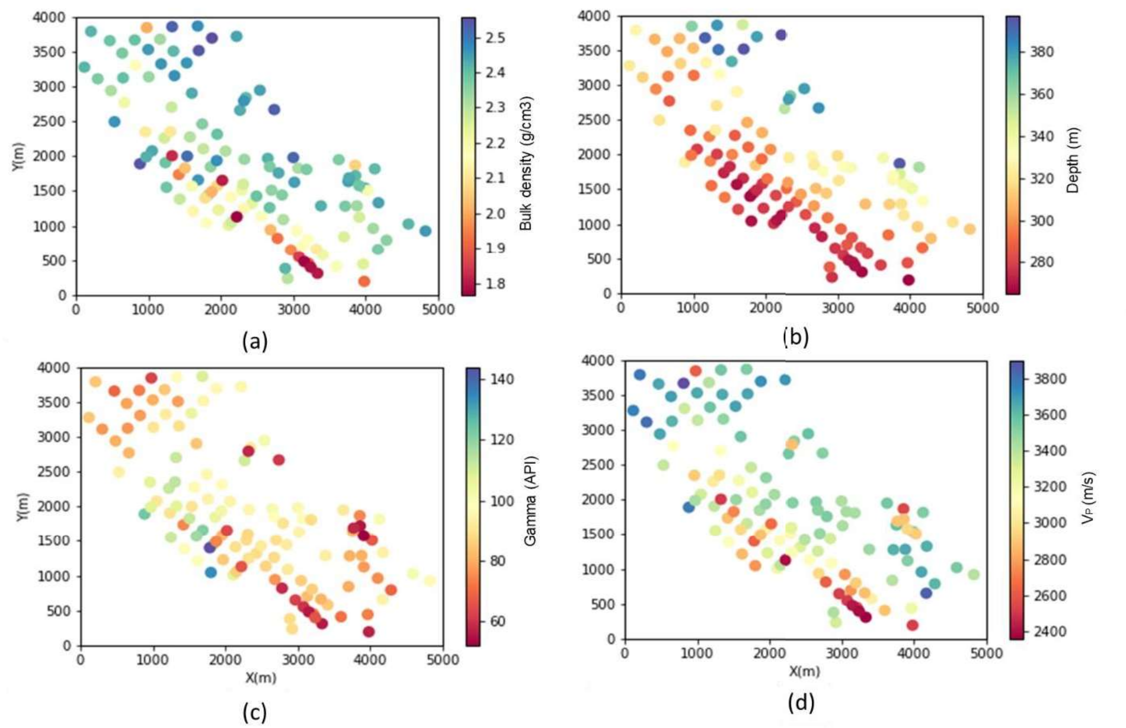


Figure 4.1. Averaged features including (a) bulk density, (b) depth, (c) gamma ray and (d) V_P in 2D space.

The cross plots of available measurements and V_P are shown in Figure 4.2 demonstrating strong correlation among variables. Bulk density, gamma ray, depth, X and Y coordinates were selected as input parameters for estimating V_P. Before training the predictive models, we standardized the data using a separate mean and standard deviation for each variable. This is necessary as the variables have different ranges and distributions. We used Equation (4.1) to standardize the data according to:

$$z = \frac{x - \mu}{SD} \tag{4.1}$$

where x is the log value, μ and SD are the mean and standard deviation of the data, respectively and z represents the standardized value. We developed four separate models using linear regression (LR), LSSVR, regression Kriging (RK) and the hybrid LSSVR-RK algorithm for estimating V_P. We explained the details of the LSSVR-RK algorithm in Section 4.4.2 due to its similarity in the training procedure. Finally, we evaluated and compared all four models to estimate V_P in the test locations.

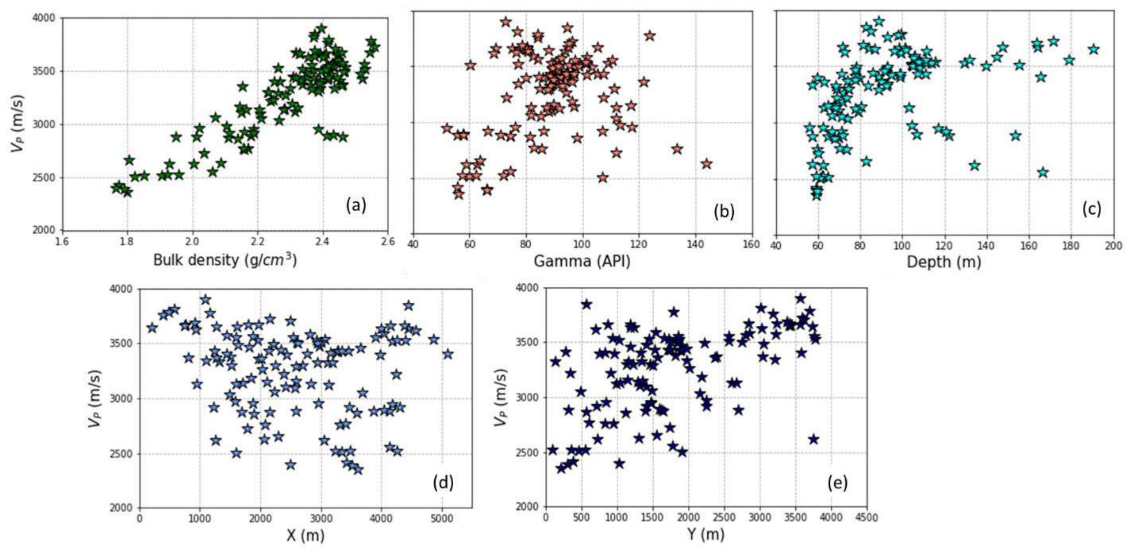


Figure 4.2. Cross plots of input and output features used for modelling: (a) bulk density vs V_P, (b) gamma ray vs V_P, (c) depth vs V_P, (d) X coordinate vs V_P and (e) Y coordinate vs V_P.

4.4. Methods

4.4.1 Hybrid model structure

Regression Kriging is a technique that combines linear regression with simple Kriging. It is important to note that regression Kriging relies on the assumption of linearity between input and target variables, which may not hold in all cases. On the other hand, the literature presents various machine learning algorithms for estimating petrophysical properties, with support vector regression (SVR) and neural networks standing out as particularly effective in geoscience applications (Mohammadpour et al., 2023; Cranganu and Breaban, 2013; Bagheripour et al., 2015; Olayiwola and Sanuade, 2021). Due to the limited features in rock physics studies, ML algorithms such as SVR capable of handling high variance are preferred for efficient learning and predictions. In this study, we chose LSSVR, an enhanced version of SVR, due to its improved computational efficiency (Wang and Hu, 2005). It is worth noting that while other ML methods may be suitable for regression analysis, our focus remains on the application of LSSVR to be within the scope of our study.

In the proposed hybrid model, we aim to generalize the linear assumption of regression Kriging by incorporating the LSSVR algorithm which is capable of capturing non-linearities between auxiliary and target variables. To optimize the accuracy of the LSSVR model, we employed an iterative process, utilizing Kriging estimation. In our proposed methodology, Kriging of residuals acts as a guiding mechanism for the LSSVR model to capture the variability and relationships between input and target variables. The step-by-step procedure for developing the hybrid model is outlined below and also shown in Figure 4.3:

1. **Data Split:** Dividing the dataset randomly into three sets for training, validation and testing: (X_{train}, Y_{train}), (X_{val}, Y_{val}), and (X_{test}, Y_{test}) where X and Y represent input and target variables, respectively;

2. **LSSVR Training:** Training the LSSVR model using the training set (X_{train}, Y_{train});

3. **Training Predictions:** Using the trained LSSVR model to make predictions on the training data: $Y_{pred\_train} = LSSVR(X_{train})$;

4. **Residual Calculation:** Calculating the residuals for the training data: $Res_{train} = Y_{train} - Y_{pred\_train}$;

5. **Semivariogram Development:** Developing a semivariogram using the residuals from step 4;

6. **Validation Predictions:** Employing the developed LSSVR model to estimate values for the validation set: $Y_{pred\_val} = LSSVR(X_{val})$;

7. **Kriging for Error Estimation:** Utilizing Kriging and the previously developed semivariogram to estimate errors/residuals for data validation: Res_{Val} = Kriging (X_{val}, semi-var);

8. **Integrating Prediction:** Calculating hybrid predictive models for the validation set: $Y_{hybrid\_val} = Y_{pred\_test} + Res_{val}$. Evaluating the model's performance, e.g. using R^2, RMSE or MAE (Y_{hybrid_val}, Y_{val});

9. **Check for Satisfactory Results:** If the model performance is satisfactory, proceed to step 13;

10. **Error-Based Data Modification:** Modifying data validation based on the estimated errors: $(X_{val}, Y_{val}) \rightarrow (X_{val}, Y_{val} - Res_{val})$;

11. **Data Fusion:** Combining the training data (X_{train}, Y_{train}) with the modified validation set $(X_{val}, Y_{val} - Res_{val})$;

12. **Data Re-split:** Dividing the merged dataset into new testing and validation sets and return to step 1;

13. **Final Testing Predictions:** Calculating the hybrid predictive models for testing set: $Y_{hybrid\_test} = Y_{pred\_test} + Res\_{test}$. Evaluating the model's performance, e.g. using R^2 $(Y_{hybrid\_test}, Y_{test})$ and RMSE or MAE $(Y_{hybrid\_test}, Y_{test})$.

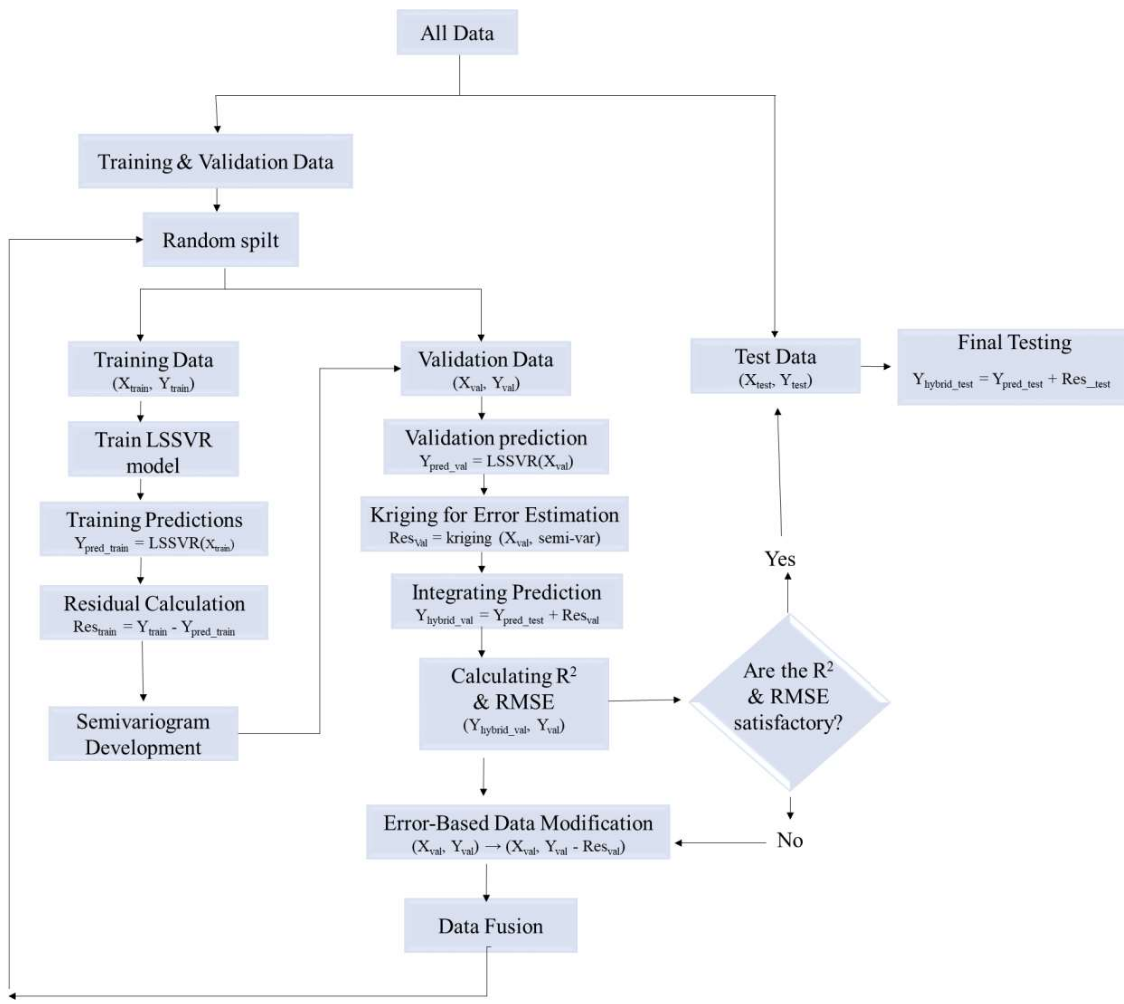


Figure 4.3. Flowchart for developing the proposed model

4.4.2 Least square Support vector regression (LSSVR)

LSSVR is an improved version of the SVR algorithm. It differs from SVR in its objective and restriction functions (Suykens et al. 2002). By solving a linear equation instead of a quadratic one, LSSVR reduces the computational cost and complexity of regression analysis (Wang and Hu 2005). The objective function in LSSVR is defined as follows:

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2 \qquad \text{Subject to } y_i = w^T.F(x) + b + e_i \tag{4.2}$$

In Equation (4.2) , F is a non-linear function that maps the input into a higher dimensional feature space, w is the weight factor, e_i represents the error in each data point, and γ is the regularisation parameter. The Lagrange multiplier is used to find the optimum values:

$$L(\omega, b, e; \alpha) = J(w, e) - \sum \alpha_k\{w^T.F(x_k) + b + e_k - y_k\} \tag{4.3}$$

where the derivatives of Lagrange multiplier are determined as:

$$\begin{aligned}
&\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^{N} \alpha_k\{w^T.F(x_k)\\
&\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^{N} \alpha_k = 0\\
&\frac{\partial L}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_{k,} \qquad k = 1, \dots, N\\
&\frac{\partial L}{\partial \alpha_k} = 0 \rightarrow w^T.F(x_k) + b + e_k - y_k = 0, \ \ k = 1, \dots, N
\end{aligned} \tag{4.4}$$

By solving the above equations and applying Mercer's condition, LSSVR can be used for nonlinear functions. The estimation function in LSSVR is given by

$$y_i = \sum_{i=1}^{l} \alpha_i K(x.x_i) + b \tag{4.5}$$

where α_i and K denote the support vectors and kernel functions, respectively.

4.4.3 Regression Kriging

RK or Kriging interpolation (after detrending) is a technique which considers both correlation with auxiliary variable and spatial variables for prediction. While simple Kriging only uses the target variable observation at sampled locations, RK employs other features measured in the sampled location. In RK, spatial properties at unsampled location are considered to include two parts: the predicted drift and the residual. Therefore, $V(z_0)$ can be considered as the velocity of sonic wave at unsampled location $z_0(x_0, y_0)$ and $V(z_1), V(z_2), V(z_3), \ldots, V(z_n)$ as measurements at locations $z_1, z_2, z_3, \ldots, z_n$. $V(z_0)$ can be assumed as a combination of trend or external drift $\widehat{m}(z_0)$ and residual $\hat{e}(z_0)$ according to:

$$V(z_0) = \widehat{m}(z_0) + \hat{e}(z_0) \tag{4.6}$$

where, $\widehat{m}(z_0)$ is the trend or correlation between V and available auxiliary variables which is commonly fitted using a linear regression analysis while $\hat{e}(z_0)$ is the residual component calculated from simple Kriging. RK assumes that the correlation between the unknown and auxiliary variable is linear and thus:

$$V(z_0) = \sum_{k=0}^{p} \mathcal{B}_k . q_k(z_0) + \sum_{i=0}^{n} w_i(z_i) . e(z_i) \tag{4.7}$$

$$q_k(z_0) = 1 \tag{4.8}$$

where B_k is the estimated linear regression coefficient, $q_k(z_0)$ is the k_{th} auxiliary variable at location z_0, p is the number of predictors, $w_i(z_i)$ is the calculated weight using semivariogram function and $e(z_i)$ is the regression residual. Eq. (7) can be presented in matrix notation as follows:

$$V(z_0) = {q_0}^T \mathcal{B} + \lambda_0^T . e \tag{4.9}$$

Where similar to what mentioned about equation 4.8, q_0 is vector of predictors at z_0, $\mathcal{B}$ is vector of estimated coefficients, λ_0 is vector of kriging weights and e is vector of residuals. The drift model's coefficients can be estimated in an iterative process using generalised least squares (GLS) method. Therefore, the drift model should be determined using ordinary least squares method in the first step; then the covariance (semivariance) of the residual is used to calculate the GLS coefficients according to:

$$\hat{\beta}_{GLS} = (q^T . C^{-1} . q)^{-1} . q^T . C^{-1} . \tag{4.10}$$

Where C is the matrix of residual covariance as below:

$$C = \begin{bmatrix} C(s_1, s_1) & \cdots & C(s_1, s_n) \\ \vdots & \ddots & \vdots \\ C(s_n, s_1) & \cdots & C(s_n, s_n) \end{bmatrix} \tag{4.11}$$

The resulting GLS coefficients can be then used to re-compute the residual and so on.

4.4.4 Nested semivariogram model

To use Kriging, we must switch from using the empirical semivariogram to a semivariogram model. This change is necessary because the Kriging algorithm requires access to semivariogram values across various distances, and the chosen model must meet specific numerical criteria for the Kriging equations to work. There are several common techniques for semivariogram modelling including exponential, spherical and Gaussian. Figure 4.4 provides an example of these three semivariogram models. Each semivariogram has three key characteristics: sill, range and nugget. Sill represents the semivariance values at which the variogram levels are off. Range indicates the distance at which the semivariogram reaches the sill value or where the autocorrelation becomes negligible. The nugget represents variability at distances smaller than the sample spacing, typically attributed to measurement errors and while theoretically, the nugget should be zero, in practice, it can be higher.

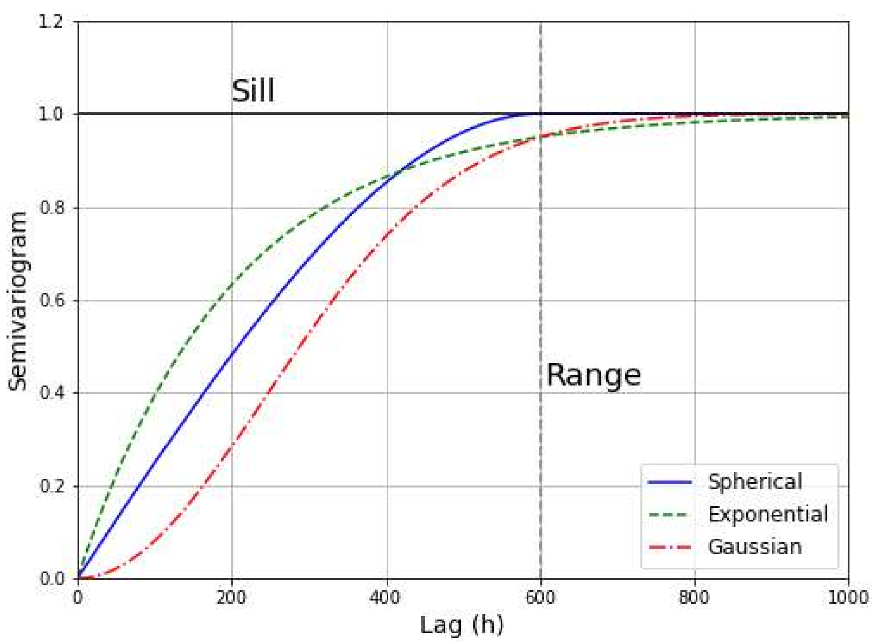


Figure 4.4. Schematic representation of common seamivariogram models.

To make the hybrid model complete, the ability to fit semivariogram and predict data trends is needed. Nested semivariogram model is a powerful technique which has been used in geostatistics to model spatial variability in a dataset. Such an approach involves the combination of two or more basic semivariogram models to create a more complex model that captures the nested structure of data. The models used in the nested semivariogram approach can be chosen based on specific properties of data and can include exponential, spherical and Gaussian models (Webster and Oliver, 2007).

$$\gamma(h) = \sum_{i=1}^{nst} \Gamma_i(h) \tag{4.12}$$

where is the nested semivariogram model resulting from the summation of each semivariogram.

Each variogram structure, denoted as, is defined by a geometric anisotropy model that includes the orientation and range of the major and minor directions. In two-dimensional (2D) models, the geometric anisotropy model is represented by an azimuth and ranges, azi, $amaj$, and $amin$. It is worth noting that the minor range is orthogonal to the major range direction. The geometric anisotropy model assumes that the range in all off-diagonal directions is elliptical in shape, with the major and minor axes aligned with and set to the major and minor ranges of the variogram accordingly as formulated below:

$$h_i = \sqrt{(\frac{r_{maj_i}}{a_{maj_i}}) + (\frac{r_{maj_i}}{a_{maj_i}})} \tag{4.13}$$

4.5. V_P predictive model development

4.5.1 Developing the hybrid LSSVR-RK model

To assess the effectiveness of proposed methodology detailed in Section 4.4.1, which combines the advantages of LSSVR and RK, we developed a predictive model for V_P using measurements of bulk density, porosity, gamma ray as well as X and Y coordinates. Following steps outlined in Section 4.4.1, we partitioned our dataset into 81 data points for training, 20 for validation and 20 for testing purposes.

The proposed methodology involves an iterative process to optimize the LSSVR model for predicting the target variable. At each iteration, a separate LSSVR and semivariogram model must be fitted to the target datasets. For training the LSSVR model, we selected the radial basis function (RBF) from among several available kernel functions. We also performed random searches within the parameter ranges of [0.01, 10] for regularization and [0.01, 50] for RBF kernel parameter to generate the most accurate predictive model, as suggested by Suykens and

Gestel (2003). The regularization and kernel parameters determined as (1.33, 0.3) and (1, 0.2) for LSSVR and hybrid model respectively. It is essential to recognize that these parameters govern the model's learning capability, predictive accuracy and generalization potential, as noted by Anemangely et al. (2019). It is essential to consider the role of regularization and kernel parameters in LSSVR models. Regularizations plays a crucial role in balancing the trade-off between model complexity and generalization performance. A smaller value of regularization parameter encourages a larger margin and results in a simpler decision boundary, which helps prevent overfitting by penalizing large coefficients in the model. Conversely, a larger regularization parameter value allows the model to fit the training data more closely but may lead to overfitting. On the other hand, the choice of kernel parameters in the Radial Basis Function (RBF) kernel, influences the flexibility and complexity of the decision boundary. A higher kernel parameter value leads to more complex decision boundaries that can capture intricate patterns in the data, potentially resulting in overfitting.

While the LSSVR model captures relationships between V_P and the five auxiliary variables (bulk density, gamma ray, depth, X and Y coordinates) to a certain extent, we acknowledge the possibility of an additional spatial variable term not to be captured by the statistical regression alone. We assumed that the residual of LSSVR model represents a combination of these spatial terms. Thus, similar to the concept of regression Kriging, we modelled the spatial term using Kriging, utilizing LSSVR predictions' residuals at all training points.

For the development of Kriging models, we calculated semivariograms in each iteration, based on certain assumptions. 45 and 135 degrees directions were considered as the two principal directions of spatial continuity, with lag spacing determined by data spacing along major and minor directions. The lag tolerance was set equal to the lag distance and an angle tolerance of approximately 22.5 degrees was applied, following common practice (Pyrcz & Deutsch 2014). Semivariogram modelling was automated, employing a nested model that combined two semivariogram models linearly. The selection of critical parameters, including ranges, model types (spherical, exponential, or Gaussian), nugget values and weights for each structure, was accomplished through an iterative method known as "sequential least squares programming" (Kraft 1994). Such a method ensures precise fitting of empirical semivariograms, allowing the variogram model to accurately represent underlying spatial dependent structure in the data.

According to the proposed methodology, the iterative process continues until a satisfactory result is achieved, as determined by the calculated R^2 and RMSE for the validation data (step

8, Section 4.4.1). In our modelling approach, we chose to halt the iteration when R^2 showed signs of stabilization, specifically when the difference between the R^2 of two consecutive iterations fell below 0.001. Based on this criterion, the validation R^2 reached a stable value after 10 iterations, yielding an R^2 value of 0.930 and RMSE of 0.28. The results of R^2 and RMSE for each iteration are depicted in Figure 4.5.

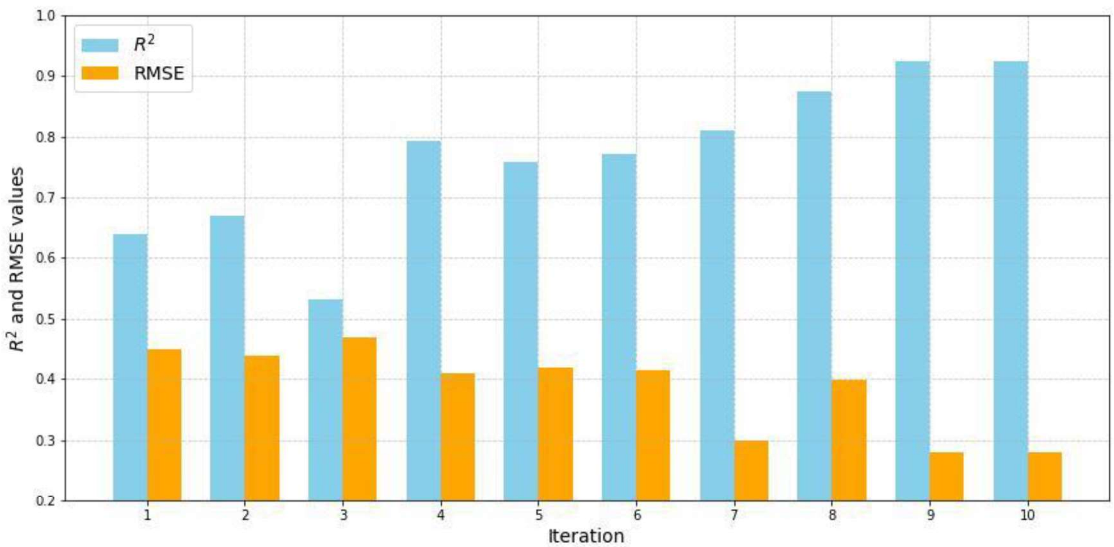


Figure 4.5. The R^2 and RMSE values resulted from several iterations.

As semivariogram modelling was performed automatically in each iteration; ten separate models were internally fitted to the empirical semivariograms within the model. The final empirical semivariogram and the fitted model are presented in Figure 4.6, showing semivariograms under major and minor directions.

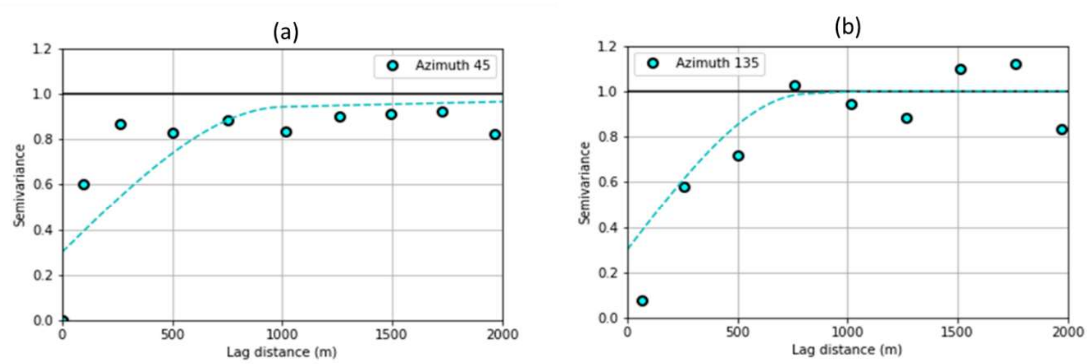


Figure 4.6. Empirical and fitted semivariograms under (a) major (Azimuth 45) and (b) minor (Azimuth 135) directions.

4.5.2 Model evaluation

To evaluate the model's performance, we trained three additional models: LR, LSSVR, and RK aiming to assess the hybrid model's performance relative to each individual algorithm. Figure 4.7 compares the predictions resulted from all four models. Utilizing testing data unseen by any of the models, we evaluated their performance based on predicted values for the target parameter at test points and determined R^2 scores for comparative analysis.

The linear regression model (Figure 4.7a) exhibited the least favourable performance with an R^2 of 0.73. This improved to 0.81 with the adoption of LSSVR, a logical outcome given the non-linear relationship between inputs and target parameters in our study (see Figure 4.2). LSSVR's enhanced ability to capture such non-linear relationships was evident. Similarly, the model combining linear regression and either Kriging or RK (Figure 4.7b) demonstrated improved performance compared to the sole linear regression model. This highlights the significance of considering spatial changes among input and target variables. Ultimately, the hybrid model, an optimized LSSVR model guided by Kriging, emerged as the most accurate performer among the evaluated models with R^2 of 0.93 outperforming all other models.

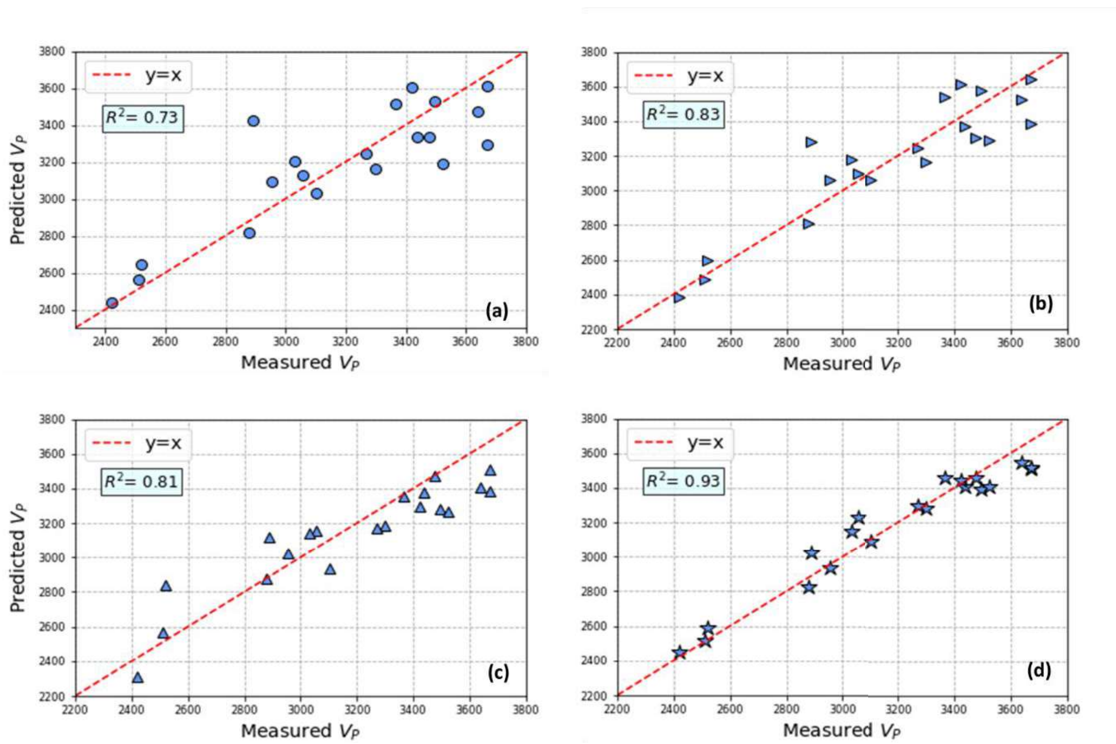


Figure 4.7. Comparing the predictions resulted from the developed models with the actual measurements including (a) LR, (b) RK, (c) LSSVR and (d) hybrid model.

4.6. Conclusion

The utilization of machine learning algorithms has gained popularity in recent years for developing predictive models in geophysical and petrophysical studies. However, the previous machine learning models did not account for the spatial autocorrelation between data. Geostatistical interpolation methods, such as simple Kriging, have been used to model spatial autocorrelations and spatial variable estimation, but they cannot extract valuable information from other available measurements. Hybrid interpolation techniques, such as regression Kriging (RK), can improve accuracy by using linear regression (LR) to capture the correlation between the target variable and any available auxiliary variable and then interpolating the residual from the developed LR model. In this study, a hybrid model inspired by RK was developed through coupling Least Squared Support Vector Regression and simple Kriging to capture the non-linearities between data sources and estimate sonic compressional wave velocity as accurate as possible. The hybrid model effectively captured both the correlation between auxiliary and target variables and the spatial autocorrelation of data. The developed model was tested on a dataset consisting of 122 boreholes, each containing measurements for bulk density, gamma ray, depth and sonic P-wave velocity from a coal mine located at the Bowen Basin of Queensland, Australia. The developed hybrid model was found to outperform compared to other models, demonstrating its effectiveness in geophysical and petrophysical studies dealing with spatial variability and non-linearity.

4.7. References

Abbas, A. K., Flori, R. E., & Alsaba, M. (2018) Estimating rock mechanical properties of the Zubair shale formation using a sonic wireline log and core analysis. Journal of Natural Gas Science and Engineering, 53, 359-369.

Al-Dousari M, Garrouch A A, Al-Omair O (2016) Investigating the dependence of shear wave velocity on petrophysical parameters. J Petrol Sci Eng 146:286-296

Al Khalifah, H., P. W. J. Glover, and P. Lorinczi (2020). Permeability prediction and diagenesis in tight carbonates using machine learning techniques. Mar Pet Geol 112 :104096.

Amini, M.A., et al, (2019). Analysis of deterministic and geostatistical interpolation techniques for mapping meteorological variables at large watershed scales. Acta Geophysica, 67 (1), 191–203.

Anemangely M, Ramezanzadeh A, Amiri H, Hoseinpour S-A (2019) Machine learning technique for the prediction of shear wave velocity using petrophysical logs. J Petrol Sci Eng 174:306-327

Asoodeh M, Bagheripour P (2012) Prediction of compressional, shear, and stoneley wave velocities from conventional well log data using a committee machine with intelligent systems. Rock Mech Rock Eng 45(1):45-63

Al-Dousari M, Garrouch A A, Al-Omair O (2016) Investigating the dependence of shear wave velocity on petrophysical parameters. J Petrol Sci Eng 146:286-296

Behnia D, Ahangari K, Moeinossadat S R (2017) Modeling of shear wave velocity in limestone by soft computing methods. Int J Min Sci Technol 27(3):423-430

Bishop, T. F. A., and A. B. McBratney. "A comparison of prediction methods for the creation of field-extent soil property maps." *Geoderma* 103, no. 1-2 (2001): 149-160.

Bagheripour, P., Gholami, A., Asoodeh, M., Vaezzadeh-Asadi, M., 2015. Support vector regression-based determination of shear wave velocity. J. Pet. Sci. Eng. 125, 95–99.

Bourennane, Hocine, Dominique King, and Alain Couturier. "Comparison of Kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities." *Geoderma* 97, no. 3-4 (2000): 255-271.

Bourennane, Hocine, Francis Douay, Thibault Sterckeman, Estelle Villanneau, Henri Ciesielski, Dominique King, and Denis Baize (2010) Mapping of anthropogenic trace elements inputs in agricultural topsoil from Northern France using enrichment factors. Geoderma 157, no. 3-4: 165-174

Bohling, G. C. Introduction to geostatistics (2007) Kansas Geological Survey Open File Report 26: 50.

Chung, C. H., Chiang, Y. M., & Chang, F. J. (2012). A spatial neural fuzzy network for estimating pan evaporation at ungauged sites. Hydrology and Earth System Sciences, 16(1), 255-266.

Chiles, J. P., & Delfiner, P. (2012). Geostatistics: modeling spatial uncertainty (Vol. 713). John Wiley & Sons.

Cressie, Noel. Statistics for spatial data. John Wiley & Sons, 2015.

Cranganu, C., Breaban, M., 2013. Using support vector regression to estimate sonic log distributions: a case study from the Anadarko Basin, Oklahoma. J. Pet. Sci. Eng. 103, 1–13.

Da Silva, A.S.A., et al., (2019) Comparison of interpolation methods for spatial distribution of monthly precipitation in the State of Pernambuco, Brazil. Journal of Hydrologic Engineering, 24 (3), 04018068.

Demyanov, V., Kanevsky, M., Chernov, S., Savelieva, E., & Timonin, V (1998) Neural network residual Kriging application for climatic data. Journal of Geographic Information and Decision Analysis, 2(2), 215-232.

Farouk S, Sen S, Ganguli S S, Abuseda H, Debnath A (2021) Petrophysical assessment and permeability modeling utilizing core data and machine learning approaches–a study from the Badr El Din-1 field, Egypt. Mar Pet Geol 133:105265

Gazis, I.Z. and Greinert, J., (2021) Importance of spatial autocorrelation in machine learning modeling of polymetallic nodules, model uncertainty and transferability at local scale. Minerals, 11(11), p.1172.

Goovaerts, P. (1999) Geostatistics in soil science: state-of-the-art and perspectives. Geoderma, 89(1-2), 1-45.

Handhal, A.M., Al-Abadi, A.M., Chafeet, H.E. and Ismail, M.J (2020). Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. Mar Pet Geol, 116, p.104347.

Han, Feilong, Hongbing Zhang, Jianwen Rui, Kuiye Wei, Dailu Zhang, and Wei Xiao (2020). Multiple point geostatistical simulation with adaptive filter derived from neural network for sedimentary facies classification." Mar Pet Geol 118 :104406.

Hengl, Tomislav, Gerard BM Heuvelink, and David G. Rossiter. "About regression-Kriging: From equations to case studies." Computers & geosciences 33, no. 10 (2007): 1301-1315.

Hengl, Tomislav, David G. Rossiter, and Alfred Stein. "Soil sampling strategies for spatial prediction by correlation with auxiliary maps." Soil Research 41, no. 8 (2003): 1403-1422.

Hudson, Gordon, and Hans Wackernagel. "Mapping temperature using Kriging with external drift: theory and an example from Scotland." *International journal of Climatology* 14, no. 1 (1994): 77-91.

Hashemi, M., & Karimi, H. A. (2020). Weighted machine learning for spatial-temporal data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 3066-3082.

He, Jianhua, Andrew D. La Croix, Jiahao Wang, Wenlong Ding, and J. R. Underschultz (2019). Using neural networks and the Markov Chain approach for facies analysis and prediction from well logs in the Precipice Sandstone and Evergreen Formation, Surat Basin, Australia. Mar Pet Geol 101: 410-427.

Kalambukattu, Justin George, Suresh Kumar, and R. Arya Raj (2018) Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. Environ. Earth Sci. 77: 1-14.

Karimpouli S, Fattahi H (2018) Estimation of P-and S-wave impedances using Bayesian inversion and adaptive neuro-fuzzy inference system from a carbonate reservoir in Iran. Neural Comput Appl 29(11):1059-1072

Kraft, Dieter. "Algorithm 733: TOMP–Fortran modules for optimal control calculations." ACM Transactions on Mathematical Software (TOMS) 20, no. 3 (1994): 262-281.

Lv A, Cheng L, Aghighi M.A, Masoumi H, Roshan H (2021) A novel workflow based on physics-informed machine learning to determine the permeability profile of fractured coal seams using downhole geophysical logs. Mar Pet Geol 131:105171.

Li, Xiuquan, Yile Ao, Shuang Guo, and Liping Zhu (2019) Combining regression Kriging with machine learning mapping for spatial variable estimation." IEEE Geoscience and Remote Sensing Letters 17, no. 1: 27-31.

Mehrgini B, Izadi H, Memarian H (2019) Shear wave velocity prediction using Elman artificial neural network. Carbonates and Evaporites 34(4):1281-1291

Mohammadpour, M., Roshan, H., Arashpour, M., & Masoumi, H. (2023). Effect of spatial variability of downhole geophysical logs on machine learning exercises. International Journal of Coal Geology, 277, 104333.

Mitas, Lubos, and Helena Mitasova. "Spatial interpolation." Geographical information systems: principles, techniques, management and applications 1, no. 2 (1999).

Meyer, H.; Reudenbach, C.; Hengl, T.; Katurji, M.; Nauss, T (2018) Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Softw., 101, 1–9.

Meyer, H.; Reudenbach, C.; Wöllauer, S.; Nauss, T. (2019) Importance of spatial predictor variable selection in machine learning applications—Moving from data reproduction to spatial prediction. Ecol. Modell., 411, 108815.

Maxwell K, Rajabi M, Esterle J (2019) Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. Int J Coal Geol 214:103284

Mutton A J (2003) Queensland coals. Physical and chemical properties, colliery and company information.

Odeh, Inakwu OA, A. B. McBratney, and D. J. Chittleborough. "Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging." Geoderma 67, no. 3-4 (1995): 215-226.

Onalo D, Adedigba S, Khan F, James L.A, Butt S (2018) Data driven model for sonic well log prediction. J Petrol Sci Eng 170:1022-1037

Olayiwola, T., Sanuade, O.A., 2021. A data-driven approach to predict compressional and shear wave velocities in reservoir rocks. Petroleum. 7 (2), 199–208.

Pyrcz, M. J., & Deutsch, C. V. (2014). Geostatistical reservoir modeling. Oxford University Press, USA.

Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N.; et al. (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 2020, 11, 4540

Rajabi M, Bohloli B, Ahangar E.G (2010) Intelligent approaches for prediction of compressional, shear and Stoneley wave velocities from conventional well log data: A case study from the Sarvak carbonate reservoir in the Abadan Plain (Southwestern Iran). Comput Geoci 36(5):647-664

Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.;huiller,W.; et al. (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40, 913–929.

Saad, B., Negara, A. and Syed Ali, S (2018) Digital rock physics combined with machine learning for rock mechanical properties characterization. In Abu Dhabi International Petroleum Exhibition and Conference (p. D011S008R001). SPE.

Santibanez, Sebastian F., Marius Kloft, and Tobia Lakes (2015) Performance analysis of machine learning algorithms for regression of spatial variables. A case study in the real estate industry. In 13th International Conference of Geo Computation, Dallas, USA, pp. 20-23.

Seo, Youngmin, Sungwon Kim, and Vijay P. Singh (2015) Estimating spatial precipitation using regression kriging and artificial neural network residual kriging (RKNNRK) hybrid approach. Water Resources Management 29 : 2189-2204.

Schratz, P.; Muenchow, J.; Iturritxa, E.; Richter, J.; Brenning, A. (2019) Hyperparameter tuning and performance assessment of statisticaland machine-learning algorithms using spatial data. Ecol. Modell. 2019, 406, 109–120.

Sergeev, A.P., Buevich, A.G., Baglaeva, E.M. and Shichkin, A.V(2019) Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. Catena, 174, pp.425-435.

Saporetti, C. M., D. L. Fonseca, L. C. Oliveira, E. Pereira, and L. Goliatt (2022). Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields. Mar Pet Geol 143: 105783.

Shekhar, Shashi, Pusheng Zhang, and Yan Huang. Spatial data mining." Data mining and knowledge discovery handbook (2010): 837-854.

Suykens J.A, De Brabanter J, Lukas L, Vandewalle J (2002) Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48(1-4):85-105

Tarasov, D.A., Medvedev, A.N., Sergeev, A.P., Shichkin, A.V. and Buevich, A.G (2017) A hybrid method for assessment of soil pollutants spatial distribution. In AIP Conference Proceedings (Vol. 1863, No. 1). AIP Publishing

Tan, M., Song, X., Yang, X., & Wu, Q (2015) Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: A comparative study. J. Nat. Gas Sci. Eng, 26, 792-802.

Webster, Richard, and Margaret A. Oliver. Geostatistics for environmental scientists. John Wiley & Sons, 2007.

Wang P, Peng S (2019) On a new method of estimating shear wave velocity from conventional well logs. J Petrol Sci Eng 180:105-123

Wang H, Hu D (2005) Comparison of SVM and LS-SVM for regression. 2005 International conference on neural networks and brain, IEEE.

Zheng, Dongyu, Sixuan Wu, and Mingcai Hou (2021). Fully connected deep network: An improved method to predict TOC of shale reservoirs from well logs. Mar Pet Geol 132 :105205.

Zhou Y, Zhao X, Jiang C, Liu S, Han Z, Wang G (2021) Permeability prediction of multi-stage tight gas sandstones based on Bayesian regularization neural network. Mar Pet Geol 133:105320

Zhao, L., Liu, J., Xu, M., Zhu, Z., Chen, Y. and Geng, J (2024) Rock-physics-guided machine learning for shear sonic log prediction. Geophysics, 89(1), pp.D75-D87.

Zhang, M., Leon, C.D., and Migliaccio, K (2018) Evaluation and comparison of interpolated gauge rainfall data and gridded rainfall data in Florida, USA. Hydrological Sciences Journal, 63 (4), 561–582.

Xu J, Tan M, Wang X, Wu C (2017) Predicting acoustic-wave velocities and fluid sensitivity to elastic properties in fractured carbonate formation. Interpretation 5(1):SB69-SB80

Zhang, Long, Zetian Zhang, Ru Zhang, Mingzhong Gao, and Jing Xie (2021) The ultrasonic P-wave velocity-stress relationship and energy evolution of sandstone under uniaxial loading-unloading conditions. Adv. Mater. sci. eng. 2021: 1-11.

Chapter 5

5. Ultrasonic to sonic shear wave velocity conversion for rock mechanical characterisation using an integrated machine learning and physics

5.1. Summary

In this chapter, a new methodology was developed using a combination of AI and physical equations to predict the trend and magnitude of V_S from that of ultrasonic. The methodology involves training a predictive model based on LSSVR on data from 178 core samples including UV_P, porosity, and density as input and UV_S as output. The developed model was then used to predict the trend of SV_S using the same downhole geophysical input as those of laboratory measurements. Once the sonic shear wave velocity trend is obtained, a frequency dispersion physical model is used to convert the trend to actual values. The obtained result is then compared with available sonic shear wave velocity to assess the accuracy and robustness of the adopted methodology where a good agreement between predicted and measured velocities is observed. The results of this study confirm that the adopted methodology using a combination of AI and physics can be used to obtain accurate sonic shear wave velocity when laboratory measurements are available.

The content of this chapter has been submitted to "*International Journal of Rock Mechanics and Mining Sciences*" as a research paper.

5.2. Introduction

Accurate estimation of shear wave velocities is essential for various applications in rock engineering, hydrogeological, geophysical and geotechnical studies (Lv et al., 2022; Riedel et al., 2014, Pan et al. 2013, Wennberg et al. 2023). Measuring sonic wave velocity in mining engineering projects offers several significant benefits that enhance both the efficiency and safety of mining operations. Sonic wave velocity plays a crucial role in accurately characterizing the mechanical properties of rock masses, such as elasticity and strength, which are essential for predicting rock behaviour under various stress conditions. This detailed characterization improves geotechnical modelling, leading to more accurate predictions and better-designed mine structures. In addition, sonic wave velocity data helps optimize drilling and blasting operations, resulting in cost reductions, improved fragmentation, and minimized damage to surrounding rock. This optimization not only makes resource extraction more efficient but also ensures the stability of underground openings, slopes, and other mine structures by informing the design of safer support systems (Hatherly, 2013; Miah et al., 2020).

Combination of V_S andV_P wave velocities along with density can be used for the prediction of rock dynamic elastic properties (Li et al. , 2020; Benavente et al., 2020, Pappalardo and Mineo, 2022, Torabi et al., 2018), identifying lithology (Pickett, 1963; Chang et al., 2006;; Najibi et al., 2015; Parvizi et al., 2015; Elkatatny et al., 2018), seismic analysis and assessing overpressure condition (Domenico, 1984; Eastwood and Castagna, 1983; Johnston and Christensen, 1993; Potter et al., 1996; Hamada, 2004; Bailey and Dutton, 2012; Klarner and Klarner, 2012). While sonic shear wave velocity (slowness) is widely measured in oil and gas wells, it is rarely measured in mining industry due to involved cost and unawareness of its benefits (Hatherly, 2013; Hatherly et al., 2016; Karacan, 2009). In absence of sonic shear measurement, attempts have been made to estimate it indirectly using other geophysical logs (Mavko and Bandyopadhyay, 2009; Hossain et al., 2012) such as density, porosity and V_P (Han et al., 1986; Eberhart-Phillips et al., 1989; Castagna and Backus, 1993; Jørstad et al., 1999; Abbas et al., 2018; Khandelwal and Singh, 2009).

On the other hand, rock physical properties including density, porosity, V_P and V_S are often measured within the laboratory environment on retrieved core samples. The wave velocity measured using downhole sonic tool (e.g. dipole), however, has a much lower frequency (e.g. 10-15 kHz) compared to laboratory ultrasonic measurement (with a frequency range of about 100 kHz to 5 MHz). Different conditions and frequencies associated with sonic and ultrasonic measurements can result in different sonic velocities (SV_P, SV_S) compared to ultrasonic velocities (UV_P, UV_S). Similar to the measured field properties, a number of correlations have been proposed to correlate the laboratory based ultrasonic velocities to other rock physical properties such as density and porosity (He et al. 2021; Khandelwal, 2013; Sarkar et al., 2012). While both field and laboratory measurements capture the relationships between various rock physical properties and wave velocities (ultrasonic or sonic); their correlations have been found to be not identical due to dispersion and other environmental conditions such as pressure and temperature involved (Hossain et al., 2012).

With the advances in AI, various ML methods have been used to develop predictive correlations between petrophysical parameters and downhole geophysical logs, specifically, a number of sonic shear wave predictive models based on empirical or machine learning methods have been developed for sonic shear velocity prediction from other geophysical logs (Abbas et al., 2018; Onalo et al., 2018; Anemangely et al., 2019). Different ML methods including artificial neural network, support vector regression, fuzzy logic and Gaussian processes have been utilized to find such correlations (Mohammadpour et al., 2023; Rajabi et al., 2010;

Asoodeh and Bagheripour, 2012; Onalo et al., 2018; Anemangely et al., 2019; Mehrgini et al., 2019; Wang and Peng, 2019; Al-Dousari et al., 2016; Behnia et al., 2017; Al-Anazi and Gates, 2010).

The ultrasonic wave velocity is commonly a part of rock mechanical testing programs particularly in mining projects due to its non-destructive nature and low cost (Chawre, 2018), however, the sonic shear wave measurement is almost non-existence. While several empirical or ML methods have been used to link ultrasonic measurements to other rock properties (Khandelwal, 2013; Sarkar et al., 2012), the extension of ultrasonic to sonic wave velocity has not been conducted nor systematically investigated. To address this research gap, a number of laboratory measurements including porosity, density, UV_P and UV_S are collected from samples retrieved form three neighbouring mines located in Bowen basin, Queensland, Australia. Multivariable Linear Regression (MLR) and LSSVR algorithms are then used to develop predictive models for UV_S based on porosity, density and UV_P as input parameters. To develop an inclusive model, all laboratory input and output parameters are standardised using their mean and standard deviation values. 80% of data is used for the training and the remaining 20% for validation. The validated models are then employed to predict SV_S trend using the field measurements. A complete set of logging measurements extracted from a borehole in Bowen basin is utilised as the testing dataset. The standardised field properties including density, corrected neutron porosity, SV_P are used as input parameters for SV_S trend prediction. The resulting trend is then converted to actual values using a physical equation correcting for dispersion effect. Comparison between the predicted and the measured SV_S data demonstrated successful applicability of the laboratory-based developed model for field scale estimation.

5.3. Sonic log and ultrasonic measurement

Ultrasonic laboratory testing, borehole sonic logging, and seismic survey are three main techniques commonly used to measure wave velocities in geological formations across scales. While sonic and seismic measurements are performed at field-scale, ultrasonic measurement is performed within the laboratory environment (see Figure 5.1). It is noteworthy that the focus of this study is on sonic and ultrasonic measurements. A high amplitude wave pulse is emitted from a downhole tool's transmitter during sonic logging (e.g. dipole to measure both compressional and shear wave responses). This wave travels through drilling fluid, refracts to the formation, and is later received by the tool's receiver/s. The wave travel time (slowness) and velocity are then calculated (Serra, 1983; Kennedy, 2015).

Similar to sonic logging, the ultrasonic system measures the velocity of wave using transmitter/s and receiver/s on a core sample under laboratory conditions (Ersoy et al., 2020). The frequency associated with ultrasonic waves has a wide range often varying between 100

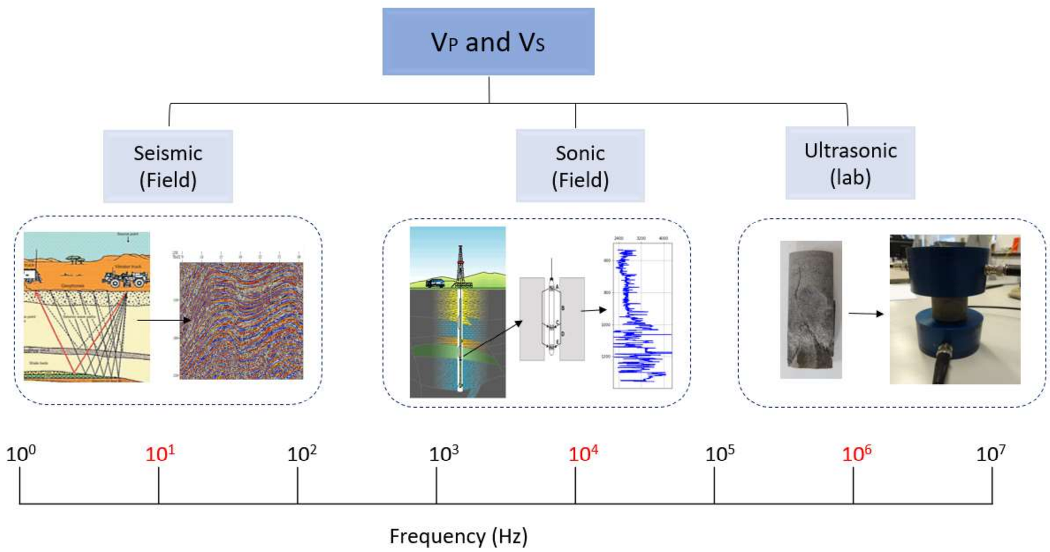


Figure 5.1. V_P and V_S measurements within seismic, sonic and ultrasonic frequencies (Aminzadeh et al 2013; Roshan et al 2016)

kHz to 5 MHz which is significantly higher than a typical range of sonic wave frequency (with an average frequency of 10-15 kHz) (Cheng and Johnston, 1981; Jizba, 1992; Fjær, 2009; Olsen et al., 2008). Ultrasonic systems are commonly used for estimation of rock elastic properties (e.g. dynamic Young's modulus and Poisson's ratio) using measured shear and compressional wave velocities and equation of motion (Takahashi et al., 2006). It has been reported that the measured wave velocities and estimated elastic properties by ultrasonic systems have higher values compared to the sonic logging measurements (Bennett et al., 1991, Elkatatny et al., 2018). Several mechanisms have been reported as potential causes for such differences including change in stress environment, temperature variation and different frequencies used in different measurement systems (Richardson et al., 1991). It is, however, noted that our analysis centres on Vs of water saturated rocks that is much less affected by the temperature (up to 60°C) and stress variation (up to 20 MPa) (Wang and Nur, 1988; Zhang et al., 2021). Therefore, the frequency dependency should still be the main controlling factor affecting wave velocity between different acoustic tools having different frequencies. While the ultrasonic wave (Howarth, 1984; Jizba, 1992; Olsen et al., 2008; Elkatatny et al., 2018) and sonic wave velocities (Warpinski et al., 1998) have been extensively employed for dynamic elastic property estimation in many fields, the field scale sonic shear wave measurement especially in mining operations is almost non-existence. Yet the sonic shear wave provides the unique opportunity to estimate a continuous profile of dynamic rock elastic properties along the borehole. Therefore, having a proper predictive model to estimate sonic shear wave velocities from available ultrasonic measurements is highly sought.

5.4. Data analysis and model selection

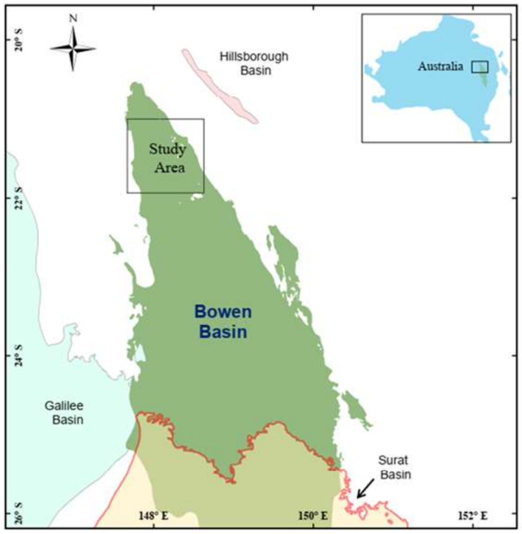
A large number of various MLalgorithms have been used in engineering geology for correlating rock physical properties with geophysical parameters. The two most popular algorithms based on an extensive literature review are the SVR and ANN (Anemangely et al., 2019; Zhang et al., 2020; Wang et al., 2020). It is noteworthy that while these two algorithms have been used frequently, no proper explanation regarding their suitability for such applications have been provided.

Due to the limited features available for analysis in rock physics studies, it has been noted that the ML algorithms capable of handling high variance are desirable to achieve efficient learning and predictions. In this study, we opted to utilize the LSSVR which is an improved version of SVR. This decision was based on the high computational requirements of SVR for optimization, which have been significantly improved in the LSSVR method (Wang and Hu, 2005). It is important to note that various ML methods have the potential for regression analysis in this context. However, for the scope of this work, which primarily focuses on the conversion of sonic to ultrasonic data rather than identifying the best ML algorithm, LSSVR and MLR were used. A brief explanation of these two algorithms is provided later.

5.4.1 Study area and its geological information

Two separate datasets containing i) the core samples extracted from three neighbouring mines located in Bowen basin, Queensland, Australia (Figure 5.2a) were used for laboratory ultrasonic compressional and shear wave velocity measurements (dataset-1) and ii) downhole logging data extracted from a nearby borehole (dataset-2). It should be noted that the laboratory ultrasonic test results were provided to us by the mining company; we did not conduct any laboratory measurements ourselves.

(a)



(b)

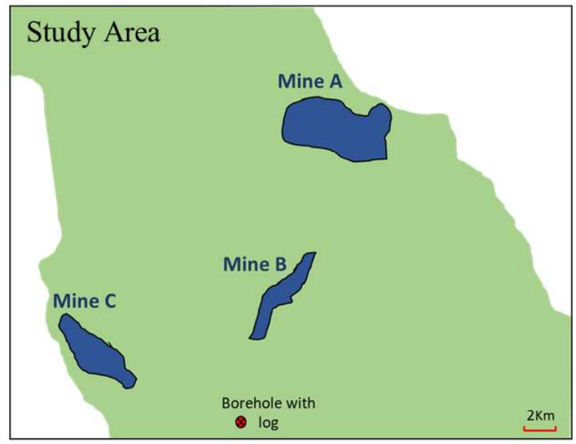


Figure 5.2a) Approximate location of the study mines in Bowen basin and b) the locations of the mines where core data were collected as well as the borehole with a complete downhole logging data (Maxwell et al., 2019).

The laboratory measurements were obtained from 178 core samples from which, 154 samples were from Mine A, 17 samples from Mine B and 7 samples from Mine C (mostly sandstone and siltstone) with depth ranging from 500 m to 700 m (Figure 5.2b). The ultrasonic test was conducted at room temperature and atmospheric pressure using an ultrasonic system with an approximate frequency of 2 MHz. This dataset (dataset-1) was used to train and validate the developed models.

While density, neutron porosity and SV_P from downhole geophysical logs were measured in the majority of the boreholes in the study mines, no SV_S was reported. Hence, a borehole with a log suit of density, neutron porosity, SV_P and SV_S (dataset-2) located near the study mines in the same basin was used to assess the predictability of the developed models. This borehole was selected due to availability of its shear wave measurement. The logging data extracted from this borehole at the depth of 600m to 900m was used in the analysis.

5.4.2 Data preparation and input parameters selection

It is common to have outliers in the collected data due to error in the measurements. Therefore, the Tukey's statistical method was utilised to remove the possible outliers from the datasets (data points that were anomalous and significantly different from most of the other observations). In the Tukey's method (Madhubabu et al., 2016; Anemangely et al., 2019;

Mehrad et al., 2020), the Interquartile Range (IQR) is defined as the distance between Q1 and Q3 (first and third quartiles) and the inner and outer boundaries are determined to separate the main observations and outliers using IQR:

$$\text{Inner boundary} = Q1 - 1.5IQR \quad (5.1)$$

$$\text{Outer boundary} = Q3 + 1.5IQR \quad (5.2)$$

As a result, any observations outside inner and outer boundaries were considered as outliers. Figure 5.3 shows the box plot for all laboratory measurements and identified outliers. The standardisation of data was performed for model development. There are several methods for standardisation including min-max normalization and power transformer scaler amongst others. Since two sources of data (field and lab measurements) were available, standardization process through mean and standard deviation (Eq. 5.3) was utilized. It is noteworthy that each dataset was standardised separately using their own mean and standard deviation:

$$z = \frac{x - \mu}{SD} \quad (5.3)$$

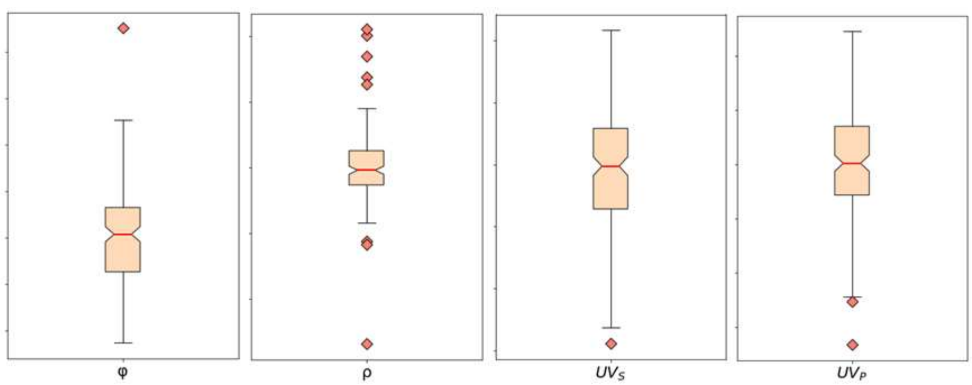


Figure 5.3. Box plot of laboratory measurements; red diamonds represent the outliers.

Furthermore, the feature selection process is required prior to training a ML algorithm (Miah et al., 2020; Onalo et al., 2018; Zhang et al., 2020; Al-Dousari et al., 2016). The aim of feature selection is to assess the sensitivity of input parameters and rank them with respect to their dependency on output parameters. To select the feature combination, the correlation matrix (Figure 5.4) consisting of the correlation coefficients amongst all possible pairs of parameters (possible inputs and output parameters of the model) was calculated and the proper features combinations were selected accordingly. As the number of available features was limited and they show high correlation with the output parameter, all the features including density (ρ), porosity (φ) and UV_P were used for training purpose.

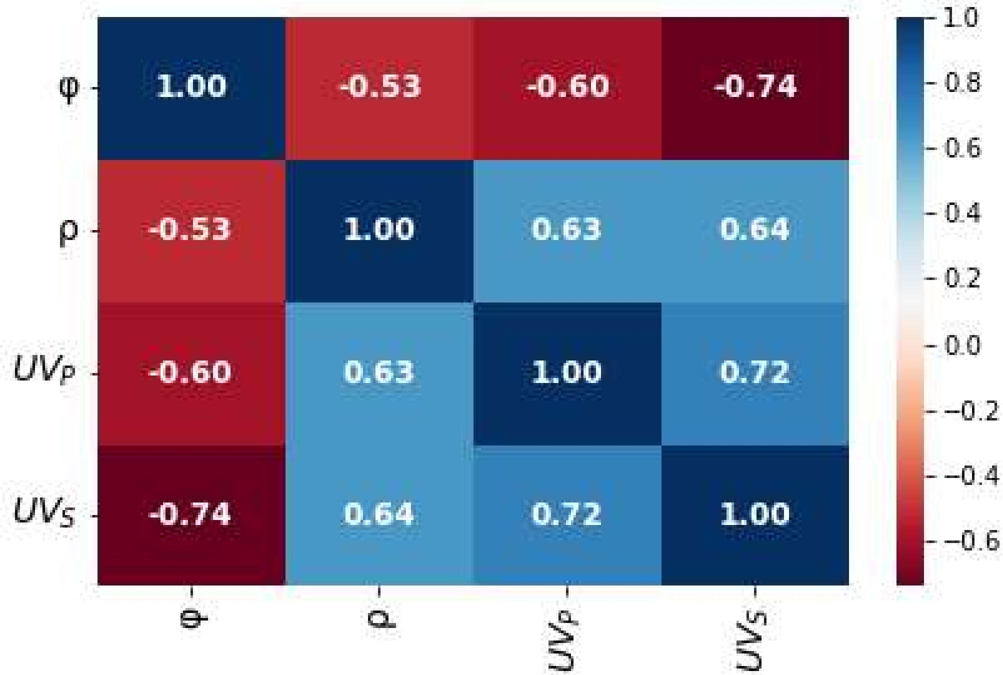


Figure 5.4. Correlations matrix amongst the various pairs of physical parameters measured within the laboratory environment.

5.4.3 Least Squares Support Vector Regression (LSSVR)

LSSVR is an improved version of SVR algorithm whose objective and restriction functions are slightly different to that of SVR (Suykens et al., 2002). LSSVR solves linear equation instead of quadratic hence reduces the computational cost and complexity of regression analysis (Wang and Hu, 2005). The objective function is defined as:

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2 \qquad \text{Subject to } y_i = w^T.F(x) + b + e_i \tag{5.4}$$

where F is a non-linear function for mapping input into a higher dimensional feature space, w is the weight factor, e_i is the error in each data point and γ is the regularisation parameter. The Lagrange multiplier is used to find the optimum values:

$$L(\omega, b, e; \alpha) = J(w, e) - \sum \alpha_k \{w^T . F(x_k) + b + e_k - y_k\} \quad (5.5)$$

where the derivatives of Lagrange multiplier are determined as:

$$\begin{aligned}
&\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^{N} \alpha_k \{w^T . F(x_k) \\
&\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^{N} \alpha_k = 0 \\
&\frac{\partial L}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_{k,} \qquad\qquad k = 1, \ldots, N \\
&\frac{\partial L}{\partial \alpha_k} = 0 \rightarrow w^T . F(x_k) + b + e_k - y_k = 0, \quad k = 1, \ldots, N
\end{aligned} \quad (5.6)$$

Solving the above equations and applying Mercer's condition results in an estimation function resulting in LSSVR to be used for nonlinear function:

$$y_i = \sum_{i=1}^{l} \alpha_i K(x . x_i) + b \quad (5.7)$$

where α_i and K are the support vectors and kernel functions, respectively.

5.4.4 Multiple Linear Regression

Multiple linear regression can identify the possible correlation between the input and output data linearly. MLR uses a linear correlation between several independent variables ($x_1, x_2, x_3, \ldots., x_n$) and a dependent variable (f). A multivariable linear equation can be expressed as:

$$f = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots. + w_n x_n \quad (5.8)$$

where $w_{0,}, w_{1,} w_{2,} \ldots and\ w_n$ are coefficients for each independent variable.

.

5.5. Predictive Models

All measurements including density, porosity, UV_P and UV_S obtained from the dataset-1 (laboratory) were processed using Tukey's method to remove possible outliers. Each property was then standardised using its own mean and standard deviation to prepare the data for modelling. The φ, ρ, $UV_P * \rho$ were considered for estimating UV_S. The cross plots presented in Figure 5.5 reveal correlations between the input and output parameters. Notably, the density data exhibits more dispersion compared to other combinations, implying that UV_S is strongly correlated with porosity and UV_P than with density. This relationship is further highlighted in the correlation matrix (Figure 5.4), where the correlation coefficient values for porosity, density and UV_P with UV_S are 0.74, 0.64, and 0.72, respectively. As a result, it can be concluded that porosity and UV_P have greater impacts on UV_S than density.

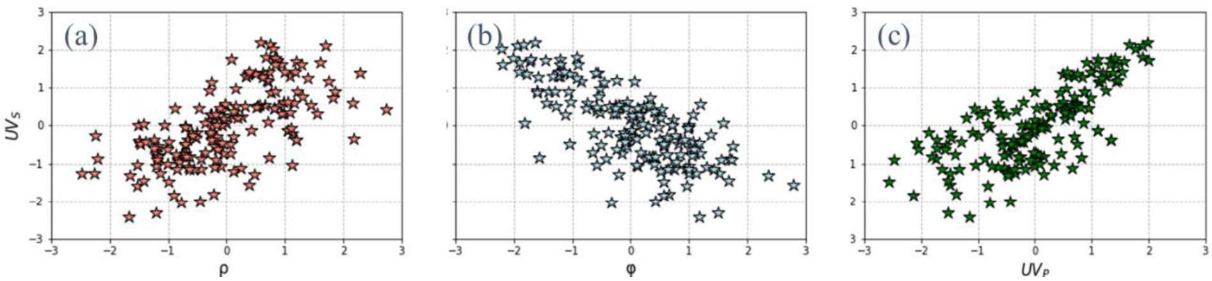


Figure 5.5. Correlation plots of three selected parameters from the laboratory measurements. (a)-UV_S (b) -UV_S and (c)UV_P - UV_S

5.5.1 Training and validation using laboratory data

The dataset-1 was mainly used to develop the models, thus 20% of dataset-1 was randomly selected for validation and the remaining 80% for the training, as a common practice in ML modelling (Mehrgini et al., 2019).

Selection of an appropriate kernel function is crucial during the LSSVR modelling (Asoodeh and Bagheripour, 2012; Anemangely et al., 2019). Amongst several available kernel functions, radial basis function (RBF) was chosen by trial-and-error. Two variables including σ and γ were then optimised using a random search amongst [0.01, 10] and [0.01, 50], respectively to generate the most suitable predictive model (Suykens and Gestel, 2003). It is noted that σ^2 and γ are the regularization and RBF kernel parameters (squared bandwidth), respectively which control the generalization, prediction and learning ability of the model (Anemangely et al., 2019). MLR was also employed to determine the best possible linear correlation fitted to the training data.

The developed models were then verified using the selected validation data. The coefficient of determination (R^2) and root-mean square error (RMSE) were utilised as the statistical measures to evaluate the performance of the developed models (Table 5.1). Figures 5.6a and b demonstrate the predictability of resulting models based on LSSVR and MLR, respectively. It is clear that LSSVR resulted in a higher R^2 and lower RMSE compared to the model obtained from MLR highlighting its superiority. It is noteworthy that applying a laboratory-based model for prediction of logging measurements is the main purpose of this study rather than selecting the best algorithm for the model development. Therefore, while LSSVR led to a better prediction, for comparison purposes both algorithms were used for the field downhole logging data estimation in the next section.

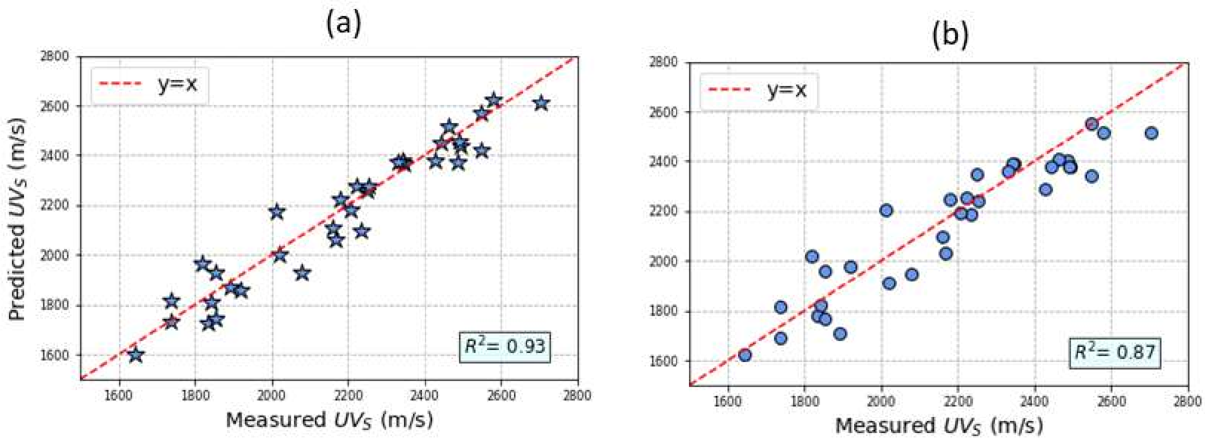


Figure 5.6. Comparing the predicted and measured UV_S data based on (a) LSSVR and (b) MLR

Table 5.1. Comparing the accuracy of UVS predictive models obtained from LSSVR and MLR

model	R^2	RMSE
LSSVR	0.93	0.26
MLR	0.87	0.34

5.5.2 Testing the model using the field data

In the previous section, two predictive models were developed based on LSSVR and MLR to estimate UV_S using the laboratory data from three neighbouring mines. Now, the predictability of these proposed models to estimate sonic logs data is assessed. To do so, the logging measurements from a borehole with the log suits of density, neutron porosity, SV_P and SV_S were used. The same standardisation procedure as that conducted on the laboratory data was performed on the logging data resulting in density, porosity and SV_P as inputs and SV_S as output. Subsequently, the developed models were tested using the logging data (dataset-2) extracted from a sandstone interval of the nominated borehole (see Figure 5.7). As pointed earlier, the lithology of the core samples which were used for the initial model development were mainly sandstone and siltstone. Therefore, for consistency, the coal intervals were excluded from the logging data of the analysis. Since all the input parameters were standardised, the developed models estimate the standardised SV_S through the logging interval, thus predicting a trend of SVs than actual values. Figure 5.8a and b show the performance of both models in predicting the trend of SV_S.

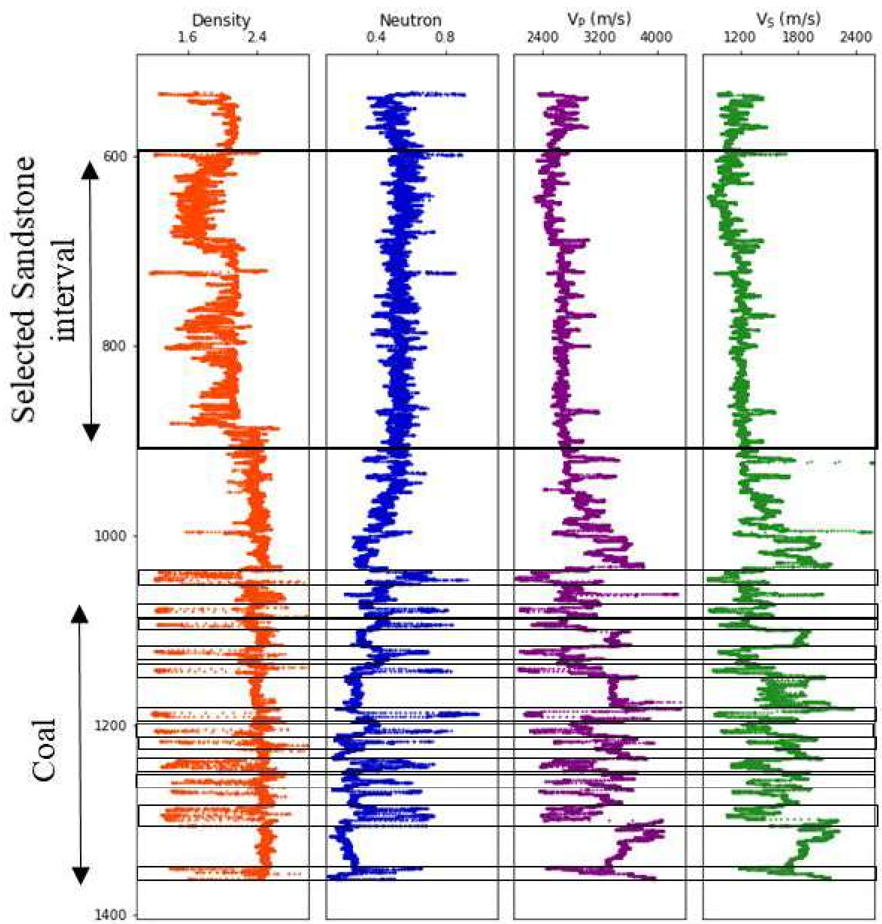


Figure 5.7. Density, Neutron porosity, SV_P and SV_S of the test borehole

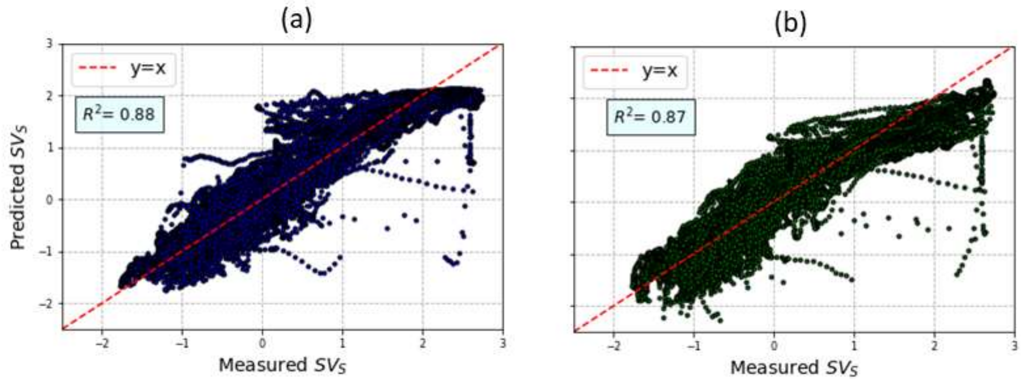


Figure 5.8. Comparing the standardised predicted and measured SV_S data based on (a) LSSVR and (b) MLR

5.5.3 Selecting mean and standard deviation for field data

While both LSSVR and MLR could estimate the trend of SV_S (the standardised SV_S), the predicted values need to be un-standardised for further assessment of the developed models. The un-standardising procedure, however, required a valid mean and standard deviation. In this study, it was postulated that the standard deviation of logging measurements for shear wave velocity falls within the standard deviation of laboratory measurements ofUV_S. The postulation was based on the inclusion of various core samples in the training data which captures the required variability. In addition, it is known that the variation in wave velocity during the ultrasonic measurements is higher than sonic measurements (Nolen-Hoeksema et al., 1995). To gain confidence on the postulation used, we additionally compared the standard deviation of the ultrasonic measurements to the available sonic shear wave measurements (used for comparison) and observed that the standard deviation of sonic shear wave measurements (SD=294.9) sits within the ultrasonic laboratory measurements (SD =303.9). The mean value can be either assumed based on a typical range of SV_S in the study area or extracted from other boreholes in the neighbourhood regions if available. In case of absolute unavailability of SV_S, theoretical models based on physical principles such as squirt flow can be used as an alternative method to estimate the mean shear wave velocity at the desired frequency.

5.5.4 Squirt flow model for mean estimation

Squirt flow is amongst the well-known theories for modelling velocity dispersion in rocks. This theory can be used to estimate shear wave velocity at desired frequencies using rock physical and mechanical properties.

$$V_P = \sqrt{Real(\, K_r \; + \tfrac{3}{4} G_m)/\rho} \qquad (5.9)$$

$$V_S = \sqrt{Real(\, G_m)/\rho} \qquad (5.10)$$

where K_r and G_m are the saturated rock bulk modulus and shear modulus of modified frame respectively. K_r, and G_m can be estimated using bulk modulus of dry rock (K) and mineral phase (K_S), shear modulus of dry rock (G), rock density (ρ) and porosity (φ) at a specific frequency. The detailed steps for implementing squirt flow model can be found in Dvorkin et al., (1995). In order to estimate SV_S at logging frequencies using this theory, rock physical and mechanical properties were extracted from laboratory measurements on a core sample in dataset-1. The logging frequency was 12 KHz. As the squirt flow model estimates both shear and compressional wave velocity, it is conventional to use one of the estimates for determining the tuning parameter (Z). Since SV_P was available in the study borehole, it was used as matching parameter i.e. Z was iterated until measured and estimated SV_P matched. Each parameter was varied by ±10% around its nominal value to account for possible measurement errors and other sources of uncertainty. By simulating 100 possible combinations of parameter values, a distribution of estimated SV_S values were generated that reflects the range of possible outcomes due to uncertainties in the parameters. From this distribution, the mean and the error of the estimated SV_S values were computed. Table 5.2 shows the parameters which were used for estimating SV_S in the squirt flow model and was resulted in estimated SV_S of 1320 m/s with potential error of 43 m/s. This practice should provide an accurate basis for the analysis considering that the investigated interval from the logging suit was mainly sandstone.

Table 5.2. The mean value for rock physical and mechanical properties of the selected core sample and the estimated mean UV_S

Property	Value
K (GPa)	9.44 ± 10%
K_S (GPa)	47.2 ± 10%
G (GPa)	6.07 ± 10%
(Kg/m^3)	2.1 ± 10%
	0.22 ± 10%
UV_S (m/s)	1320 ± 43

5.5.5 Model evaluation and field application

Using the mean value from squirt flow model and standard deviation from the laboratory measurements, the SVs estimation was un-standardised and compared with logging data. The resulting R^2 based on LSSVR and MLR algorithms are 0.88 and 0.87, respectively which are similar to what was obtained in Section 5.2.1 (see Figure 5.9). This confirms the accuracy and robustness of the proposed workflow for the prediction of sonic shear wave from the laboratory ultrasonic shear wave data. As a result, it is true to state that any missing sonic log data can be indirectly estimated from the laboratory physical and ultrasonic measurements with good accuracy using the proposed workflow based on LSSVR and MLR. The estimated SV_S through both models were also shown in Figure 5.10 to reflect the trend of prediction at different depths.

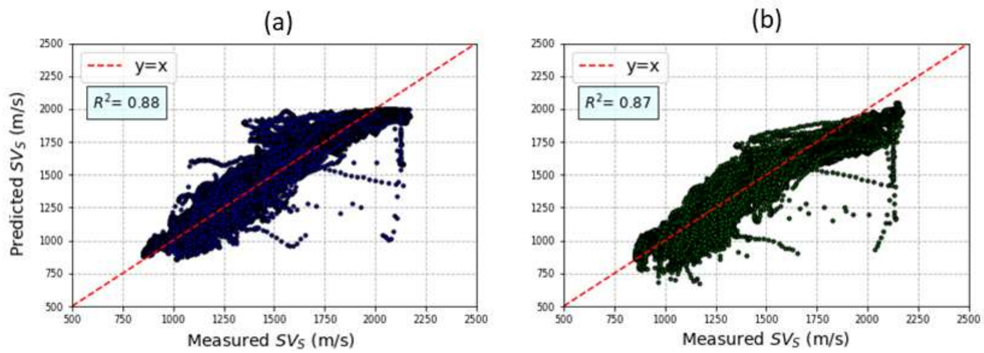


Figure 5.9. Comparing the predicted and measured SV_S data based on (a) LSSVR and (b) MLR

As can be seen in Figure 5.8 (a, b) and 5.9 (a, b), the predicted versus measured SV_S mostly follow the linear trend confirming the suitable performance of the developed models. However, minor deviations from the linear trend can be observed at higher velocities (greater than 2000 m/s). This trend is also reflected in Figure 5.10, where the deviation between the predicted and measured values increases with depth. Assuming no bias/error in the measurements (and outliers removed), several processes can lead to such deviation. While the frequency dependency of wave velocity has been already taken into consideration, the difference between the field and laboratory measurements of acoustic properties still needs attention. Amongst different in-situ conditions, the stress, pore pressure and temperature can play considerable roles.

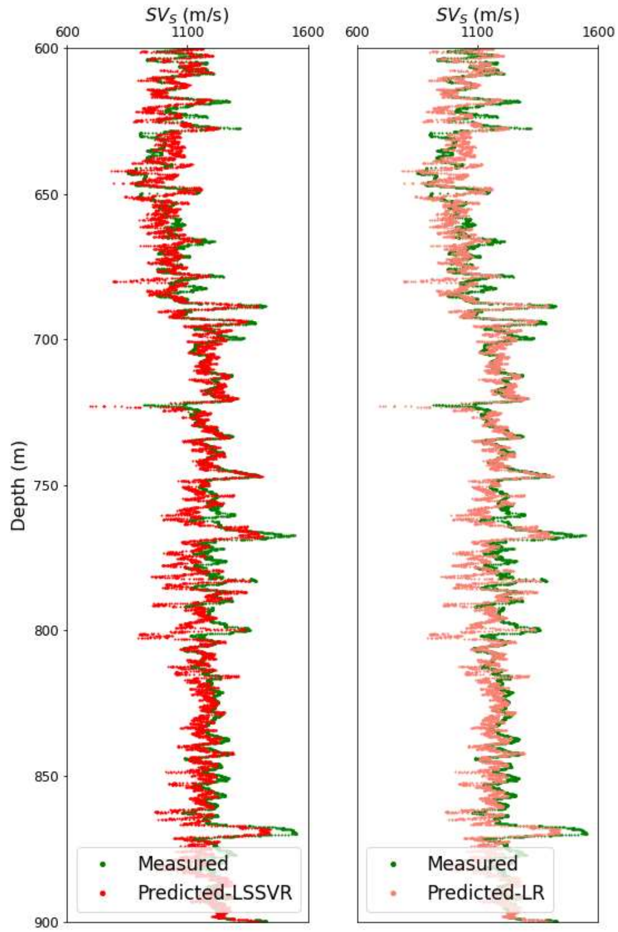


Figure 5.10. Predicted and measured SV_S data at different depths based on (a) LSSVR and (b) MLR

The field logging tools perform the measurements at in-situ stress and temperature conditions, while the laboratory ultrasonic measurements are often conducted under atmospheric conditions. For example, the effect of stress on the microcracking process of intact rock and drilling induced damage were reported as processes causing differences between mechanical properties measured at the field and laboratory conditions in sedimentary rocks (Ide, 1936, Jizba,1992; Walsh, 1996). It is noted that the impact of stress and temperature is amplified at deeper intervals observed from our results (Fig. 5.9) when deviations occur at higher velocities (representing deeper rocks with lower porosity). At shallower depths, both the in-situ stress and temperature of the logged interval are lower and closer to the laboratory conditions leading to less/no deviation. It is thus important to consider these effects when interpreting logging data from deeper intervals, as the measured properties may not accurately reflect the conditions within the laboratory.

5.6. Conclusion

Compressional and shear wave velocities can be measured in the laboratory or field environments. Different conditions and frequencies associated with sonic and ultrasonic measurements can lead to discrepancies in measured sonic (SV_P, SV_S) and ultrasonic velocities (UV_P, UV_S). In this study, a set of laboratory ultrasonic and sonic log data were used to develop an integrated ML and physics model. 178 core samples with measured density, porosity, ultrasonic compressional and shear wave velocities (UV_P and UV_S) were collected from three neighbouring mines in Bowen basin, Australia. Hence, two predictive models based on laboratory data were developed to estimate UV_S based on LSSVR and MLR algorithms. The developed predictive models were tested to assess their SV_S predictability where squirt flow theoretical model was used to convert the mean ultrasonic velocity to sonic velocity. Subsequently, it was demonstrated that the developed models based on LSSVR and MLR can estimate field measured SV_S with R^2 of 0.93 and 0.87, respectively indicating their suitable predictability. The results from this study show that the proposed workflow can be used to estimate any missing sonic log data (e.g. SV_P, SV_S) from the laboratory ultrasonic measurements with high accuracy.

5.7. References

Abbas A K, Flori R E, Alsaba M (2018) Estimating rock mechanical properties of the Zubair shale formation using a sonic wireline log and core analysis. J Nat Gas Sci Eng 53:359-369

Al-Anazi, A., & Gates, I. D. (2010). A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. J Eng Geol ,*114*(3-4), 267-277.

Al-Dousari M, Garrouch A A, Al-Omair O (2016) Investigating the dependence of shear wave velocity on petrophysical parameters. J Petrol Sci Eng 146:286-296

Anemangely M, Ramezanzadeh A, Amiri H, Hoseinpour S-A (2019) Machine learning technique for the prediction of shear wave velocity using petrophysical logs. J Petrol Sci Eng 174:306-327

Asoodeh M, Bagheripour P (2012) Prediction of compressional, shear, and stoneley wave velocities from conventional well log data using a committee machine with intelligent systems. Rock Mech Rock Eng 45(1):45-63

Bagheripour P, Gholami A, Asoodeh M,Vaezzadeh-Asadi M (2015) Support vector regression based determination of shear wave velocity. J Petrol Sci Eng 125:95-99

Bailey T, Dutton D (2012). An empirical vp/vs shale trend for the kimmeridge clay of the central North Sea. 74th EAGE Conference and Exhibition incorporating EUROPEC 2012, European Association of Geoscientists & Engineers.

Benavente, D., Galiana-Merino, J. J., Pla, C., Martinez-Martinez, J., & Crespo-Jimenez, D. (2020). Automatic detection and characterisation of the first P-and S-wave pulse in rocks using ultrasonic transmission method. J Eng Geol, *266*, 105474.

Behnia D, Ahangari K, Moeinossadat S R (2017) Modeling of shear wave velocity in limestone by soft computing methods. Int J Min Sci Technol 27(3):423-430

Bennett R H, Bennett R H, Bryant W R, Chiou W, Faas R, Hulbert M H, Kasprowicz J, Bouma A, Li H, Lomenick T,1991. Microstructure of fine-grained sediments: From mud to shale, Springer Science & Business Media.

Castagna J P, Backus M M,1993. Offset-dependent reflectivity—Theory and practice of AVO analysis, Society of Exploration Geophysicists.

Chang C, Zoback M D, Khaksar A (2006) Empirical relations between rock strength and physical properties in sedimentary rocks. J Petrol Sci Eng 51(3-4):223-237

Chawre B (2018) Correlations between ultrasonic pulse wave velocities and rock properties of quartz-mica schist. J Rock Mech Geotech Eng 10(3):594-602

Cheng C, Johnston D H (1981) Dynamic and static moduli. Geophys Res Lett 8(1):39-42

Cranganu C, Breaban M (2013) Using support vector regression to estimate sonic log distributions: a case study from the Anadarko Basin, Oklahoma. J Petrol Sci Eng 103:1-13

Domenico S N (1984) Rock lithology and porosity determination from shear and compressional wave velocity. Geophysics 49(8):1188-1195

Dvorkin J, Mavko G, and Nur A (1995) Squirt flow in fully saturated rocks. Geophysics 60(1): 97-107.

Eastwood R L, Castagna J P (1983). Basis for interpretation of Vp/Vs ratios in complex lithologies. SPWLA 24th Annual Logging Symposium, OnePetro.

Eberhart-Phillips D, Han D-H, Zoback M (1989) Empirical relationships among seismic velocity, effective pressure, porosity, and clay content in sandstone. Geophysics 54(1):82-89

Ersoy, H, Karahan, M, Babacan, A E, & Sünnetci, M O (2019) A new approach to the effect of sample dimensions and measurement techniques on ultrasonic wave velocity. J Eng Geol.*251*, 63-70.Elkatatny S, Mahmoud M, Mohamed I, Abdulraheem A (2018) Development of a new correlation to determine the static Young's modulus. J Pet Explor Prod Technol 8(1):17-30

Fjær E (2009) Static and dynamic moduli of a weak sandstone. Geophysics 74(2): WA103-WA112

Hamada G (2004) Reservoir fluids identification using Vp/Vs ratio? Oil & Gas Science and Technology 59(6):649-654

Han D-h, Nur A, Morgan D (1986) Effects of porosity and clay content on wave velocities in sandstones. Geophysics 51(11):2093-2107

Hatherly P, 2013. Overview on the application of geophysics in coal mining. Int J Coal Geol, 114, pp.74-84.

Hatherly P, Medhurst T, Zhou B (2016) Geotechnical evaluation of coal deposits based on the Geophysical Strata Rating. Int J Coal Geol 163:72-86

He, W, Chen, Z, Shi, H, Liu, C, & Li, S (2021) Prediction of acoustic wave velocities by incorporating effects of water saturation and effective pressure. J Eng Geol, *280*, 105890.

Hossain Z, Mukerji T, Fabricius I.L (2012) V_p-V_s relationship and amplitude variation with offset modelling of glauconitic greensand. Geophys Prospect 60(1):117-137

Howarth D (1984) Apparatus to determine static and dynamic elastic moduli. Rock Mech Rock Eng 17(4):255-264

Ide J M (1936) Comparison of statically and dynamically determined Young's modulus of rocks. Proceedings of the National Academy of Sciences of the United States of America 22(2):81

Jizba D L (1992) Mechanical and acoustical properties of sandstones and shales. Doctor of Philosophy thesis, Stanford University.

Johnston J.E. and Christensen N.I. (1993) Compressional to shear velocity ratios in sedimentary rocks. Int. J. Rock Mech. Sci & Geomech. Abstr. 30,751-754.

Jørstad A, Mukerji T, Mavko G (1999) Model-based shear-wave velocity estimation versus empirical regressions. Geophys Prospect 47(5):785-797

Karacan C.Ö (2009) Elastic and shear moduli of coal measure rocks derived from basic well logs using fractal statistics and radial basis functions. Int J Rock Mech Min 46(8):1281-1295

Kennedy M, 2015. Practical petrophysics, Elsevier.

Khandelwal M, & Singh, T. N. (2009). Correlating static properties of coal measures rocks with P-wave velocity. Int J Coal Geol , 79(1-2), 55-60.

Khandelwal M (2013) Correlating P-wave velocity with the physico-mechanical properties of different rocks. Pure and Applied Geophysics 170(4):507-514

Klarner S, Klarner O (2012) Identification of paleo-volcanic rocks on seismic data. Updates in volcanology—a comprehensive approach to volcanological problems. Intech, Rijeka:181-206

Li, G, Li, G, Wang, Y, Qi, S, & Yang, J (2020) A rock physics model for estimating elastic properties of upper Ordovician-lower Silurian mudrocks in the Sichuan Basin, China, J Eng Geol. *266*, 105460.

Lv A, Aghighi M.A, Masoumi H, Roshan H (2022) The effects of sorbing and non-sorbing gases on ultrasonic wave propagation in fractured coal. Int J Coal Geol 249:103906.

Madhubabu N, Singh P, Kainthola A, Mahanta B, Tripathy A, Singh T (2016) Prediction of compressive strength and elastic modulus of carbonate rocks. Measurement 88:202-213

Mavko G, Bandyopadhyay K (2009) Approximate fluid substitution for vertical velocities in weakly anisotropic VTI rocks. Geophysics 74(1): D1-D6

Maxwell K, Rajabi M, Esterle J (2019) Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. Int J Coal Geol 214:103284

Mehrad M, Bajolvand M, Ramezanzadeh A, Neycharan J.G (2020) Developing a new rigorous drilling rate prediction model using a machine learning technique. J Petrol Sci Eng 192:107338

Mehrgini B, Izadi H, Memarian H (2019) Shear wave velocity prediction using Elman artificial neural network. Carbonates and Evaporites 34(4):1281-1291

Miah M.I, Ahmed S, Zendehboudi S, Butt S (2020) Machine Learning Approach to Model Rock Strength: Prediction and Variable Selection with Aid of Log Data. Rock Mech Rock Eng 53(10):4691-4715

Mohammadpour, M, Roshan, H, Arashpour, M and Masoumi, H (2023) Effect of spatial variability of downhole geophysical logs on machine learning exercises. Int J Coal Geol , 277, p.104333.

Najibi A.R, Ghafoori M, Lashkaripour G.R, Asef M.R (2015) Empirical relations between strength and static and dynamic elastic properties of Asmari and Sarvak limestones, two main oil reservoirs in Iran. J Petrol Sci Eng 126:78-82

Olayiwola T, Sanuade O.A (2020) A data-driven approach to predict compressional and shear wave velocities in reservoir rocks. Petroleum

Olsen C, Christensen H.F, Fabricius I.L (2008) Static and dynamic Young's moduli of chalk from the North Sea. Geophysics 73(2): E41-E50

Onalo D, Adedigba S, Khan F, James L.A, Butt S (2018) Data driven model for sonic well log prediction. J Petrol Sci Eng 170:1022-1037

Pan, J., Meng, Z., Hou, Q., Ju, Y. and Cao, Y., 2013. Coal strength and Young's modulus related to coal rank, compressional velocity and maceral composition. *Journal of Structural Geology*, *54*, pp.129-135.

Pappalardo, G., & Mineo, S. (2022). Static elastic modulus of rocks predicted through regression models and Artificial Neural Network. J Eng Geol, *308*, 106829.

Parvizi S, Kharrat R, Asef M.R, Jahangiry B, Hashemi A (2015) Prediction of the shear wave velocity from compressional wave velocity for Gachsaran Formation. Acta Geophysica 63(5):1231-1243

Pickett G.R (1963) Acoustic character logs and their applications in formation evaluation. J Pet Technol 15(06):659-667

Potter C.C, Miller S, Margrave G.F (1996) Formation elastic parameters and synthetic PP and PS seismograms for the Blackfoot field. CREWES Res. Rep. 8:37-31-37–18

Rajabi M, Bohloli B, Ahangar E.G (2010) Intelligent approaches for prediction of compressional, shear and Stoneley wave velocities from conventional well log data: A case study from the Sarvak carbonate reservoir in the Abadan Plain (Southwestern Iran). Comput Geoci 36(5):647-664

Richardson M.D, Muzi E, Troiano L, Miaschi B (1991). Sediment shear waves: A comparison of in situ and laboratory measurements. Microstructure of Fine-Grained Sediments, Springer: 403-415.

Riedel M, Goldberg D, Guerin G (2014). Compressional and shear-wave velocities from gas hydrate bearing sediments: Examples from the India and Cascadia margins as well as Arctic permafrost regions. Mar Pet Geol 58:292-320.

Sarkar K, Vishal V, Singh T (2012) An empirical correlation of index geomechanical parameters with the compressional wave velocity. Geotech Geol Eng 30(2):469-479

Serra O (1983) Fundamentals of well-log interpretation.

Suykens J.A, De Brabanter J, Lukas L, Vandewalle J (2002) Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48(1-4):85-105

Suykens J.A, Gestel V. T, Least Squares Support Vector Machines (2003), World Scientific.

Torabi, Anita, B. Alaei, and T. S. S. Ellingsen (2018). "Faults and fractures in basement rocks, their architecture, petrophysical and mechanical properties." *Journal of Structural Geology* 117: 256-263.

Takahashi T, Takeuchi T, Sassa K (2006) ISRM suggested methods for borehole geophysics in rock engineering. Int J Rock Mech Min 43(3):337-368

Walsh J (1966) Seismic wave attenuation in rock due to friction. J Geophys Res 71(10):2591-2599Wang H, Hu D (2005) Comparison of SVM and LS-SVM for regression. 2005 International conference on neural networks and brain, IEEE.

Wang H, Hu D (2005) Comparison of SVM and LS-SVM for regression. 2005 International conference on neural networks and brain, IEEE.

Wang J, Cao J, Yuan S (2020) Shear wave velocity prediction based on adaptive particle swarm optimization optimized recurrent neural network. J Petrol Sci Eng 194:107466

Wang P, Peng S (2019) On a new method of estimating shear wave velocity from conventional well logs. J Petrol Sci Eng 180:105-123

Wang Z, Nur A (1988) Effect of temperature on wave velocities in sands and sandstones with heavy hydrocarbons. SPE Reservoir Engineering 3(01):158-164

Warpinski N, Peterson R, Branagan P, Engler B, Wolhart S (1998). In situ stress and moduli: Comparison of values derived from multiple techniques. SPE annual technical conference and exhibition, OnePetro.

Wennberg, O. P., F. De Oliveira Ramalho, M. Virgolino Mafia, F. Lapponi, A. S. Chandler, LE Gomis Cartesio, and David W. Hunt (2023). "The characteristics of natural open fractures in acoustic borehole image logs from the pre-salt Barra Velha formation, Santos Basin, Brazil." *Journal of Structural Geology* 167:104794.

Zhang L, Zhang Z, Zhang R, Gao M, Xie J (2021) The Ultrasonic P-Wave Velocity-Stress Relationship and Energy Evolution of Sandstone under Uniaxial Loading-Unloading Conditions. Advances in Materials Science and Engineering 2021:1-11

Zhang Y, Zhong H-R, Wu Z-Y, Zhou H, Ma Q-Y (2020) Improvement of petrophysical workflow for shear wave velocity prediction based on machine learning methods for complex carbonate reservoirs. J Petrol Sci Eng 192:107234

Chapter 6

6. Estimation of rock elastic properties using the predictive models: example of application

6.1. Summary

Chapter 6 serves as a complementary section, offering a practical demonstration to the workflows and predictive models outlined in the preceding chapters. Here, the shear and compressional wave velocities were estimated by employing the developed predictive models across five randomly selected boreholes. Subsequently, these estimated velocities were utilized to construct continuous profiles of elastic modulus and Poisson's ratio. To validate the accuracy of these estimations, the results were meticulously compared with the laboratory measurements obtained from core samples.

This chapter not only illustrates a significant application of the predictive models but also serves as a validation mechanism, affirming the precision and reliability of the developed predictive models. Comparison with the laboratory measurements provides a tangible benchmark, showcasing the practical efficacy of the developed workflows and predictive models in estimating sonic wave velocities and addressing data unavailability challenges.

6.2. Introduction

Characterizing the mechanical properties of rock is a crucial aspect in the design of structures on or within rock masses. The accurate estimation of Young's modulus (E) and Poisson's ratio (ν) holds paramount importance for ensuring the safe operation of rock engineering projects. Traditionally, the uniaxial compressive test, involving the measurement of axial and radial deformations, has been the conventional method for estimating these parameters, following the guidelines of the International Society for Rock Mechanics (ISRM, 2007; Masoumi et al., 2015; Lv et al., 2021; Roshan et al., 2018). However, challenges associated with laboratory experiments, such as core sample recovery difficulties, testing costs, and time constraints, have prompted the exploration of indirect (non-destructive) methods for property estimation (Raaen et al., 1996).

One prevalent indirect method for estimating Young's modulus and Poisson's ratio involves geophysical techniques, particularly downhole sonic measurements. This prediction often relies on empirical relationships linking E solely to compressional wave velocity or utilizes physics-based solutions of elastic wave propagation to calculate dynamic E and ν (dynamic here refer to estimate from field measurements) based on compressional and shear wave velocities. This dynamic data can then be converted to static measurements (Takahashi et al., 2006). Such relationships enable the construction of continuous profiles of E and ν along a borehole using downhole geophysical data. This approach contrasts with the limitations of constructing profiles solely from laboratory experiments, given the challenges of accessing recovered core samples across the entire borehole and the potential mismatch with actual in-situ conditions.

Previous chapters extensively discussed the development of predictive models for estimating V_PV_S through various methods. In this chapter, the application of these developed models for estimating profiles of Young's modulus and Poisson's ratio are briefly explored. To validate the accuracy of V_P and V_S estimation, the estimated profiles are compared with the laboratory measurements obtained from core samples. This comparative analysis serves to enhance our understanding of the efficacy and reliability of the predictive models in capturing the complex mechanical properties of rock.

6.3. Young's modulus and Poisson's ratio estimation

Five test boreholes which were selected originally in section 3.2 (See Figure 3.1) as test data set were considered to be used for estimation of Young's modulus and Poisson's ratio using Equation 5.1.

$$\nu = \frac{\left(\frac{V_P}{V_S}\right)^2 - 2}{2\left(\frac{V_P}{V_S}\right)^2 - 2} \tag{6.1}$$

$$E = \frac{\rho V_S^{\,2}\left(3V_P^{\,2} - 4V_S^{\,2}\right)}{V_P^{\,2} - V_S^{\,2}} \tag{6.2}$$

Here, ρ represents density. Each of the five boreholes provided measurements for density, gamma, depth, and V_P. To estimate the profiles of V_P and V_S in these boreholes, the developed models from Chapter 3 and Chapter 5 were utilised. The resulting values were then integrated into the equations above for the estimation of dynamic elastic properties. Figure *6.1* to Figure *6.5* visually present the predicted sonic wave velocities and dynamic elastic properties derived from this comprehensive approach for the five test boreholes. The location of a sandstone layer within each test boreholes is specified by red zone in all figures. This holistic method not only showcases the application of the developed models but also demonstrates the potential for accurate estimation of dynamic elastic properties in real-world geological settings.

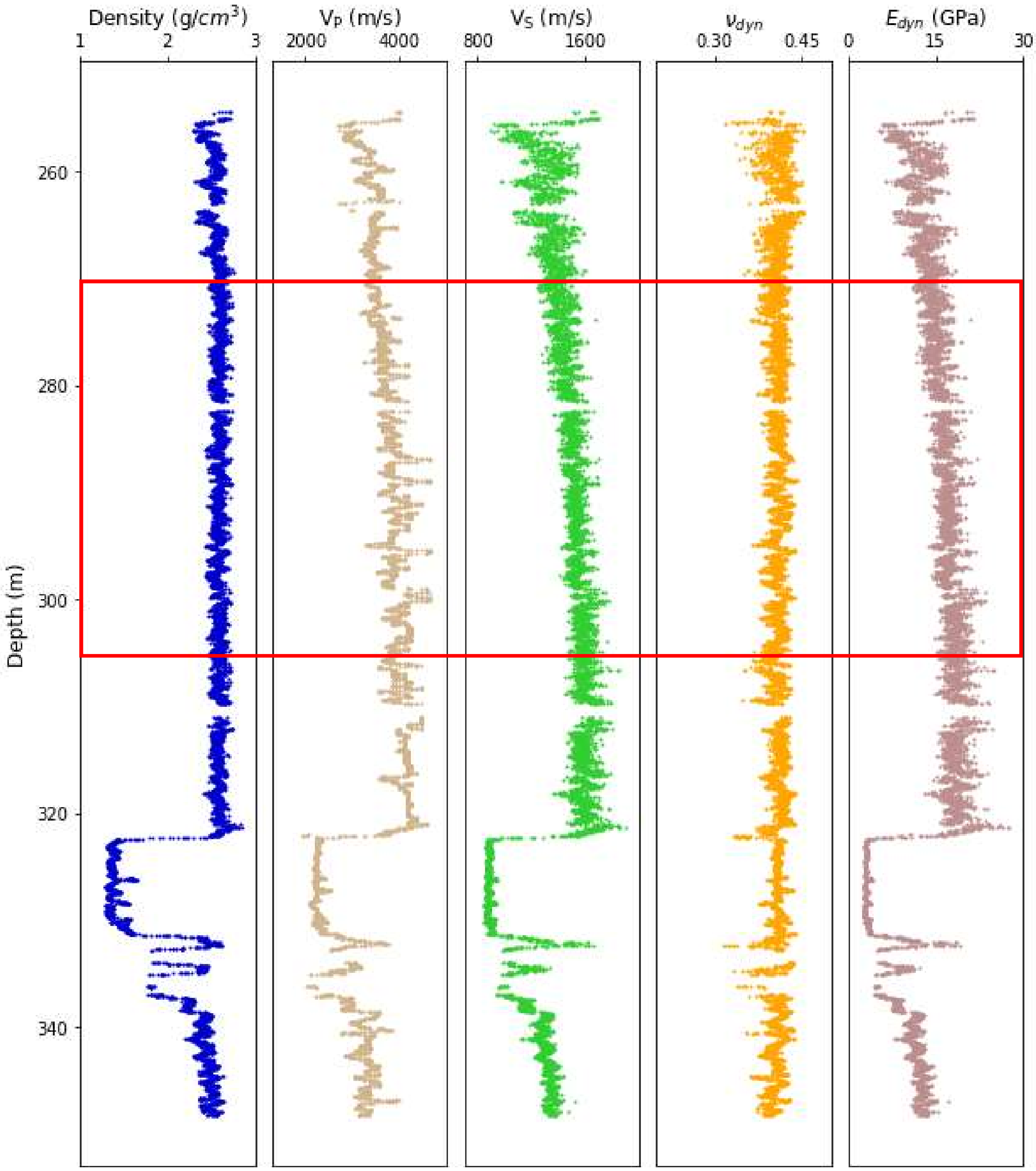


Figure 6.1. Density log and predicted V_P, V_S, E_{dyn} and ν_{dyn} for test borehole number 1

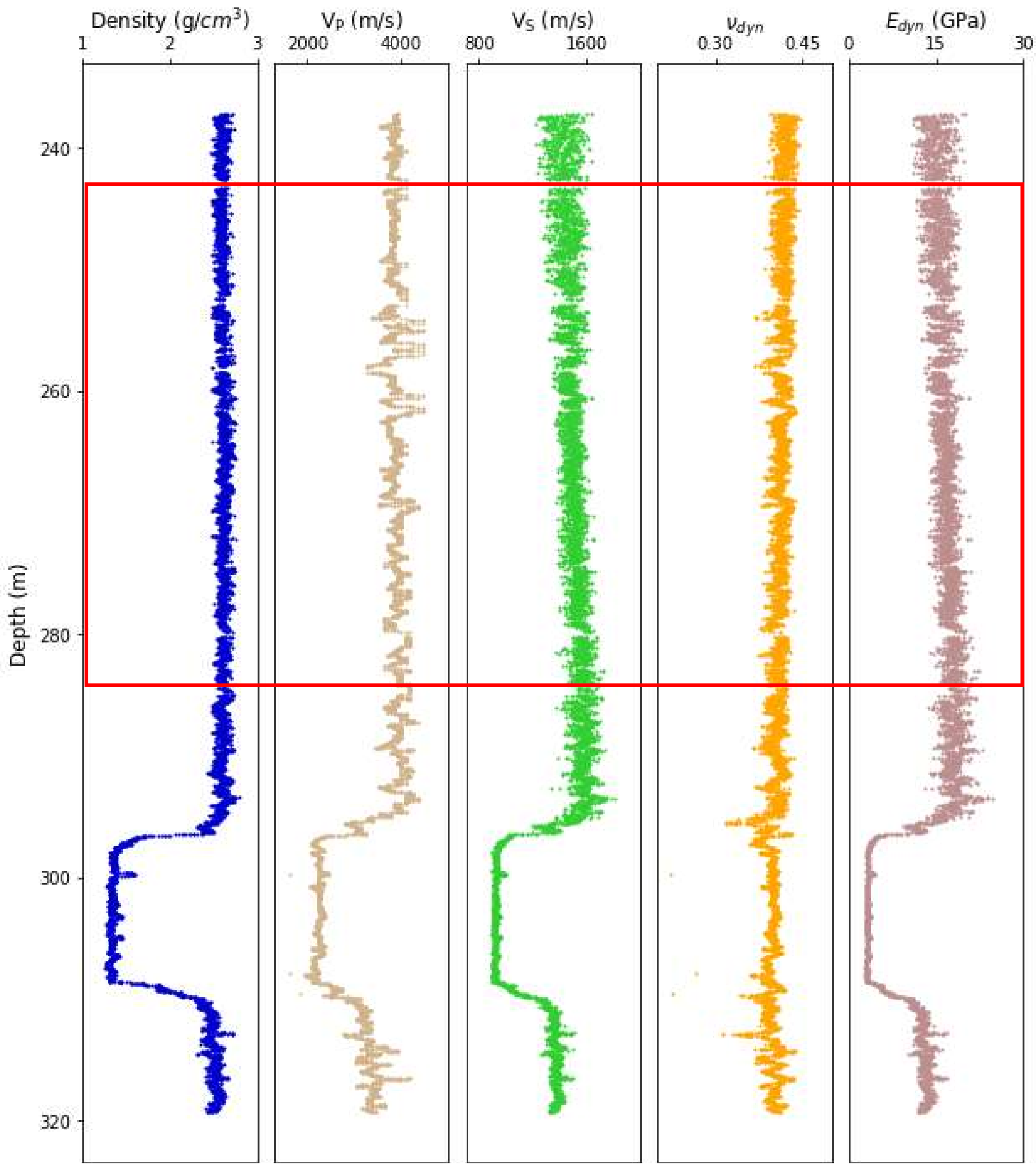


Figure 6.2 Density log and predicted V_P, V_S, E_{dyn} and ν_{dyn} for test borehole number 2

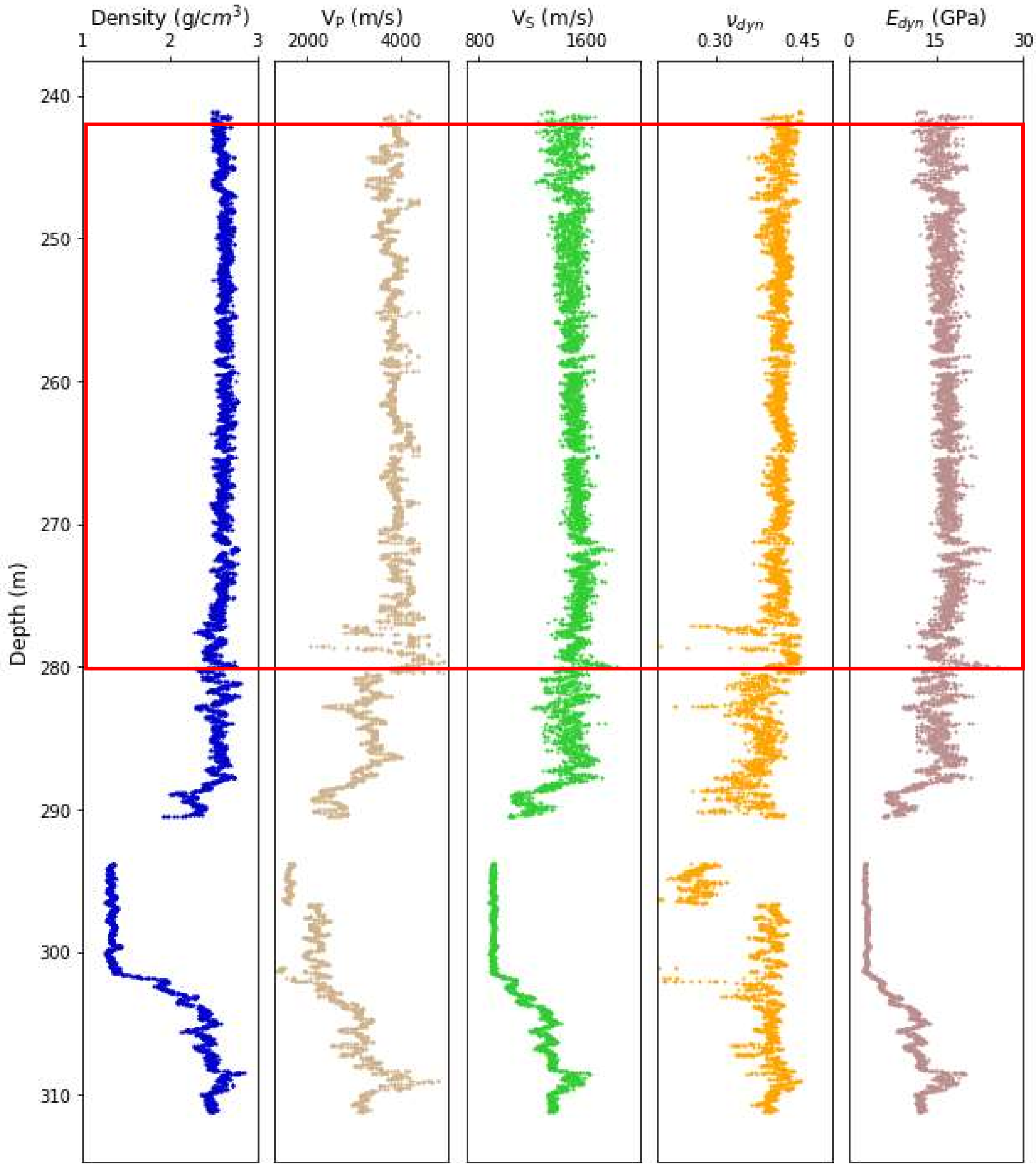


Figure 6.3 Density log and predicted V_P, V_S, E_{dyn} and ν_{dyn} for test borehole number 2

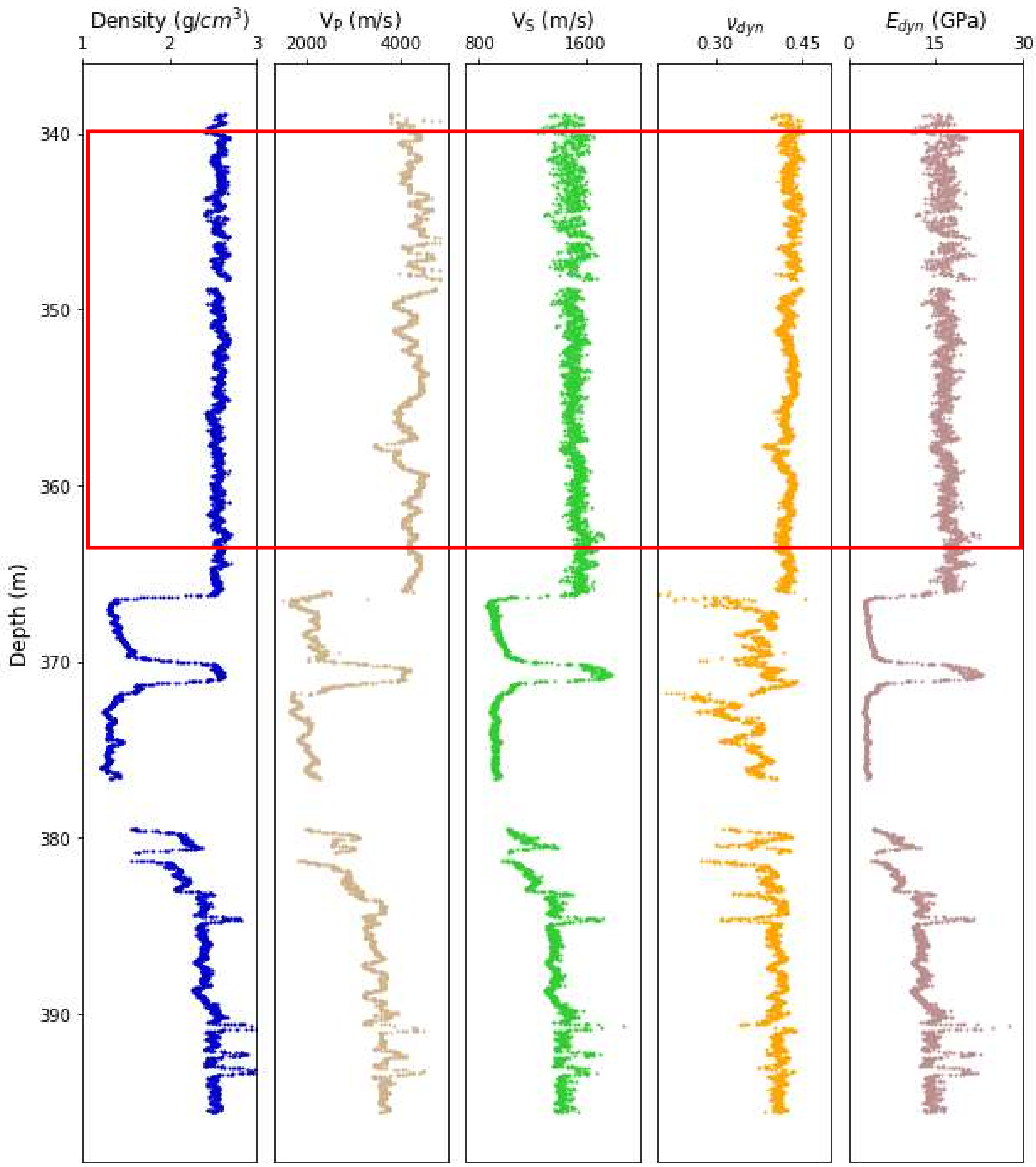


Figure 6.4 Density log and predicted V_P, V_S, E_{dyn} and ν_{dyn} for test borehole number 3

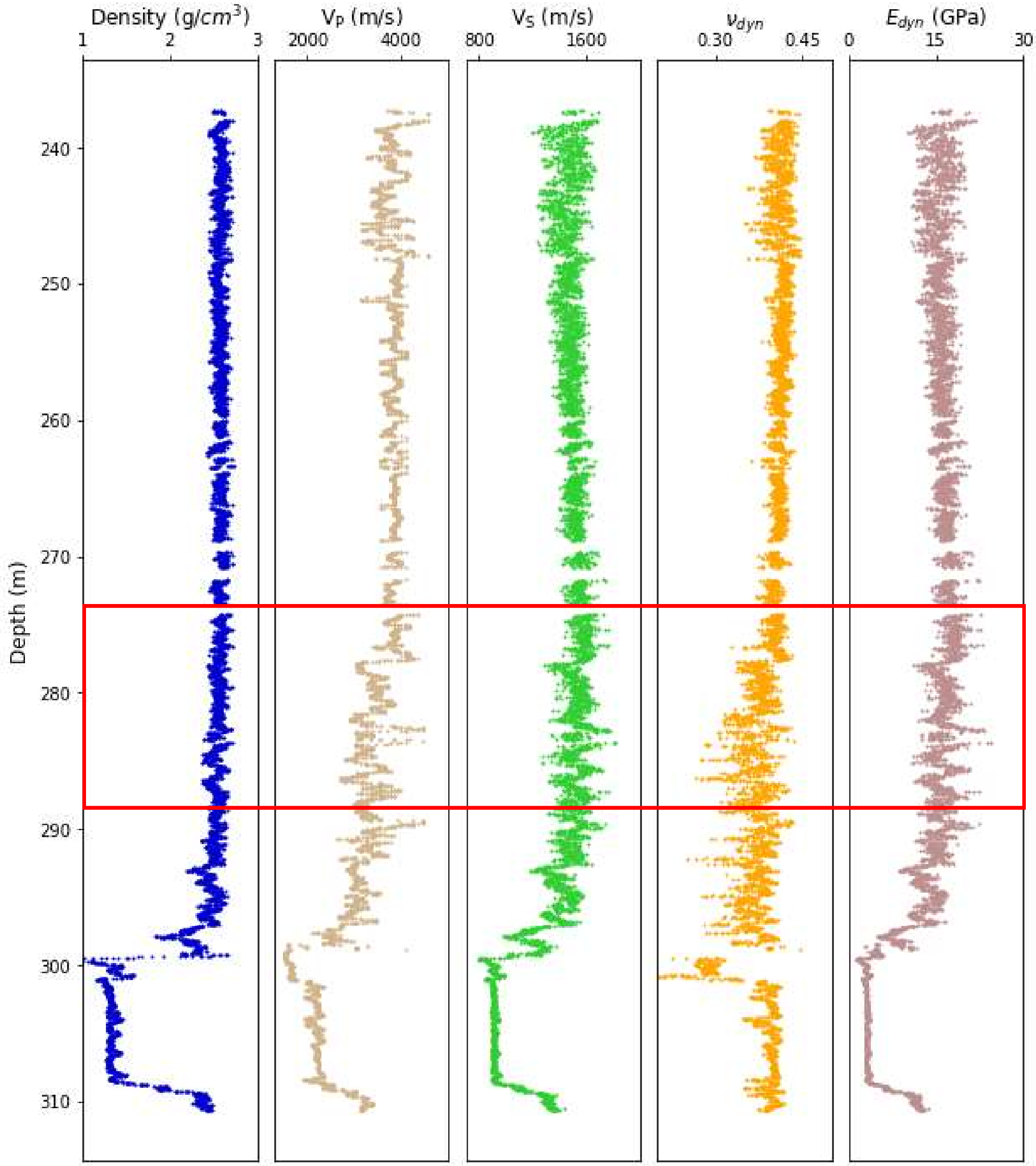


Figure 6.5 Density log and predicted V_P, V_S, E_{dyn} and ν_{dyn} for test borehole number 3

6.4. Comparison between the estimated dynamic elastic properties and static measurements

As explained in section 6.2, *E* and ν were estimated for five test boreholes. The estimated *E* and ν were then averaged over a single sandstone layer and presented in Table 6.1.

Table 6.1 Estimated elastic properties (dynamic) over the sandstone interval in the test boreholes

Test borehole	E_{dyn} (GPa)	ν_{dyn}
T_1	16.5	0.41
T_2	14.3	0.41
T_3	16.6	0.41
T_4	16.6	0.42
T_5	16.2	0.37

In order to assess the validity of the precited elastic properties (dynamic), they need to be compared with the measured elastic properties (static) obtained from the uniaxial compressive tests within the laboratory environment. The experiments were conducted on six sandstone core samples extracted from sandstone formation of mine A (see Table 6.2). All the tests were performed according to ISRM (2007). The axial deformation of tested rock samples was measured through on sample strain gauge and the ram movement of loading frame). The experimental results along with all the graphs and photos for all six core samples were presented in Figure 6.6 to Figure 6.11The deviation in the measurement of axial deformation by different methods have been extensively discussed in various studies (Masoumi 2013; Masoumi et al. 2014; Roshan et al. 2018b; Alejano et al. 2021).

Table 6.2 Lab measured elastic properties (static) obtained from sandstone formation in mine A

Sample number	E_{stat} (GPa) (Strain gauge)	ν_{stat} (Strain gauge)	E_{stat} (GPa) (Loading frame) (GPa)	UCS (MPa)
S_1	18.6	0.27	5.8	50.9
S_2	17.3	0.20	6.6	53.1
S_3	14.8	0.21	6.8	48.8
S_4	17.9	0.22	7.4	62.3
S_5	16.3	0.25	7.6	62.3
S_6	15.1	0.26	8.1	57.8

From Table 6.2 it is clear that there is a discrepancy between the E_{stat} calculated from the strain gauge and the loading frame. Masoumi (2013) and Masoumi et al. (2014) extensively investigated such a discrepancy for sandstone who demonstrated that the estimated E_{stat} from the on-sample transducer can be up to three times larger than that calculated from the ram movement of loading frame. The reported E_{stat} values from the two methods in Table 6.2 are match with that reported by Masoumi (2013) and Masoumi et al. (2014) for sandstone.

Some discrepancies exist between predicted dynamic and measured static mechanical properties of rocks. Previous studies have shown that dynamic elastic properties are usually higher than static ones for sedimentary rocks (Zisman, 1933; Ide, 1936). However, this ratio is equal to one for elastic materials like steel (Ledbetter, 1993). The effects of stress and confining pressure on microcracks and rock microstructure have been identified as major reasons for the differences between dynamic and static mechanical properties in sedimentary rocks, particularly sandstone. Microcracks may close under the high-stress conditions where dynamic properties are measured, resulting in increased wave velocity and, consequently, higher elastic properties (Brace, 1965; Walsh, 1966; Jizba, 1991; Brotons et al., 2014; Ciccotti et al., 2004). For sandstone samples, the dynamic-to-static elastic moduli ratio ranges from 0.8 to 1 at high confining pressures and from 1 to 1.7 at low confining pressures. Another factor contributing to these discrepancies is pore fluid saturation. It is expected that pore fluid supports some of the elastic loading of rock during high-frequency dynamic measurements, whereas in static test conditions, the pore fluid is drained out of the stressed regions (Yale and Jamieson Jr, 1994). Thus, differences in the nature and testing conditions between dynamic and static measurements account for the observed discrepancies.

Although no conversion from dynamic to static mechanical properties are yet conducted, a comparison between Tables 6.2 and 6.3 reveals that the E_{dyn} values estimated by the sonic logs are very close to E_{stas} values calculated from the strain gauge. The dynamic mechanical properties of intact rocks are measured from a core sample within ultrasonic frequencies (100 kHz to 1 MHz) (Cheng and Johnston 1981; Jizba, 1992; Fjær 2009; Olsen et all 2008), while the downhole sonic log measurements are performed with a typical frequency of 10 kHz (van Heerden 1987; Elkatatny et al. 2018). Zisman (1933) and Ide (1936) demonstrated that in the sedimentary rocks, the dynamic elastic properties are generally greater than the static ones due to dispersion, while in homogenous materials having no porosity such as steel, the dynamic and static elastic properties are identical (Ledbetter 1993). Tavener et. al (2017) developed two empirical correlations between dynamic and static elastic properties for sedimentary formations within the Bowen basin as follows:

$$E_{stat} = 0.32\ E_{dyn} \tag{6.3}$$

$$v_{stat} = 0.7\ v_{dyn} + 0.06 \tag{6.4}$$

Using Equations 6.3 and 6.4, the estimated dynamic elastic properties from the log data (Table 6.1) were converted to the static elastic properties for the five selected test boreholes as shown in Table 6.2.

Table 6.3 The resulting static elastic properties over the sandstone interval in the test boreholes through empirical correlations between dynamic and elastic properties suggested by Tavener et. al (2017)

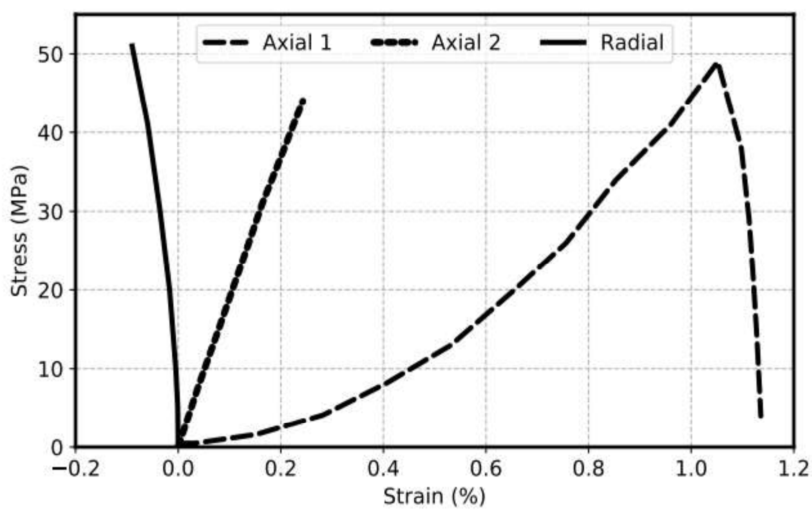
Test borehole	Converted E_{stat} (GPa)	Converted ν_{stat}
T1	5.3	0.35
T2	4.6	0.35
T3	5.3	0.35
T4	5.3	0.36
T5	5.2	0.32

Comparing the static elastic properties in Tables 6.2 and 6.3 shows that the converted E_{stat} values by Equation (6.3) are reasonably in good agreement with the calculated E_{stat} from the ram movement of loading frame. Whereas earlier it was shown that the estimated E_{stat} values from the strain gauge are close to the E_{dyn} values obtained from the sonic log data directly. Latter is consistent with the findings by Olsen et al. (2008) who argued that the use of on-sample transducer (or strain gauge) can minimise the error during the deformation measurement which can be associated with the boundary conditions (Masoumi et. al 2015) or intrinsic microstructure and pre-exiting microcracks in sedimentary rocks (Korinets and Alehossein, 2002).

It is however a significant finding from this study that E_{dyn} might not need conversion to static values that are measured based on ram movement and not on sample displacement. Clearly on sample displacement is more representative of rock mechanical response to loading and it is close to dynamic measurement without any dispersion correction. It should be however noted that the samples used here are dry thus dispersion effect should be minimal. Static elastic properties measured from both testing methods (ram movement and strain gauge) have been presented here. This is important because it is often unclear which method has been used to generate the dynamic to static correlation. By including both methods, we ensure a comprehensive comparison and a clearer understanding of the relationship between dynamic and static properties. This helps in accurately interpreting the results and provides a more reliable basis for further studies and practical applications.

(a1)

(b1)

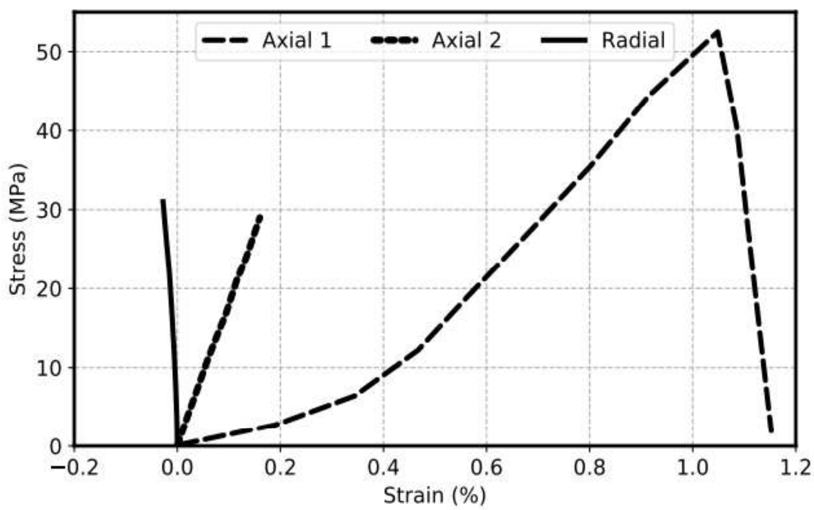


(c1)

Figure 6.6 Samples 1- tested under uniaxial compression: (a) before and (b) after test. (c) Axial 1: measurement of axial deformation by the ram movement of loading frame, Axial 2 and Radial: measurements of axial and radial deformations by strain gauges, respectively.

(a2)

(b2)

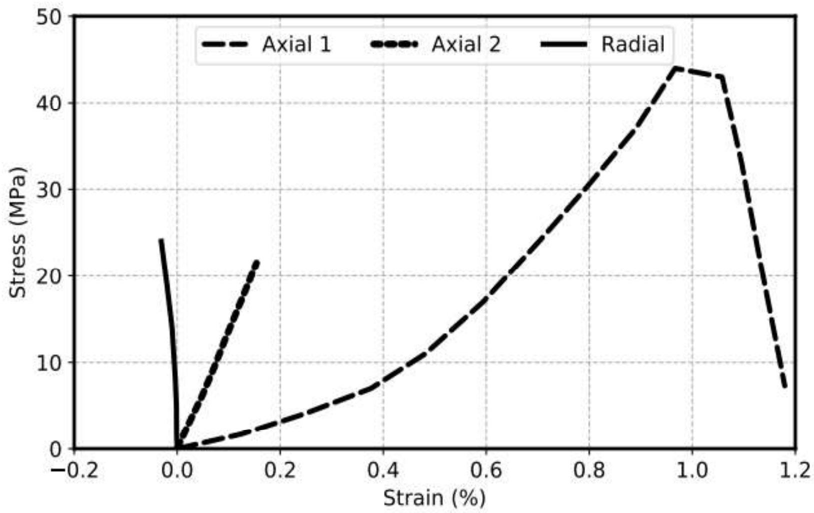


(c2)

Figure 6.7 Samples 2- tested under uniaxial compression: (a) before and (b) after test. (c) Axial 1: measurement of axial deformation by the ram movement of loading frame, Axial 2 and Radial: measurements of axial and radial deformations by strain gauges, respectively.

(a3)

(b3)

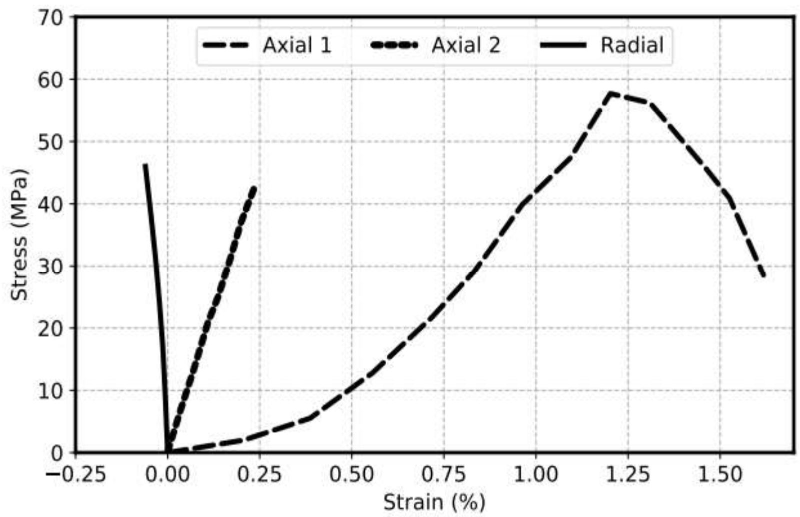


(c3)

Figure 6.8.Samples 2- tested under uniaxial compression: (a) before and (b) after test. (c) Axial 1: measurement of axial deformation by the ram movement of loading frame, Axial 2 and Radial: measurements of axial and radial deformations by strain gauges, respectively

(a4)

(b4)

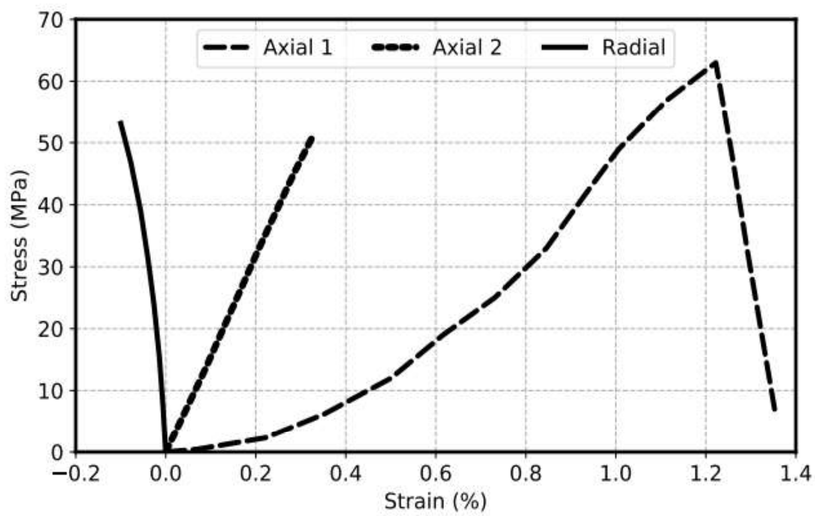


(c4)

Figure 6.9. Samples 3- tested under uniaxial compression: (a) before and (b) after test. (c) Axial 1: measurement of axial deformation by the ram movement of loading frame, Axial 2 and Radial: measurements of axial and radial deformations by strain gauges, respectively.

(a5)

(b5)

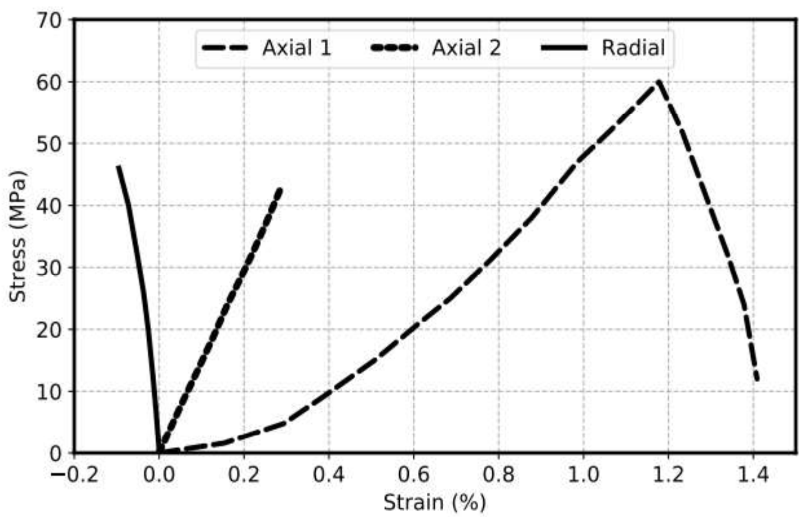


(c5)

Figure 6.10. Samples 4- tested under uniaxial compression: (a) before and (b) after test. (c) Axial 1: measurement of axial deformation by the ram movement of loading frame, Axial 2 and Radial: measurements of axial and radial deformations by strain gauges, respectively.

(a6)

(b6)



(c6)

Figure 6.11. Samples 5- tested under uniaxial compression: (a) before and (b) after test. (c) Axial 1: measurement of axial deformation by the ram movement of loading frame, Axial 2 and Radial: measurements of axial and radial deformations by strain gauges, respectively.

6.5. References

ISRM (2007) The complete suggested methods for rock characterization, testing and monitoring: 1974–2006 ISRM. In: Ulusay R, Hudson JA (eds) Suggested methods prepared by the commission on testing methods. ISRM Turkish National Group, Kozanofset, Ankara

Masoumi H, Saydam S, Hagan P C (2015) A modification to radial strain calculation in rock testing. Geotech Test J 38(6):813-822

Lv A, Aghighi M A, Masoumi H, Roshan H (2021a) The Effective Stress Coefficient of Coal: A Theoretical and Experimental Investigation. Rock Mech Rock Eng:1-17

Roshan H, Lv A, Xu Y, Masoumi H, Regenauer-Lieb K (2018a) New generation of hoek cells. Geotech Test J 42(3):747-760

Roshan H, Masoumi H, Zhang Y, Al-Yaseri A Z, Iglauer S, Lebedev M, Sarmadivaleh M (2018b) Microstructural effects on mechanical properties of shaly sandstone. J Geotech Geoenviron 144(2):06017019

Raaen A, Hovem K, Joranson H, Fjaer E (Year) of Conference. FORMEL: A step forward in strength logging. SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers.

Takahashi T, Takeuchi T, Sassa K (2006) ISRM suggested methods for borehole geophysics in rock engineering. Rock Mech Rock Eng 43(3):337-368

Masoumi H (2013) Investigation into the mechanical behaviour of intact rock at different sizes. Doctor of Philosophy thesis, University of New South Wales, Sydney, Australia

Masoumi H, Bahaaddini M, Kim G, Hagan P (2014). Experimental investigation into the mechanical behavior of Gosford sandstone at different sizes. 48th US Rock Mechanics/Geomechanics Symposium (ARMA), 1-4 June, Minneapolis, US, vol 2, pp 1210,1215.

Alejano L, Estévez-Ventosa X, González-Fernández M, Walton G, West I, González-Molano N,Alvarellos J (2021) A Method to Correct Indirect Strain Measurements in Laboratory Uniaxial and Triaxial Compressive Strength Tests. Rock Mech Rock Eng:1-28

Masoumi H, Bahaaddini M, Kim G, Hagan P (2014). Experimental investigation into the mechanical behavior of Gosford sandstone at different sizes. 48th US Rock

Mechanics/Geomechanics Symposium (ARMA), 1-4 June, Minneapolis, US, vol 2, pp 1210,1215.

Cheng C, Johnston D H (1981) Dynamic and static moduli. Geophys Res Lett 8(1):39-42

Jizba D L (1992) Mechanical and acoustical properties of sandstones and shales. Doctor of Philosophy thesis, Stanford University.

Fjær E (2009) Static and dynamic moduli of a weak sandstone. Geophysics 74(2):WA103-WA112

Olsen C, Christensen H F, Fabricius I L (2008) Static and dynamic Young's moduli of chalk from the North Sea. Geophysics 73(2):E41-E50

van Heerden W (1987). General relations between static and dynamic moduli of rocks. International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts, Elsevier.

Elkatatny S, Mahmoud M, Mohamed I, Abdulraheem A (2018) Development of a new correlation to determine the static Young's modulus. Journal of Petroleum Exploration and Production Technology 8(1):17-30

Zisman W (1933) Comparison of the statically and seismologically determined elastic constants of rocks. Proceedings of the National Academy of Sciences of the United States of America 19(7):680

Ide J M (1936) Comparison of statically and dynamically determined Young's modulus of rocks. Proceedings of the National Academy of Sciences of the United States of America 22(2):81

Ledbetter H (1993) Dynamic vs. static Young's moduli: a case study. Mater Sci Eng 165(1):L9-L10

Tavener E, Flottmann T, Brooke-Barnett S (2017) In situ stress distribution and mechanical stratigraphy in the Bowen and Surat basins, Queensland, Australia. Geological Society, London, Special Publications 458(1):31-47

Masoumi H, Saydam S, Hagan P C (2015) A modification to radial strain calculation in rock testing. Geotech Test J 38(6):813-822

Olsen C, Christensen H F, Fabricius I L (2008) Static and dynamic Young's moduli of chalk from the North Sea. Geophysics 73(2):E41-E50

Chapter 7

7. Conclusions and Recommendations

7.1. Conclusions

Compressional (V_P) and shear wave velocities (V_S) play pivotal roles in diverse engineering applications, spanning geological, hydrogeological, geophysical, and geomechanical studies. The limited availability of measurements for these parameters highlights the need for developing estimation models. The utilization of Machine Learning (ML) to estimate petrophysical properties, including V_P and V_S, from downhole geophysical logs has gained notable attention in the geoscience community. This study centres on evaluating the suitability of ML algorithms for creating predictive models for spatially variable data, specifically V_P and V_S. The primary ML algorithm employed was least squares support vector regression (LSSVR) due to its popularity and flexibility for training and optimization. However, the selection of the most accurate algorithm falls outside the scope of this research. The proposed framework involves: I) establishing a workflow for assessing spatial variability and selecting data from downhole geophysical logs in ML exercises, II) introducing a novel ML-assisted kriging technique designed for spatially variable petrophysical data, and III) developing an integrated ML and rock physics model for converting laboratory ultrasonic data to field-scale sonic measurements.

The evaluation process initiated by examining an extensive dataset containing geophysical logs from 50 boreholes in a coal mine within the Bowen Basin, QLD, Australia. Employing the K-mean algorithm for clustering analysis aimed to identify homogeneous clusters, but no distinct clusters were evident. Out of the 50 boreholes, 45 were allocated for training predictive models, with the remaining 5 for validation. To address data variability and introduce randomness, multiple training sets were generated using a random number generator. Each set varied in the number of included boreholes, ranging from 1 to 40, with each set undergoing five repetitions for stochastic considerations, resulting in 30 distinct training datasets. These datasets were further partitioned into segments from 10% to 100% to determine the optimal data amount required from each borehole, effectively capturing vertical variations. This strategy sought a balance between accuracy and computational efficiency, yielding 300 final training datasets instrumental in model training. The careful curation of these datasets ensures model robustness and adaptability to diverse inputs. Utilizing the least squares support vector regression method, these datasets trained the models, which were subsequently evaluated using separate test data. Based on the analysed data, the following conclusions can be drawn:

- In this study, data from one borehole can significantly differ from another, indicating higher spatial variability under horizontal direction than the vertical direction (considering all rock types together). This is supported by the average standard deviation of V_P under horizontal direction being approximately 560.76, while the vertical standard deviation ranges between 270 and 460.10 across different rock types. The high spatial variability horizontally implies that the local geological factors play a crucial role in subsurface properties.
- The results intriguingly suggest that the vertical variation in the study dataset is either negligible or 10% randomly selected data can capture it. Therefore, increasing the portion of training data does not enhance predictive model accuracy.
- The R^2 trends of all predictive models indicate that horizontal variation is more critical than the vertical variation for model development. In other words, when training a ML model, giving priority to including data from various boreholes is more important than collecting additional data points from a restricted number of boreholes.

To delve deeper into the influence of spatial variability on the development of rock physical properties, a hybrid approach integrating machine learning and geostatistical simulation was proposed. Drawing inspiration from regression kriging, a hybrid model was crafted by combining LSSVR with simple kriging. This hybrid model aims to capture non-linear relationships between data sources and precisely estimate sonic compressional wave velocity. It effectively addresses both the correlation between auxiliary and target variables and the spatial autocorrelation of the data. The model was developed using data from an expanded dataset of 122 boreholes within the same study area, enhancing its generalizability.

The proposed methodology entails an iterative optimization process for the LSSVR model to predict the target variable. In each iteration, distinct LSSVR and semivariogram models are fitted to the target datasets. While the LSSVR model captures relationships between V_P and auxiliary variables to a certain extent, the potential presence of additional spatial terms is acknowledged. The residual of the LSSVR model is assumed to represent a combination of these terms, modelled using kriging in a manner similar to regression kriging. Semivariogram modelling is automated, utilizing a nested model that linearly combines two semivariogram models. The iterative process persists until a satisfactory result is achieved, as determined by the calculated R^2 for the validation data. Iteration concludes when the difference between the R^2 of two consecutive iterations falls below 0.001, signifying stabilization. This modelling approach ensures a robust integration of machine learning and geostatistical methods, contributing to a good understanding of spatial variability's impact on rock physical properties. The proposed methodology was also compared with sole kriging and sole LSSVR and proved to outperform both of them.

In addition to exploring the methodology for constructing predictive models, this study delved into alternative data sources and substitutes for sonic logs. Given the scarcity of measurements for sonic shear wave velocity in the mining industry, acquiring sufficient data for the development of ML or geostatistical predictive models poses a challenge. To overcome this limitation, a dataset comprising laboratory ultrasonic measurements and introduced was leveraged an innovative model to convert these lab measurements to field sonic data. This transformation was achieved through a hybrid approach incorporating LSSVR and squirt flow models. A total of 178 core samples, encompassing measured density, porosity, ultrasonic compressional and shear wave velocities (UV_P and UV_S), were collected. Subsequently, two predictive models based on laboratory data were formulated to estimate UV_S using both the LSSVR and MLR algorithms. The efficacy of these predictive models was assessed by testing their ability to predict (SV_S, with the mean ultrasonic velocity converted to sonic velocity using the squirt flow theoretical model. The outcomes demonstrated that the developed models, particularly the LSSVR-based model, exhibited a high predictability for field-measured SV_S, achieving an R^2 of 0.93. The MLR-based model also demonstrated considerable predictability with an R^2 of 0.87. This study's results affirm that the proposed workflow can accurately estimate missing sonic log data (e.g., SV_P, SV_S) from laboratory ultrasonic measurements, showcasing its practical utility with high accuracy.

7.2. Recommendations for future work

The following recommendations are presented for future research as a continuation of the contributions provided in this thesis:

- The workflow devised in Chapter 3, tailored to our specific study area, can be extended to different locations with distinct geological characteristics. By applying the workflow in varied settings, researchers can assess the workflow's adaptability and effectiveness across diverse geological conditions. This comparative analysis would contribute to the development of a more universally applicable workflow, taking into account location-specific variations under horizontal and vertical variability.
- In Chapter 4, data were averaged over a geological layer, and a 2D semivariogram was developed. To further enrich the modelling process, future research could explore the

development of semivariograms under the vertical direction. This extension would bring additional spatial data into the hybrid model, fostering a more comprehensive understanding of spatial variability and enhancing the overall regression modelling.

- In chapter 5, due to the unavailability of shear logs in the study mine, test data were extracted from a neighbouring mine. Future research could validate the methodology in a study area where both laboratory and full sonic log data are available. This comparison is expected to provide valuable insight into the model's performance under different data scenarios, potentially enhancing its accuracy and reliability.
- In Chapter 5, the ML model was integrated with a simple theoretical squirt flow model. Future research could explore the integration of more advanced rock physics models to further refine the predictive capabilities of the model. By incorporating sophisticated rock physics frameworks, the model's accuracy in predicting sonic compressional wave velocity can be improved, offering a more nuanced understanding of the complex relationships within the data.
- In Chapter 6, we briefly discussed the application of developing predictive models for a sandstone layer in the study area. Future research should aim to thoroughly investigate this application by addressing several key aspects. Firstly, there is a need to understand the impact of geological discontinuities, such as fractures and faults, on the predictive models. These features can significantly influence the mechanical properties and behavior of rock masses, and their inclusion is essential for improving model accuracy. Furthermore, advanced imaging techniques, such as micro-CT scanning and seismic imaging, should be employed to obtain high-resolution data on rock discontinuities and heterogeneities, which can be integrated into the predictive models to improve their accuracy.
- In this study, machine learning (ML) algorithms were used for regression and clustering purposes. However, there is potential for using ML algorithms for image processing in subsurface areas. This could be valuable for seismic interpretation, which would help improve the developed models. Additionally, downhole image logs could provide meaningful insights if interpreted using robust ML models. A workflow that includes seismic and downhole image log interpretation could make a significant contribution to the field.

- While this thesis primarily focused on the application of Machine Learning (ML) for sonic wave velocities, there is potential for extending the study to encompass other petrophysical and spatial variable data. A comparative analysis of the results across various datasets could offer valuable insights into the versatility and effectiveness of ML methodologies in different domains within the geoscience