Development of soil spectroscopy calibration and prediction models for precision agriculture within South Africa

A Kock

orcid.org/ 0000-0001-5349-9798

Thesis accepted in fulfilment of the requirements for the degree *Doctor of Philosophy in Science with Environmental Sciences* at the North-West University

Promoter: Prof GM van Zijl

Graduation: June 2025

PREFACE

This dissertation is presented as a compilation of six chapters, with Chapters 3 through 5 formatted as individual publications. At the time of submission, these chapters are under preparation for submission to peer-reviewed journals. Presenting the thesis in this format results in some repetition across chapters, and significant efforts have been made to minimise redundancies without compromising the integrity and completeness of each paper. Specifically, performance metrics equations were provided for all three publications as it is standard practice in soil spectroscopy to include them to avoid confusion on how these metrics are used and calculated.

Chapter 1 provides an overarching introduction to soil spectroscopy, its relevance to precision agriculture and sustainable land management, and outlines the problem statement, research aims, and objectives. Chapter 2 delivers a comprehensive review of the theoretical foundations and the advancements in soil spectroscopy, particularly the integration of machine and deep learning and the critical need for localised soil spectral models in South Africa. This chapter lays the groundwork for the research papers that follow. Chapters 3, 4, and 5 represent the core of this dissertation and are presented as standalone research papers. Chapter 3 investigated the integration of combined mid-infrared (MIR) and near-infrared (NIR) spectral data to enhance soil property predictions. This chapter will be submitted to Geoderma for open-access publication. Chapter 4, which is a follow-up using the findings from the previous chapter, explored the application of convolutional neural networks (CNN) for predicting soil properties using combined MIR and NIR data. This chapter will be submitted to the European Journal of Soil Science. Chapter 5 examined "spiking" global soil spectral libraries with localised data from South Africa to improve predictive accuracy. This chapter will be submitted to Catena. Finally, Chapter 6 synthesised the findings across all papers, providing a general conclusion and identifying future research directions to address the remaining challenges in soil spectroscopy and its application to precision agriculture in South Africa. This dissertation contributes to soil spectroscopy by developing innovative methods for integrating localised data, leveraging machine learning techniques, and enhancing the applicability of global soil spectral libraries to underrepresented regions. These advancements aim to support the development of precision agriculture and sustainable land management practices in resource-limited and ecologically diverse contexts.

ACKNOWLEDGEMENTS

First and foremost, I express my deepest gratitude to my Savior, Christ, for His enduring grace and strength that carried me through this journey.

To my beloved grandmother, Magda Hefer, thank you for your profound influence on my life. Your steadfast belief in me and your constant encouragement to strive for excellence instilled in me the values that have guided me to this achievement. Although your passing this year has left an irreplaceable void, your memory will forever be a source of inspiration.

To my dearest wife, Alicia Kock thank you for your loving support and patience, especially during these challenging times. Your understanding and encouragement were instrumental in keeping me focused and motivated.

I am eternally grateful to my parents and family for their love and support throughout my life. Your encouragement and belief in my abilities have been a constant source of strength.

To my supervisor and mentor, Prof George van Zijl, thank you for your invaluable guidance, expertise, and constant support. Your insights and feedback were crucial to the success of this research.

I would also like to sincerely thank my colleagues, Willie Cloete and Molebaleng Sehlahpelo, for their assistance and camaraderie—your willingness to help and offer support created a positive and encouraging research environment.

I thank Dr Jaco Koch and Dr Henno Havenga for their valuable contributions. Your expertise and insights were instrumental in shaping my work.

Finally, I would like to thank Martiens Du Plessis at NWK Limited, Dailena Pienaar at NviroTek Laboratories, and Dries Bloem at Geolab for their assistance with analysing my soil samples. I greatly appreciate your expertise and professionalism.

This research would not have been possible without the financial support of the NRF (Grant Numbers: 121302) the Water Research Commission (WRC Report No 3145/1/24). I gratefully acknowledge their funding, which enabled me to pursue this endeavour.

ABSTRACT

This thesis aimed to enhance the accuracy and applicability of soil spectroscopy prediction models for South Africa by addressing the limitations of global soil spectral libraries, for example the Open Soil Spectral Library (OSSL), which often underrepresent South African soils. The research focused on improving soil property predictions by integrating local data, employing advanced machine learning techniques, and optimising the calibration and validation of models for the unique soil conditions of the Western Highveld region. The study utilised a dataset of 772 soil samples from the Western Highveld region, encompassing fundamental agricultural soil properties pH (KCl), P, and exchangeable cations (Ca, Mg, K, Na). Mid-infrared (MIR) and near-infrared (NIR) spectral data were collected and combined to form a comprehensive dataset. Advanced pre-processing techniques were applied to improve spectral data quality, including noise reduction, scatter correction, and feature enhancement. The dataset was then split into training and validation subsets using a conditioned Latin hypercube sampling approach to ensure robust model development. Machine learning techniques were central to the methodology. Cubist regression and convolutional neural networks (CNN) were employed to develop predictive models for soil properties. CNN demonstrated superior performance in capturing complex spectral relationships compared to traditional regression methods. The thesis also investigated the "spiking" process, a targeted integration of local spectral data into the global OSSL library, to address the underrepresentation of South African soils. This approach was evaluated at various spiking levels to determine the optimal balance between local and global data. The core findings highlight the effectiveness of combining MIR and NIR spectral data for improving soil property predictions. Combined spectral models outperformed single-spectrum models, significantly reducing prediction error for properties pH (KCl) and exchangeable Ca. Tailored pre-processing techniques provided additional improvements for specific soil properties, though raw combined spectral data often performed comparably. Machine learning models, particularly CNN, showed high predictive accuracy and robustness, outperforming traditional methods in handling the complexity and variability of soil spectral data. The spiking method demonstrated that integrating local data into global spectral libraries improves prediction accuracy for underrepresented regions. Models incorporating local data achieved superior performance compared to those relying solely on global data. However, excessive spiking introduced overfitting and reduced model generalisation, underscoring the importance of balancing local and global data contributions. This thesis contributes to soil spectroscopy by developing a scalable framework for integrating local soil data into global spectral libraries and demonstrating the value of advanced machine-learning techniques in predictive modelling. By addressing the limitations of current models, the research supports establishing a South African Soil Spectral Library, enabling more

accurate soil property assessments for agricultural and environmental applications. These advancements pave the way for more effective land management and sustainable farming practices in South Africa and other underrepresented regions.

Keywords: Convolutional neural networks, Mid-infrared spectroscopy, Near-infrared spectroscopy, Soil spectral inference, Soil spectral libraries, Spiking

LIST OF ABBREVIATIONS

MIR	Mid-Infrared
NIR	Near-Infrared
OSSL	Open Soil Spectral Library
CNN	Convolutional Neural Network
ML	Machine Learning
R^2	Coefficient of Determination
RMSE	Root Mean Square Error
LCCC	Lin's Concordance Correlation Coefficient
RPIQ	Ratio of Performance to Interquartile Distance
MAE	Mean Absolute Error
cLHS	Conditioned Latin Hypercube Sampling
SNV	Standard Normal Variate
MSC	Multiplicative Signal Correction
SG	Savitzky-Golay Filtering
pH (KCl)	Soil pH measured in potassium chloride solution
Ca	Exchangeable Calcium
Mg	Exchangeable Magnesium
Na	Exchangeable Sodium
K	Exchangeable Potassium
P (Bray-1)	Phosphorus measured using the Bray-1 extraction method
FT-NIR	Fourier Transform Near-Infrared
Vis-NIR	Visible-Near Infrared
DRIFT	Diffuse Reflectance Infrared Fourier Transform
ICRAF	International Centre for Research in Agroforestry
ISRIC	International Soil Reference and Information Centre

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Background

The increasing global human population continues to drive a growing demand for food, and this causes significant challenges for the adoption of sustainable agriculture whilst mitigating climate change and increasing food production (Godfray *et al.*, 2010; Tilman *et al.*, 2011). Addressing these challenges necessitates the adoption of efficient, resource-conscious farming methods (Godfray *et al.*, 2010). Precision agriculture, which is a data-driven approach aimed at optimising natural resource utilisation and maximising crop yields, offers a promising solution (Zhang *et al.*, 2002). Implementing precision agriculture successfully relies on the accurate analysis of soil properties, which in return provides essential and precise insights for informed decision-making on land management strategies, fertilisation recommendations, and other agricultural practices (McBratney *et al.*, 2005). While reliable and widely used, conventional laboratory soil analysis methods present significant limitations. Laboratory soil analysis, which is regarded as the standard approach, is accurate but time-consuming, labour-intensive, and costly(Nanni and Demattê, 2006). Soil spectroscopy has been proven to be an innovative alternative to traditional methods for soil analysis. It offers advantages such as being rapid, cost-effective, environmentally friendly, and non-destructive, making it particularly suitable for applications in resource-constrained regions. Nocita *et al*. (2015) highlighted the potential of spectroscopy to replace traditional methods by reducing costs and enhancing efficiency in soil monitoring. Rossel *et al*. (2006) demonstrated that visible (Vis), near-infrared (NIR), and mid-infrared (MIR) spectroscopy could provide reliable predictions for multiple soil properties, reducing reliance on time-consuming and expensive chemical methods (Rossel *et al*., 2006). The working principle of soil spectroscopy relies on the interaction between electromagnetic radiation and soil constituents, which varies greatly and depends on the soil composition and properties (Wadoux *et al.*, 2021). By analysing the unique soil spectral signatures (patterns of absorption and reflection) produced by soil samples, fundamental soil properties pH, nutrient content, organic matter, and mineral composition can be inferred with high accuracy (Janik et al., 1998).

Three principal techniques dominate the field of soil spectroscopy: visible-near-infrared (Vis-NIR), NIR and MIR spectroscopy. Vis-NIR and NIR spectroscopy excel in detecting organic soil components, for example organic matter and carbon content, while MIR spectroscopy provides detailed insights into the soil mineral compositions (Stenberg et al., 2010). These methods also generate substantial datasets, which necessitates the need and application of advanced analytical tools like machine learning (ML) algorithms and neural networks (NN) to derive

meaningful patterns and enhance predictive accuracy (Padarian *et al.*, 2020). Despite its promise, implementing soil spectroscopy in regions such as South Africa has been challenging. Global soil spectral models often fail to deliver accurate predictions when applied to South African soils due to their distinct properties and environmental conditions, significantly influencing spectral responses (Kock et al., 2024). This variability highlights the need for region-specific models tailored to local soil characteristics. Developing a comprehensive South African Soil Spectral Library is critical for addressing these challenges (Kock et al., 2024). However, progress for accurate local soil property prediction models is slow due to limited access to local spectral datasets and comparatively sparse research efforts in South Africa relative to global initiatives. These constraints underscore the urgent need for dedicated research to expand local spectral data availability and improve existing soil spectroscopy prediction models, particularly for regions with diverse and unique soil characteristics, such as the Western Highveld. Focused efforts to build localised spectral models and libraries will enhance the predictive accuracy and useability of soil spectroscopy in South Africa. This progress is essential for advancing precision agriculture within the region and contributing to global efforts toward sustainable agriculture and food security (Nocita *et al.*, 2015).

1.2 Problem Statement

Traditional laboratory soil analysis methods, although reliable, are time-consuming, labour-intensive, and costly (Nanni and Demattê, 2006). These limitations restrict their widespread applicability, particularly for smallholder farmers with limited resources. In contrast, soil spectroscopy provides a non-invasive, cost-effective, and rapid alternative for determining key soil properties, making it an ideal tool for integrating into precision agriculture frameworks (Dangal *et al.*, 2019). Global soil spectral models often fail to account for the unique mineralogical, environmental, and climatic conditions characteristic of South African soils. This reduces the applicability of global soil spectroscopy prediction models for local contexts, particularly in regions like the Western Highveld of South Africa, where soil variability is significant and diverse. South African soils possess unique characteristics stemming from their distinct pedogenesis, geological formations, and specific soil conditions. These features are shaped by a combination of climate, parent material, and geomorphology, giving rise to diverse mineralogical and chemical properties. As a result, farmers are reliant on expensive soil analysis, without any alternative tools for assessing soil health and managing their soil nutrient status at scale. Currently, South Africa lacks a large comprehensive soil spectral library that is tailored to its unique and diverse soil characteristics, which limits the implementation of precision agriculture practices effectively. In

previous work (Kock et al., 2024) a soil spectral library for the Western Highveld region of South Africa was created, which utilized MIR soil spectral data and ML algorithms like Cubist to predict a limited set of soil properties routinely used for fertiliser recommendations. While successful, the limited scope of soil properties in these models and the sub-par predictive accuracy compared to traditional laboratory methods highlight the need for further research and development. The further development of a South African Soil Spectral Library would address the limitations of current soil spectral inference methods and provide a foundation for creating predictive models capable of delivering accurate soil information for end-users. Such efforts would enable cost-effective, non-invasive soil analysis tools that are essential for integrating precision agriculture into diverse farming systems (Du and Zhou, 2009). Addressing this gap in rapid soil analysis infrastructure will require collecting and integrating large, region-specific multi-spectral datasets with the application of advanced modelling techniques. By improving the predictive accuracy of the existing soil spectroscopy models, these developments will provide a scalable and sustainable alternative to conventional laboratory methods, supporting smallholder farmers, increasing crop productivity, and contributing to global efforts in sustainable agriculture.

1.3 Aim and Objectives

This study aimed to develop and evaluate accurate, region-specific soil property prediction models for the Western Highveld region of South Africa by leveraging the combined strengths of NIR and MIR spectroscopy and investigating different pre-processing techniques and ML deep learning methods on model performance. The specific objectives were to :

1. Create and assess the effectiveness of combining different spectral regions in improving the prediction accuracy of fundamental soil properties, compared to using single spectral regions (NIR or MIR alone).
2. Create and compare the performance of a machine-learning algorithm Cubist against deep learning convolutional neural networks (CNN) in predicting soil properties using the combined pre-processed NIR and MIR spectral data.
3. To investigate the effectiveness of "spiking" a global soil spectral library with soil spectral data from the Western Highveld region of South Africa in improving the prediction accuracy of local prediction and global soil property models.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

This chapter will comprehensively review the literature on soil spectroscopy and its use and applications in predicting important soil properties routinely used in precision agriculture. This review will focus on challenges, advancements, and the relevance of region-specific data. The aim is to critically examine the integration of NIR and MIR spectroscopy for enhanced soil property predictions, particularly in the South African context. Given traditional soil analysis methods' limitations for precision agriculture adoption, this review highlights how advancements in spectroscopic techniques and the development of spectral libraries address these challenges, offering the possibility of rapid, cost-effective, and accurate soil property predictions. The discussion will include the theoretical foundations of soil spectroscopy, including its role in soil analysis and the mechanisms through which electromagnetic radiation interacts with soil components. Furthermore, the review will delve into the importance of soil spectral libraries, emphasising their construction and calibration, and will explore innovative methodologies, for instance the "spiking" of global libraries with localised data, to improve regional model prediction accuracy. The inclusion of machine learning, particularly CNN, in spectroscopy-based modelling will also be evaluated as a promising approach to managing complex soil variability and enhancing prediction robustness.

This review's scope reflected the PhD research's complex objectives, including addressing the challenges related to soil variability in South Africa's Western Highveld region, integrating regional data into global prediction frameworks, and leveraging deep learning models to improve soil property predictions.

2.2 Theoretical Foundations and Key Concepts

2.2.1 Basics of Soil Spectroscopy

Soil spectroscopy involves using electromagnetic radiation (EMR) to analyse soil properties based on the interaction between EMR and soil constituents. When exposed to specific wavelengths, these interactions can cause vibrational, rotational, and electronic transitions of molecules (Atkins and De Paula, 2010). The Vis, NIR and MIR regions (sometimes the combination thereof) of the EMR spectrum are particularly valuable for the purpose of soil spectroscopy and are the most frequently used, while fewer studies have employed ultra-violet (UV), and x-ray fluorescence (XRF) (Ge et al., 2011; Li et al., 2021; O'Rourke et al., 2016; Tavares

et al., in press). NIR, which ranges from 4 000 cm^{-1} to 7 000 cm^{-1} (350 to 2500 nm) detects overtones of molecular vibrations of O-H, N-H, and C-H bonds, is valued for its rapid data collection and minimal sample preparation, making it particularly effective for examining organic and moisture-related properties (Hollas, 2004; Stenberg Viscarra Rossel, et al., 2010). MIR spectroscopy, on the other hand, ranges from 400 cm^{-1} to 4 000 cm^{-1} and focuses on fundamental molecular vibrations, offering detailed and precise information on soil components such as texture, organic matter, and carbonates (FAO, 2022; McCarty and Reeves, 2006). The combination of these approaches integrates NIR's efficiency with MIR's accuracy, yielding robust results (Nyawasha *et al.*, 2024).

2.2.2 Pre-Processing Techniques for Spectral Data

Spectral data collected from soil samples often includes noise, baseline shifts, and scatter effects, necessitating pre-processing to enhance data quality (FAO, 2022). Effective pre-processing ensures that the extracted spectral features accurately represent soil properties rather than artefacts (Vašát *et al.*, 2017).

Essential pre-processing methods include:

- Standard Normal Variate (SNV)

Standard Normal Variate (SNV) is a pre-processing technique that normalises spectral data to correct for scatter effects caused by variations in particle size, packing, or surface roughness (Wadoux *et al.*, 2021). This method works by subtracting the mean of the spectrum from each data point and then dividing the result by the standard deviation of the spectrum (Riedel *et al.*, 2018; Wadoux *et al.*, 2021). The formula for SNV is given in Equation 2-1:

Equation 2-1

$$SNV = \frac{X_i - \mu_{x_i}}{\sigma_{x_i}}$$

where X_i represents the raw spectrum for the i-th sample, μ_{x_i} is the mean of the spectrum, and σ_{x_i} is the standard deviation (Wadoux *et al.*, 2021). SNV is particularly simple yet effective for reducing light scatter in soil samples and is widely used in soil spectroscopy to ensure that spectral features are not overshadowed by physical artefacts of the sample (Hollas 2004; Wadoux et al., 2021).

- Multiplicative Signal Correction (MSC)

Multiplicative Signal Correction (MSC) addresses non-linear distortions in spectral data caused by scatter or other multiplicative effects (Canero *et al.*, 2024). It aligns each spectrum with a reference spectrum, usually the mean spectrum of the dataset, to standardise the data (Canero *et al.*, 2024). The method assumes that the raw spectrum x_i can be expressed as a linear combination of an additive term a_i and a multiplicative term m_i, given by Equation 2-2:

Equation 2-2

$$MSC = \frac{x_i - a_i}{m_i}$$

where a_i and m_i are determined by fitting each spectrum to the reference spectrum, x_r $(FAO, 2022)$. This technique effectively removes scatter effects and other multiplicative distortions, making it a valuable pre-processing step in soil spectroscopy (Canero *et al.*, 2024). MSC is often used in combination with other methods, for example Savitzky-Golay filtering (SG), to enhance data quality further (FAO, 2022; Wadoux *et al.*, 2021).

- Savitzky-Golay Filtering (SG)

Savitzky-Golay filtering (SG) is a versatile method to smooth spectral data, reduce noise, and compute derivatives for improved resolution of overlapping peaks. The technique applies a polynomial regression (with an order k) within a defined window size ($w = 2g + 1$), where g determines the number of neighbouring data points considered (Wadoux *et al.*, 2021). It can also compute the first or second derivatives of the data to enhance spectral features. While effective for noise reduction, SG filtering has a drawback: data points at the beginning and end of the spectrum are lost proportional to the window size. For example, an w = 11 (where g = 5) will result in losing 5 data points at each end. Despite this limitation, SG filtering is widely used in soil spectroscopy for both smoothing and feature enhancement (Savitzky and Golay, 1964; Wadoux et al., 2021).

- Continuum Removal

Continuum removal is a technique used to isolate absorption features in a spectrum by removing the continuous baseline, effectively highlighting subtle and overlapping absorption bands (Gomez *et al.*, 2008). This method is beneficial for visualising specific features of the spectrum, for instance those associated with soil organic matter or minerals and is often employed in exploratory data analysis (Gomez et al., 2008; Stevens and Ramirez-Lopez, 2013). While continuum-removed spectra are less commonly used for predictive modelling, the technique is

invaluable for identifying and interpreting specific absorption features in complex datasets (Stevens and Ramirez-Lopez, 2013).

- Spectral Compression and Dimension Reduction

Spectral compression and dimension reduction are essential in soil spectroscopy to handle high-dimensional datasets where many wavelengths are correlated, introducing redundancy and noise (Ng et al., 2022). Principal Component Analysis (PCA) is a widely used technique for this purpose, transforming spectral data into a smaller set of uncorrelated variables called principal components (PCs) (Safanelli *et al.*, 2023). These PCs are linear combinations of the original variables, ranked by their contribution to total variance, with the first few typically capturing the majority of information (Safanelli *et al.*, 2023). PCA involves standardising the data, computing a covariance matrix, extracting eigenvalues and eigenvectors, and projecting the data onto these eigenvectors (Safanelli *et al.*, 2023). The transformation simplifies data interpretation, reduces noise, and minimises computational demands, making PCA particularly effective for pre-processing before machine learning (Safanelli *et al.*, 2023). While it eliminates redundant information and reduces overfitting, PCA's linear nature may overlook non-linear relationships in the data. Nonetheless, it remains a powerful tool for spectral data analysis, supporting improved visualisation, clustering, and predictive modelling in soil spectroscopy (Safanelli *et al.*, 2023).

2.2.3 Dataset Splitting

Splitting soil spectral libraries into training and validation datasets is essential for developing reliable predictive models that are robust and reliable for making predictions on independent validation sets of data. Different splitting methods have been developed and used to account for the complexity and variability of spectral libraries. As demonstrated in soil spectroscopy and machine learning studies, each method offers unique advantages and applications. Simple random sampling is one of the most straightforward and widely used methods for creating training and validation sets (Reitermanova et al., 2010). In this approach, samples are divided randomly into training and validation sets, typically in ratios of 70:30 (Sila, 2016), 75:25 (Wadoux *et al.*, 2021) or even 50:50 (Seybold *et al.*, 2019). Random splitting ensures statistical representation across subsets and is computationally efficient. However, it may fail to account for spatial dependencies or systematic biases in geographically heterogeneous datasets (Reitermanova et al., 2010). Shepherd and Walsh (2002) employed random splitting to validate models predicting soil properties from a spectral library of East and Southern African soils. While effective for general performance assessment, they noted that random splits might underestimate errors for region-specific predictions.

Spatial splitting divides datasets based on geographical or environmental boundaries, assigning samples from specific regions to either the training or validation set (Wang *et al.*, 2012). This method evaluates model performance in predicting soil properties across distinct geographic areas and is especially useful for large-scale applications (Wang *et al.*, 2012). For example, Padarian et al. (2019) used spatial splitting with the LUCAS Soil Spectral Library to test the transferability of models across European regions, revealing its suitability for validating models on geographically distinct datasets. K-fold cross-validation is an iterative method where the library is divided into K subsets (folds), with each subset used as the validation set. In contrast, the remaining folds are used for training (Reitermanova et al., 2010). This ensures every sample is included in training and validation, providing a robust measure of model performance (Nyawasha *et al.*, 2024). Cluster-based splitting groups samples into clusters based on spectral or soil property similarities before splitting them into training and validation sets (Yen and Lee, 2009). This ensures that the subsets represent the full range of variability within the dataset, reducing prediction bias. Cluster-based splitting improves model robustness, especially in heterogeneous datasets, by ensuring all soil types are well-represented in training and validation sets (Ogen *et al.*, 2019). Conditioned Latin Hypercube Sampling (cLHS) is a stratified sampling method that ensures the training and validation sets cover the entire range of covariates in the dataset (Minasny and Mcbratney, 2006). This method benefits datasets with complex relationships between spectral features and soil properties (Waruru *et al.*, 2015). This cLHS was used to optimise sampling designs for Random Forest models, achieving better spatial coverage and reducing prediction errors compared to random splitting (Wadoux *et al.*, 2019). Leave-One-Out Cross-Validation (LOO-CV) excludes predefined groups, for example certain soil types or regions, entirely from the training set, using them only for validation (Xu and Goodacre, 2018). This method effectively tests model generalizability to unobserved groups (Xu and Goodacre, 2018). These splitting strategies provide researchers with diverse tools to optimise model training and validation for soil spectral libraries. By selecting methods tailored to the characteristics of their datasets and research objectives, researchers can achieve more accurate and generalisable models for soil property predictions.

2.3 Machine Learning for Soil Spectral Inference

Numerous approaches are used for calibrating prediction models and predicting soil properties from soil spectral data, ranging from statistical methods to advanced machine learning algorithms. Standard techniques include Cubist regression trees (Dangal *et al.*, 2019; Minasny *et al.*, 2013), memory-based learning (MBL) (Dangal *et al.*, 2019), Random Forest (RF) (Breiman and Cutler, 2018), Partial Least Squares Regression (PLSR) (Morellos *et al.*, 2016), Principal Component Analysis (PCA) (Mouazen *et al.*, 2010), Artificial Neural Networks (ANN) (Kuang *et al.*, 2015),

Multivariate Adaptive Regression Splines (MARS) (Clingensmith and Grunwald, 2022a), Support Vector Machines (SVM) (Shao and He, 2011), Multiple Linear Regression (MLR) (Bayer *et al.*, 2012), and CNN (Ng *et al.*, 2019).

Numerous studies have explored applying machine learning techniques to predict soil properties effectively. Rossel and Behrens (2010) compared algorithms MLR, PLSR, RF, SVM, ANN, MARS and Boosted trees (BT) for modelling soil diffuse reflectance spectra, demonstrating their potential for estimating soil SOC (SOC), clay content and pH_{water}. Similarly, Morellos et al. (2016) highlighted the advantages of methods like PLSR, Cubist and SVM for predicting soil nitrogen, SOC, and moisture content using Vis-NIR spectroscopy. Clingensmith and Grunwald (2022) evaluated techniques for instance Cubist, RF, MARS and PLSR for the nationwide predictions of soil properties including SOC, total N, total S, clay, sand, exchangeable Ca, CEC and pH for the United States of America, showing the effectiveness of machine learning models in handling large and diverse datasets. Furthermore, advanced approaches like CNN have been employed to improve prediction accuracy, as demonstrated by Yang et al. (2020), who combined CNN with Recurrent Neural Networks for robust soil property analysis. Similarly, Padarian et al. (2019) showcased the superior performance of CNN over traditional methods such as PLSR and Cubist for analysing soil spectral data across large regions. These studies underscore the versatility and effectiveness of machine learning methods in enhancing soil property predictions.

2.3.1 Cubist

Cubist is a machine learning technique that combines decision trees and instance-based learning to create predictive models (Minasny *et al.*, 2013). It builds rule-based regression models, where each rule is associated with a linear regression equation, allowing it to handle complex relationships and continuous variables effectively (Minasny *et al.*, 2013; Wadoux *et al.*, 2021). Cubist excels in capturing non-linear relationships while maintaining model interpretability. It is particularly effective for soil spectroscopy because it can integrate information from multiple variables and account for intricate patterns in spectral data. For NIR and MIR spectroscopy, Cubist is preferred due to its ability to effectively utilise both broad spectral trends and localised absorption features. A study by Clingensmith and Grunwald (2022) demonstrated Cubist's superior performance compared to PLSR, RF and MARS in predicting SOC, N, and other properties from large and diverse datasets, as its rule-based approach handles spectral variability well. Padarian *et al*. (2019) also highlighted that Cubist is adept at managing high-dimensional spectral data, outperforming linear models like PLSR in scenarios involving complex soil spectral libraries. This adaptability to spectral variability and its robustness in capturing fundamental soil properties make Cubist a preferred choice for NIR and MIR soil spectroscopy applications.

2.3.2 Convolution Neural Networks

Convolutional Neural Networks (CNN) have emerged as a powerful tool for analysing soil properties, particularly in the context of soil spectroscopy. Initially developed for image processing, CNN excel at recognising patterns in high-dimensional datasets, making them ideal for analysing spectral data (Kamilaris and Prenafeta-Boldú, 2018). They have been successfully applied to predict various soil properties, which include: SOC, N, and CEC, with high accuracy. For instance, Padarianet al. (2019) Demonstrated that CNN outperformed traditional methods like PLSR and Cubist regression in analysing the LUCAS soil database, reducing prediction error by 87% compared to PLSR and 62% compared to Cubist.

In soil spectroscopy, CNN analyse spectral data either as one-dimensional (1D) sequences or two-dimensional (2D) spectrograms (Padarian et al., 2019). For example, Ng et al. (2019) demonstrated that CNN significantly improved prediction accuracy for soil properties when combining Vis-NIR and MIR spectral data, achieving R^2 values of 0.95-0.98 for multiple soil attributes. Traditionally, CNN models are trained in a single-task learning (STL) framework, where each model is optimized to predict a single soil property independently. In contrast, multi-task learning (MTL) leverages shared representations to simultaneously predict multiple soil properties within a single model, improving efficiency and potentially enhancing predictive performance by capturing interdependencies among soil attributes (Ng et al., 2019).Click or tap here to enter text.Click or tap here to enter text.Click or tap here to enter text.. Nyawasha et al. (2024) implemented a multivariate NN model for NIR and MIR spectroscopy, demonstrating improved prediction efficiency and reduced computational requirements compared to single-task models.

2.3.3 How CNN Work in Soil Spectroscopy

CNN consist of multiple layers that transform input spectral data into meaningful features through hierarchical learning, making them well-suited for analysing complex relationships in soil spectra, the basic structure was obtained from (Padarian *et al.*, 2019):

- **Convolutional Layers**: These layers apply filters to the input data, for instance soil spectra, to extract local patterns using peaks and valleys in reflectance curves. Filters slide across the data, identifying features associated with key soil properties.
- **Activation Functions**: Functions like Rectified Linear Unit (ReLU) introduce non-linearity, enabling the model to capture complex relationships in spectral data.
- **Pooling Layers**: Pooling reduces the dimensionality of the data while retaining key information. Max pooling, for instance, focuses on the most prominent spectral features, improving computational efficiency.

- **Fully Connected Layers**: High-level features extracted from convolutional layers are combined to predict soil properties such as SOC or clay content.
- **Output Layer**: The final layer produces predictions, which, include: regression values (e.g., organic carbon percentage) or classifications (e.g., soil texture categories).

The training process involves forward propagation to generate predictions, followed by backpropagation to adjust model parameters based on prediction errors calculated by a loss function, using Mean Squared Error (MSE) (Ng *et al.*, 2019; Padarian *et al.*, 2019). Optimization algorithms like stochastic gradient descent (SGD) iteratively refine the model (Ng *et al.*, 2019; Padarian *et al.*, 2019).

2.3.4 Why CNN Excel in Soil Spectroscopy

CNN automatically extract relevant features from raw spectral data, eliminating the need for extensive pre-processing, a common requirement in traditional methods like PLSR (Ng *et al.*, 2019; Padarian *et al.*, 2019). They are also highly scalable, handling large datasets effectively, and are robust against noise, as demonstrated by their superior performance in multi-task and transfer learning scenarios (Padarian *et al.*, 2019). Their ability to extract local and abstract features from spectral data, combined with their adaptability to complex datasets, makes CNN a transformative tool for soil spectroscopy and digital soil mapping applications (Padarian *et al.*, 2019).

2.4 Statistical Metrics for Model Validation

In soil spectral inference, validating models is essential to ensure that predictions are accurate, robust, and applicable across different contexts. Statistical metrics are fundamental tools for assessing prediction quality and refining calibration techniques. This section explores commonly used metrics, highlighting key studies in soil spectroscopy that have employed them. The Root Mean Square Error (RMSE) measures the average magnitude of prediction errors by calculating the square root of the mean squared differences between observed and predicted values (Wadoux *et al.*, 2021). It directly assesses prediction accuracy in absolute terms, with lower values indicating higher accuracy. A Study by Dangal et al. (2019) have utilized RMSE to validate models predicting soil properties pH, CEC, and SOC.

Assessing systematic discrepancies, the Mean Error (Bias) evaluates the tendency of a model to overestimate or underestimate values (Wadoux *et al.*, 2021). A bias close to zero suggests minimal systematic error, offering insights into calibration adequacy and helping identify directional trends in predictions (Wadoux *et al.*, 2021). The Coefficient of Determination (R^2)

quantifies the proportion of variance in observed data explained by the model, with values approaching 1.0 indicating strong performance (Wadoux *et al.*, 2021). Often used alongside RMSE, R^2 provides a comprehensive view of relative and absolute prediction accuracy (Wadoux *et al.*, 2021). To offer a normalized measure of predictive capacity, the Ratio of Performance to Deviation (RPD) is calculated as the ratio of the standard deviation of observed values to the RMSE (Wadoux *et al.*, 2021). RPD values exceeding 2.0 indicate robust models suitable for practical application, while values below 1.4 suggest weak predictive power (Wadoux *et al.*, 2021). Addressing limitations in datasets with skewed distributions, the Ratio of Performance to Interquartile Range (RPIQ) normalizes RMSE against the interquartile range of observed data (Wadoux *et al.*, 2021). This metric is particularly valuable in soil property prediction where data heterogeneity is common. Expressed as a percentage, the Coefficient of Variation (CV) measures the variability of prediction errors relative to the mean of observed values (Metwally *et al.*, 2019). CV enables cross-dataset comparisons by standardising error magnitudes and is a versatile tool in model evaluation (Metwally *et al.*, 2019). Finally, Lin's Concordance Correlation Coefficient (LCCC) offers a holistic measure of agreement between observed and predicted values by evaluating both precision (correlation) and accuracy (closeness to the 45° line) (Wadoux *et al.*, 2021). Researchers can comprehensively evaluate soil spectral models by integrating multiple metrics, capturing various performance dimensions. This multidimensional approach facilitates the identification of strengths and limitations, the refinement of calibration methods, and the enhancement of predictive reliability, which are essential for advancing soil spectroscopy in soil property assessment and management.

2.5 Soil Spectroscopy in a South African Context

In South Africa, soil spectroscopy using NIR, MIR, and Vis-NIR techniques have been extensively employed to predict a range of soil properties. These spectral methods have significantly contributed to environmental monitoring and agricultural management across the country's diverse landscapes. Nocita et al. (2011) utilized Vis-NIR spectroscopy combined with Partial Least Squares Regression (PLSR) to estimate soil SOC (SOC) in the Albany Thicket Biome. Their study demonstrated accurate SOC predictions for surface and topsoil samples. It emphasized the potential to integrate satellite-borne hyperspectral data with field spectroscopy to scale up SOC estimations in degraded ecosystems. Similarly, Bayer et al. (2012) explored Vis-NIR spectroscopy for predicting SOC, clay content, and iron oxides in soils from the Thicket Biome. By comparing feature-based regression with PLSR, they highlighted the effectiveness of multivariate approaches for heterogeneous soil datasets, reinforcing the reliability of Vis-NIR in soil monitoring and management efforts (Bayer *et al.*, 2012). Koch et al. (2017) applied Vis-NIR spectroscopy to analyse mine tailings across four mining sites. Their evaluation of the predictive

capacity for multiple soil elements yielded RPD values ranging from fair to excellent (1.4 to > 2.0), underscoring Vis-NIR's utility in environmental applications like monitoring soil contamination and assessing mineralogical properties in mining regions. In a related study, Ambushe et al. (2015) investigated Laser-Induced Breakdown Spectroscopy (LIBS) alongside NIR and MIR spectroscopy to detect heavy metals, focusing on chromium quantification in polluted areas. Their work highlighted the complementarity of these techniques in providing accurate elemental and mineralogical data for environmental remediation. Shepherd and Walsh (2002) laid the groundwork for soil spectroscopy in South Africa by establishing a regional soil spectral library using Vis-NIR spectroscopy. Achieving robust predictions for SOC (R^2 = 0.80), clay content (R^2 = 0.80), and CEC (R^2 = 0.88), they provided a valuable resource for soil property estimation across Southern Africa, promoting the use of regional spectral libraries for agricultural and environmental applications.

More recent studies have further emphasized the value of Vis-NIR, NIR and MIR spectroscopy. Kock et al. (2024) focused on MIR spectroscopy models in the Western Highveld, demonstrating that locally calibrated models significantly outperformed global ones for soil properties like pH, calcium, and magnesium, with R^2 values exceeding 0.76. Seboko et al. (2023) employed Fourier-transform MIR spectroscopy to evaluate SOC in granitic soils, showing that database disaggregation by soil texture and depth improved prediction accuracy, achieving RPD values above 2.0. Additionally, Mathadeen et al. (2013) utilized Diffuse Reflectance Infrared Fourier Transform (DRIFT) spectroscopy covering NIR and MIR ranges to predict soil properties in sugarcane fields. They achieved accurate calibrations for clay, SOC, and exchangeable cations, with R^2 values exceeding 0.85, underscoring diffuse reflectance Fourier transformed (DRIFT) spectroscopy's robustness in agricultural systems.

Collectively, these studies illustrate significant advancements in soil spectroscopy within South Africa. They emphasize the reliability and efficiency of Vis-NIR, NIR and MIR techniques when integrated with advanced statistical tools and region-specific spectral libraries. The growing adoption of these methods enhances their utility for sustainable soil management, precision agriculture, and environmental monitoring, paving the way for further innovations in the field.

2.6 Combining Spectral Regions for Soil Spectroscopy

Integrating different spectral regions, such as NIR and MIR, has emerged as a promising approach to improve the accuracy and scope of soil property predictions. By harnessing the complementary strengths of these spectral ranges, researchers can capture a broader array of chemical and physical soil characteristics. A comprehensive review by Soriano-Disla et al. (2014) evaluated the performance of visible, NIR, and MIR spectroscopy in predicting soil properties.

They found that while MIR spectroscopy generally provided superior results for chemical properties like CEC and total N, NIR offered advantages in portability and suitability for field conditions. They suggested that combining these spectral ranges enhanced predictive accuracy across diverse soil properties, facilitating applications in both laboratory and field settings. The benefits of integrating multiple spectral ranges are further illustrated by O'Rourke et al. (2016). They demonstrated that using Vis-NIR, MIR, and portable X-ray fluorescence (pXRF) spectroscopy in tandem improved soil geochemistry characterization. By combining outputs from Vis-NIR and MIR spectra, the number of accurately predicted properties increased from 15 to 25, including trace elements and pH. This underscores the value of a comprehensive spectral approach for soil analysis. Exploring the predictive capabilities of fused spectra, Xu et al. (2020) compared NIR, MIR, and their combined data for soil classification. Their research revealed that integrating NIR and MIR spectra using outer product analysis (OPA) yielded the highest classification accuracy at 68.4%. This fusion helped overcome individual limitations like overfitting and signal variability, highlighting the potential of combined spectral data. Advanced machine learning models have also been applied to combined spectral regions. Ng et al. (2019) investigated the fusion of Vis-NIR and MIR data using CNN. Their study showed that integrating spectral data significantly enhanced prediction accuracy for properties SOC, clay content, and CEC, achieving R^2 values exceeding 0.95. Data fusion techniques like spectral concatenation and two-channel input methods produced robust models suitable for large-scale applications (Ng *et al.*, 2019). Shao and He (2011) emphasized the complementarity of the NIR and MIR regions, noting that both contain substantial information on soil nutrients like N, P, and K. While combining the two spectral regions can provide a more comprehensive spectral profile, their findings suggest that the accuracy of predictions may not always significantly improve compared to using individual regions alone. However, the integration of both regions can still offer advantages, such as improved robustness and the ability to measure a wider range of soil properties. Collectively, these studies suggest that combining spectral regions, along with advanced analytical models, can enhance the precision and versatility of soil analysis. Leveraging complementary spectral information and integrating advanced analytical models allows researchers to overcome the limitations of single spectral regions, paving the way for more precise and versatile soil analysis.

2.7 Deep Learning and CNN in Soil Spectroscopy for Determining pH, Phosphorus, and Exchangeable Cations

The application of deep learning, particularly CNN, has revolutionized soil spectroscopy by enhancing the prediction of soil properties pH, P, and exchangeable cations. Ng et al. (2019) demonstrated that CNN could significantly improve predictions over traditional methods like PLSR and Cubist models. Using the Kellogg Soil Survey Laboratory (KSSL) dataset, their CNN models

achieved R^2 values as high as 0.98 for pH and CEC, underscoring CNN' ability to outperform conventional models through automatic feature learning and robust performance on large datasets (Ng *et al.*, 2019). Similarly, Padarian et al. (2019) employed CNN for soil property prediction using the LUCAS soil database, which includes extensive physicochemical measurements like pH, P, and CEC. Their multi-task CNN approach significantly reduced prediction errors compared to PLSR and regression tree models, achieving RPD values greater than 2.0 for pH and CEC. This study highlighted CNN' efficiency in simultaneous multi-property predictions, enhancing their utility for large-scale applications (Padarian *et al.*, 2019). Hosseini et al. (2023) integrated CNN with satellite-derived remote sensing data for soil property mapping, achieving high prediction accuracy for pH and phosphorus. Their hybrid CNN- Recurrent Neural Network (RNN) model reduced RMSE to as low as 0.0206 for pH and 0.1078 mg kg^{-1} for P, demonstrating the deep learning models' potential to harness multi-source data for precise soil property estimation (Hosseini *et al.*, 2023). This approach underscores CNN' versatility in combining spectral data with other environmental datasets. Additionally, Rathore and Nath Singh (2022) applied CNN for nutrient classification, including phosphorus and pH, using soil imagery. Their model outperformed traditional machine learning algorithms like SVM, achieving higher classification accuracies across varying soil types. This highlights CNN adaptability for diverse input data formats and soil properties (Rathore and Singh, 2022). Overall, the adoption of CNN in soil spectroscopy has significantly enhanced the accuracy and efficiency of predicting critical soil properties. Their ability to model complex relationships within spectral data positions them as indispensable tools for advancing precision agriculture and sustainable soil management.

2.8 Spiking in Soil Spectroscopy: Applications and Advancements

Spiking is a technique in soil spectroscopy that enhances the predictive performance of calibration models when applied to new or diverse datasets. By incorporating a small number of samples from the target dataset into an existing spectral library, spiking improves model adaptability and reduces bias, especially in heterogeneous soil environments. Numerous studies across various regions and soil properties have demonstrated its value in achieving accurate predictions. For instance, the transferability of Vis-NIR spectroscopy models for estimating SOC between regions in Central China was significantly improved by spiking with approximately 33 - 48% of the target samples, leading to better SOC predictions across different sites (Hong *et al.*, 2018). Similarly, enhancing Vis-NIR spectroscopic models for predicting SOC and total N (TN) across soil layers in a mixed forest ecosystem was achieved by reducing RMSE and absolute bias through spiking, with even better performance when using extra-weighting techniques (Jiang *et al.*, 2017). In large-scale applications, adapting regional-scale NIR models for SOC prediction in Zhejiang Province, China, was accomplished by spiking with small subsets of local samples, which significantly

reduced prediction errors, particularly when utilizing smaller-scale spectral libraries. This demonstrates the scalability of spiking for localized predictions (Li *et al.*, 2020). Additionally, improving prediction accuracy for key soil properties such as total N, total C, and soil moisture content across different geographical scales in Europe was achieved by spiking with local samples in mobile Vis-NIR spectroscopy, attaining R^2 values as high as 0.98 for total N and total C and highlighting the importance of localized calibration in mobile soil analysis (Nawar and Mouazen, 2017). Comparative studies have shown that spiking can be more effective than local calibration strategies. Using large-scale soil spectral libraries like the LUCAS database, spiking with as few as 15 local samples drastically reduced prediction bias and improved RPD values for SOC estimates in Germany, demonstrating the cost-effectiveness of this method for regional applications (Seidel *et al.*, 2019). Collectively, these studies underscore the significance of spiking in adapting soil spectral models to diverse and heterogeneous datasets. By integrating local samples into regional or global spectral libraries, spiking enhances model accuracy, reduces bias, and facilitates the practical application of soil spectroscopy across varied landscapes. This approach is particularly effective in addressing the challenges posed by variability in soil properties and environmental conditions.

2.9 Overview of Soil Spectral Libraries and their Role in Prediction Modelling

Soil spectral libraries are essential resources in soil spectroscopy, providing extensive datasets that enable the development and validation of predictive models for soil properties. These libraries shown in Table 2-1 integrate spectral data with corresponding laboratory-measured soil properties, serving as reference frameworks for applications in environmental monitoring, agricultural management, and soil health assessment. Soil spectral libraries are essential resources in soil spectroscopy, providing extensive datasets that enable the development and validation of predictive models for soil properties. Pioneering this concept, a dataset of over 1,000 topsoil samples from eastern and southern Africa was used to predict properties such as soil pH, SOC, and CEC (Shepherd and Walsh, 2002). This work demonstrated that large and diverse spectral libraries enhance prediction accuracy, even in heterogeneous landscapes. At a global level, a Vis-NIR spectral library incorporating diverse soil samples from various regions was developed to enable accurate prediction of properties like soil SOC, clay content, and pH (Viscarra Rossel *et al.*, 2016). This harmonized dataset addresses challenges related to data heterogeneity and limited local samples, underscoring the importance of global spectral libraries for worldwide soil studies. The LUCAS topsoil database has been instrumental in modelling soil properties across Europe. Utilizing this library, studies have achieved high prediction accuracies for SOC, CEC, and clay content through advanced machine learning methods like CNN (Padarian *et al.*, 2019). This highlights the potential of large-scale datasets in enhancing predictive

modelling. Introduced as a centralized, open-access resource, the Open Soil Spectral Library (OSSL) integrates multiple spectral datasets to facilitate reproducible soil calibration models and extend soil spectroscopy's reach to underserved regions (Safanelli *et al.*, 2023). Its development underscores the importance of community engagement in building comprehensive spectral libraries. Local spectral libraries have also proven invaluable for site-specific applications. By stratifying a local Vis-NIR spectral library by soil class and land use, prediction accuracy for soil SOC in subtropical soils was improved, highlighting the effectiveness of tailoring libraries to specific environmental and pedological conditions (Moura-Bueno *et al.*, 2019). Moreover, spectral libraries guide model calibration strategies. Quantifying the uncertainty in predictions derived from local and regional spectral libraries emphasizes the importance of leveraging large-scale datasets to enhance the reliability of soil property estimates(Breure *et al.*, s.a.). These studies collectively demonstrate that soil spectral libraries are indispensable for advancing soil property prediction. Through global harmonization, regional adaptations, or localized calibrations, these libraries enable precise, scalable, and cost-effective soil analyses, playing a pivotal role in soil science research and practical applications.

Table 2-1: Overview of prominent soil spectral libraries imported into the Open Soil Spectral Library (OSSL).

Library Name	**n**	**Spectral Wavelength (cm^{-1})**	**Location**	**Reference**
ICRAF-ISRIC Soil Spectral Library	4073	350 - 2500 nm (Vis-NIR), 600-4000	Global (ICRAF and	(ICRAF) et al., 2021; Safanelli et al., 2023)
KSSL (USDA-NRCS) Library	19807	350 - 2500 nm (Vis-NIR), 600-4000	USA	(Safanelli *et al.*, 2023)
LUCAS Soil Spectral Library	40764	350 - 2500 nm (Vis-NIR), 600-4000	Europe	(Centre *et al.*, 2020; Safanelli *et al.*, 2023)
AfSIS1 (Africa Soil Information Service 1)	1904	350 - 2500 nm (Vis-NIR)	Africa	(Safanelli *et al.*, 2023; Vågen *et al.*, 2020)
AfSIS2 (Africa Soil Information Service 2)	151	600 - 4000 cm^{-1} (MIR)	Africa	(Hengl *et al.*, 2021; Safanelli *et al.*, 2023)
CAF (Central African Forest Soils)	1578	600 - 4000 cm^{-1} (MIR)	Congo Basin	(Safanelli *et al.*, 2023; Summerauer *et al.*,
Garrett Library (New Zealand Forest Soils)	184	600 - 4000 cm^{-1} (MIR)	New Zealand	(Garrett *et al.*, 2022; Safanelli *et al.*, 2023)
Schiedung Library (Boreal Soils - University of Zurich)	259	600 - 4000 cm^{-1} (MIR)	Boreal Region	(Safanelli *et al.*, 2023; Schiedung *et al.*,
Serbian Soil Spectral Library	135	600 - 4000 cm^{-1} (MIR)	Serbia	(Jović *et al.*, 2019; Safanelli *et al.*, 2023)
Neospectra Soil Spectral Library	2106	1350 - 2550 nm (NIR)	USA, Ghana, Kenya,	(Sanderman *et al.*, 2023)

n = number of samples

CHAPTER 3 USING SOIL SPECTRAL INFERENCE AND COMBINED SPECTRA TO PREDICT SOIL PROPERTIES IN SOUTH AFRICA

3.1 Abstract

Accurate prediction of soil properties is vital for optimizing land management, agricultural productivity, and environmental sustainability. In South Africa, the unique soil characteristics and environmental diversity challenge the application of generalized predictive models, creating a knowledge gap in soil spectroscopy. This study investigated the integration of visible-near-infrared (Vis-NIR) and mid-infrared (MIR) spectral data to improve soil property predictions. Using a dataset of 772 soil samples from the Western Highveld region of South Africa, spectral data were pre-processed through three strategies: raw spectral analysis, blanket pre-processing, and tailored pre-processing. Prediction models were developed using the Cubist machine learning algorithm, focusing on the soil properties pH (KCl), P (Bray-1), and exchangeable cations (Ca, Mg, K, Na). The combined Vis-NIR and MIR spectral data demonstrated significant improvements in predictive accuracy for properties like pH, Ca, and Mg, achieving reductions in root mean square error (RMSE) of up to 41.67% compared to single-spectrum models. However, predictions for Na (RMSE = 9,16 mg kg LCCC = 0.28) and K (RMSE = 47,55 mg kg LCCC = 0,73) remained suboptimal, indicating challenges with spectral interference and noise. The results highlight the complementary strengths of Vis-NIR and MIR spectroscopy, with raw combined spectra providing robust predictive capabilities for most soil properties. This study contributes a framework for integrating spectral data, emphasizing the value of targeted pre-processing for challenging properties. These findings advance soil spectroscopy, offering practical solutions for precision agriculture in South Africa. By enhancing soil property predictions, this research supports sustainable farming practices, efficient resource use, and the development of a comprehensive South African Soil Spectral Library.

3.2 Introduction

Determining accurate information on soil properties is essential for effective land management, agricultural productivity, and environmental sustainability. It provides essential information regarding soil fertility, nutrient content, and other properties that directly influence crop productivity and sustainability. Rachman (2020) emphasizes that understanding soil conditions, including fertility levels and nutrient status, is vital for implementing precision agricultural practices and making informed fertilizer recommendations. The unique soil characteristics and environmental conditions present in South Africa pose significant challenges for applying generalized soil

property prediction models developed in other regions, which can lead to discrepancies in predictive accuracy when compared to traditional laboratory analyses (Janik *et al.*, 2009). Despite advancements in soil spectral inference using visible-near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy (Bai *et al.*, 2022; Naimi *et al.*, 2022), there remains a notable gap in the predictive capabilities of these models (López *et al.*, 2019; Sharififar *et al.*, 2019). The development of a South African Soil Spectral Library is still in its infancy due to limited soil spectral data and databases that are publicly available, with local studies being significantly fewer compared to those from other countries (Kock et al., 2024).

The inherent complexity of soil, characterized by spatial and temporal variability is influenced by factors which include but not limited to soil composition, soil moisture content, and environmental conditions, necessitates the development of tailored calibration methods for local contexts (Huang *et al.*, 2016; Lobsey *et al.*, 2017). Traditional single-spectral models often fail to capture the intricate relationships between various soil properties, leading to unrealistic predictions, commonly referred to as “pedological chimera” (Stenberg et al., 2010). For instance, while NIR spectroscopy is proficient at detecting organic matter and moisture, it often struggles with mineral-based properties, whereas MIR spectroscopy excels in capturing molecular vibrations related to soil mineralogy and nutrient status (Rossel *et al.*, 2006). However, the individual application of these spectral regions has limitations, and despite the advancements in soil spectral inference, a consensus remains that soil spectral models have not yet achieved the predictive accuracy of traditional laboratory analyses (Tsakiridis *et al.*, 2019). The combined utilization of MIR and NIR remains relatively unexplored, indicating a critical area for future research.

Recent studies have investigated the potential benefits of fusing or conjugating Vis-NIR and MIR spectral data to enhance predictive accuracy. For example, Li et al. (2021) demonstrated that the integration of these spectral regions could improve the prediction of soil nutrients, although the expected enhancements in model performance were not consistently realized (Dorantes *et al*., 2022). This highlights a critical knowledge gap regarding the optimal integration of NIR and MIR spectral data for soil property prediction. The potential advantages of spectral data fusion may lie in leveraging the strengths of each spectral region, which could improve the predictive capabilities of models for key soil properties P, K, and Na, which are notoriously difficult to predict using MIR alone (Chang *et al.*, 2001; Chao *et al.*, 2021).

The choice of pre-processing techniques applied to spectral datasets and specific spectral regions plays a crucial role in optimizing model performance. Different pre-processing methods can mitigate noise, enhance spectral features, and optimize the signal-to-noise ratio, which influences the robustness and reliability of soil property predictions (Maia, 2023). For example,

tailored pre-processing techniques that account for the unique characteristics of each spectral region may significantly improve model accuracy compared to blanket approaches (Mouazen et al., 2010; Okparanma and Mouazen, 2013). However, the effectiveness of these techniques can vary depending on the specific soil properties being predicted and the spectral characteristics of the samples.

This study aimed to evaluate the integration of NIR and MIR spectral data for predicting essential soil properties, including pH (KCl), P (Bray-1), and ammonium acetate extractable cations, (typically measured for precision agriculture in South Africa). The primary objective of this chapter is to assess the benefits of combining NIR and MIR spectral data and determine whether their integration improves prediction accuracy compared to using either spectral region alone. This work will also contribute to the ongoing research toward a working South African soil spectral library that functions comparably to established libraries for example the Brazilian soil spectral library (BSSL) (Demattê *et al.*, 2019). We hypothesized that the combined use of these spectral regions will enhance the predictive capability of models built using the Cubist machine learning algorithm, known for its robustness and interpretability (Mammadov *et al.*, 2022; Vasques *et al.*, 2008). To investigate this hypothesis, we employed a comprehensive approach that encompassed three distinct pre-processing strategies: (1) utilizing raw spectral data, (2) applying tailored pre-processing techniques to NIR and MIR separately, and (3) employing a blanket pre-processing approach across the entire combined spectral dataset.

The findings of this research have the potential to inform sustainable agricultural practices and precision farming initiatives. Rapid and reliable soil property assessments can enable farmers to make informed decisions about fertilizer application, irrigation scheduling, and crop selection, leading to optimized resource utilization and minimized environmental impact (Shepherd and Walsh, 2002; Xie et al., 2015). Furthermore, the development of robust and interpretable spectral models can facilitate the widespread adoption of spectroscopy-based soil analysis, empowering emerging farmers across the agricultural sector.

3.3 Materials and Methods

3.3.1 Study Area

The soil samples analysed in this study originated from the Western Highveld region of South Africa's North-West Province, a major grain-producing area characterized by diverse soil types and environmental conditions. As described in Kock et al. (2024), this region exhibits a wide range of temperatures and precipitation levels, classifying it as an arid to semi-arid zone (Buontempo *et al.*, 2020). The geological diversity of the area, with varying formations in the western and eastern

parts, contributes to the variability in soil types, including redoximorphic soils, clays, shallow soils, and alluvial soils. The vegetation is classified as Western Highveld Grassland, this region supports the cultivation of various crops, with maize and sunflower being the primary crops grown in the specific study areaClick or tap here to enter text..

3.3.2 Soil Sample Data Acquisition

A soil sample dataset containing 772 topsoil samples originating from the Western Highveld region of South Africa located within the North West Province of the country was used (Kock et al., 2024). This dataset contains several soil properties including pH (KCl), P (Bray-1) and Exchangeable cations (Ca, Mg, K, Na) analysed using the 1M ammonium acetate extraction method. These samples were donated to the study by regional commercial service providers NWK Ltd. and GWK cooperation (Kock et al., 2024). Currently, no other soil property data were obtained for these samples. The soil properties selected for this study are critical for agricultural management in South Africa, as they are routinely used in fertilizer recommendations and overall soil health determination. Accurate predictive models for these properties are essential to support effective nutrient management and optimize crop production, addressing both regional soil variability and the specific needs of South African farming. The dataset used also contained MIR spectral data with a spectral range of 600 cm^{-1} - 4000 cm^{-1} which were previously collected by Kock et al. (2024). The MIR spectra were collected using a Bruker Alpha II spectrometer with DRIFT module (Bruker OPTIK GmbH, 2019) with a spectral range of 3996 - 398 $\mathbf{cm^{-1}}$ and a resolution of 2 $\mathbf{cm^{-1}}$.

3.3.3 Spectral Data Expansion

The spectral library was extended to include NIR spectral data of the samples mentioned previously. The NIR spectral reflectance was measured using a handheld FT-NIR spectrometer Neospectra™ produced by Si-Ware systems (Si-ware Systems, 2023). The wavelength ranged from 7 400 cm^{-1} - 4 000 cm^{-1} (1 350 - 2 500 nm) with a resolution of 66.6 cm^{-1} (Si-ware Systems, 2023). The soil samples were dried and sieved using a 2 mm sieve and placed in a plastic holder, the scanner was placed directly on top of the soil sample and the scanning procedure was initiated using the android Neospectra™ Collect application for android. A scan time of 15 seconds was used for each sample and only one scan was used. The collected spectral data was exported to a .csv file which was used in R version 4.3.1 (2023-06-16 ucrt) with Rstudio 2023.06.1 Build version 524 (R Core Team, 2023). The NIR spectral data was converted to absorbance by changing reflectance data from the Neospectra™ scanner into absorbance values, which involves a mathematical conversion expressed as log ($1/X^R$), where X^R represents the reflectance

measurements arranged in a matrix with rows representing spectra and columns representing wavelengths (Wadoux *et al.*, 2021). The NIR spectral data was merged to the existing soil spectral library which contains the MIR spectral data using *bind* function in R.

3.3.4 Spectral Pre-processing

Prior to modelling, the raw NIR and MIR spectral data obtained from the soil spectral library, underwent an outlier removal process using the Mahalanobis Distance (Wadoux *et al.*, 2021). Samples with intensity or absorbance values exceeding a threshold of three standard deviations were excluded, ensuring a consistent dataset. This procedure resulted in 772 samples for the subsequent prediction model training and validation. The independent NIR and MIR regions were then conjugated, using the bind function in R version 4.2.0(R Core Team, 2023) to create a continues spectral data range which was used in the following steps. The spectral data in this study were pre-processed using three distinct approaches to evaluate their effectiveness when combined: The first approach deliberately omitted the commonly used spectral pre-processing techniques typically applied in soil spectral inference (McCarty and Reeves, 2006). The second approach applied a blanket pre-processing strategy. Combined NIR and MIR spectral data were processed using the Savitzky-Golay filter across the entire spectral range, with a moving window of 11 points and a polynomial order of 2 (Savitzky and Golay, 1964). Standard Normal Variate (SNV) transformation was then applied both the NIR and MIR spectral data to correct for light scattering (Nyawasha *et al.*, 2024). For the third approach, NIR and MIR spectral data were pre-processed independently. Additionally, the MIR spectra were further smoothed and denoised using the Savitzky-Golay filter, employing a moving window of 11 points and a polynomial order of 2 (Wadoux *et al.*, 2021).

3.3.5 Soil Spectral Inference Model Calibration

The respective pre-processed sets of soil spectral libraries were then split into training and independent validation datasets. Conditioned Latin hypercube sampling (cLHS) was employed to split the soil property dataset based on selected soil properties (pH, P, Ca, Mg, K and Na) (Minasny and Mcbratney, 2006). The algorithm was initialized to partition the data into 75% training and 25% independent validation set. A calibration dataset, encompassing both the soil analysis data and their corresponding combined spectral data, served as the foundation for training Cubist models for each soil property investigated in this study (Kuhn and Johnson, 2013). To ensure optimal performance, a grid search function was employed to systematically evaluate various combinations of Cubist parameters (e.g., number of committees, number of rules) for each soil property model. This meticulous optimization process allowed us to identify the parameter configuration that yielded the best predictive accuracy for each individual model.

Subsequently, the remaining models were calibrated using the same calibration dataset, but with their parameters set to the optimal values determined in the previous step. This approach ensured consistency and maximized the predictive potential of the Cubist models across all soil properties under consideration.

3.3.6 Soil Spectral Inference Model Validation

The performance of the developed spectral prediction models was rigorously evaluated using an independent validation dataset, employing a suite of statistical metrics commonly utilized within the soil spectral research community (McCarty and Reeves, 2006; Nocita et al., 2015). These metrics encompassed the coefficient of determination (R^2)(Equation 3-1), the root mean square error (RMSE)(Equation 3-2), ratio of performance to inter-quartile distance (RPIQ)(Equation 3-3), the Lin's concordance correlation coefficient (LCCC)(Equation 3-4) and mean absolute error (MAE)(Equation 3-5). R^2 quantifies the proportion of variance in the observed data explained by the model (Stenberg et al., 2010). It was calculated as the square of the Pearson correlation coefficient between observed and predicted values. RMSE measures the average magnitude of prediction errors (Wadoux *et al.*, 2021). It was calculated as the square root of the mean squared difference between observed and predicted values. RPIQ assesses the model's predictive ability relative to the natural variation in the data (Wadoux *et al.*, 2021). It was computed as the ratio of the standard deviation of the observed values to the RMSE. Lin's concordance correlation coefficient evaluates both the accuracy and precision of the model (Wadoux *et al.*, 2021). It incorporates measures of both correlation and agreement between observed and predicted values. Scatter plots were also constructed to visually compare the observed versus predicted values for each prediction model. These plots featured a 1:1 line, representing a perfect prediction, allowing for a qualitative assessment of model performance. It is crucial to note that this 1:1 line is not a line of best fit, but rather a reference for ideal prediction accuracy. The improvement of prediction models was quantified by calculating the percentage change in LCCC and RMSE compared to a baseline model, for instance a model utilizing only one spectral region as the predictor input, for each soil property, as outlined in Equation 3-6 and Equation 3-7 respectively.

Equation 3-1

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Obs_i - pred_i)^2}{\sum_{i=1}^{n}(Obs_i - \overline{Obs})^2}$$

where n is the validation sample size and Obs and $pred$ are vectors of observed and predicted values of the soil properties, respectively and which is equal to $1 - SSE/SST$ where SSE is the sum of the squared error and SST is the total sum of squares (Wadoux *et al.*, 2021).

Equation 3-2

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Obs_i - pred_i)^2}$$

Equation 3-3

$$RPIQ = \frac{(Q_3(Obs) - Q_1(Obs))}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Obs_i - pred_i)^2}}$$

where $Q_1(Obs)$ and $Q_3(Obs)$ are the first (25%) and third (75%) quantiles of the observations $Q_3(Obs) - Q_1(Obs)$ is the inter-quartile distance) and the denominator is RMSE (Wadoux *et al.*, 2021).

Equation 3-4

$$LCCC = \frac{2r\sigma_{pred}\sigma_{Obs}}{{\sigma_{Obs}}^2 + {\sigma_{pred}}^2 + (\mu_{Obs} - \mu_{pred})^2}$$

where r is Pearson's correlation coefficient, σ is the standard deviation ($r\sigma_{pred}\sigma_{Obs}$ is the covariance between observed and predicted values) and μ is the mean of the predictions or observed values (Wadoux *et al.*, 2021).

Equation 3-5

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Obs_i - pred_i|$$

Equation 3-6

$$LCCC\ improvement\ \% = \frac{(New\ Value - Baseline\ Value)}{Baseline\ Value}\ x\ 100$$

Equation 3-7

$$RMSE\ improvement\ \% = \frac{(Baseline\ Value - New\ Value)}{Baseline\ Value}\ x\ 100$$

3.4 Results and Discussion

3.4.1 Combined Spectral Dataset

The existing soil spectral library, which includes soil property data for pH, P (Bray-1), and extractable Ca, Mg, K, and Na, was enhanced by integrating the MIR spectral data from 772 soil samples with their corresponding NIR spectral data. This combination resulted in a comprehensive spectral dataset utilized for developing prediction models. Prior to model creation, the dataset was scrutinized for outliers, leading to the removal of eleven spectral outliers likely caused by scanning inconsistencies. The resultant joined spectra of NIR and MIR spectral data is displayed in Figure 3-1 with outliers removed, while Figure 3-2 shows the results from joined spectra that has undergone blanket and tailored pre-processing. Comparing the blanked approach to the tailored approach as seen in Figure 3-2 similarities are observed between the two spectrographs, however upon closer inspection of the MIR region, stronger peaks are observed compared to the blanked pre-processed spectra of the same region. While this appears to be insignificant, this could influence better retention of important spectral information by the calibration models.

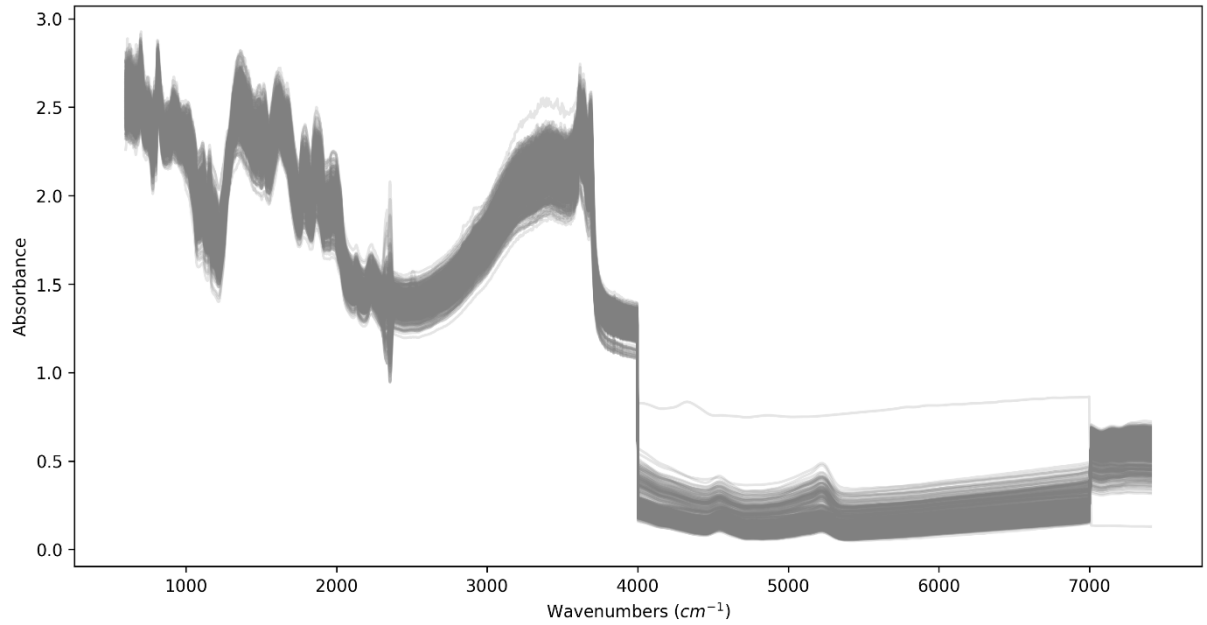

Figure 3-1: Raw spectral data of the soil samples, illustrating the combined NIR (4 000 cm^{-1}- 7 400 cm^{-1}) and MIR (400cm^{-1}- 4 000 cm^{-1}) spectral regions. The variability in spectral signatures reflects the diverse composition and properties of the soil samples.

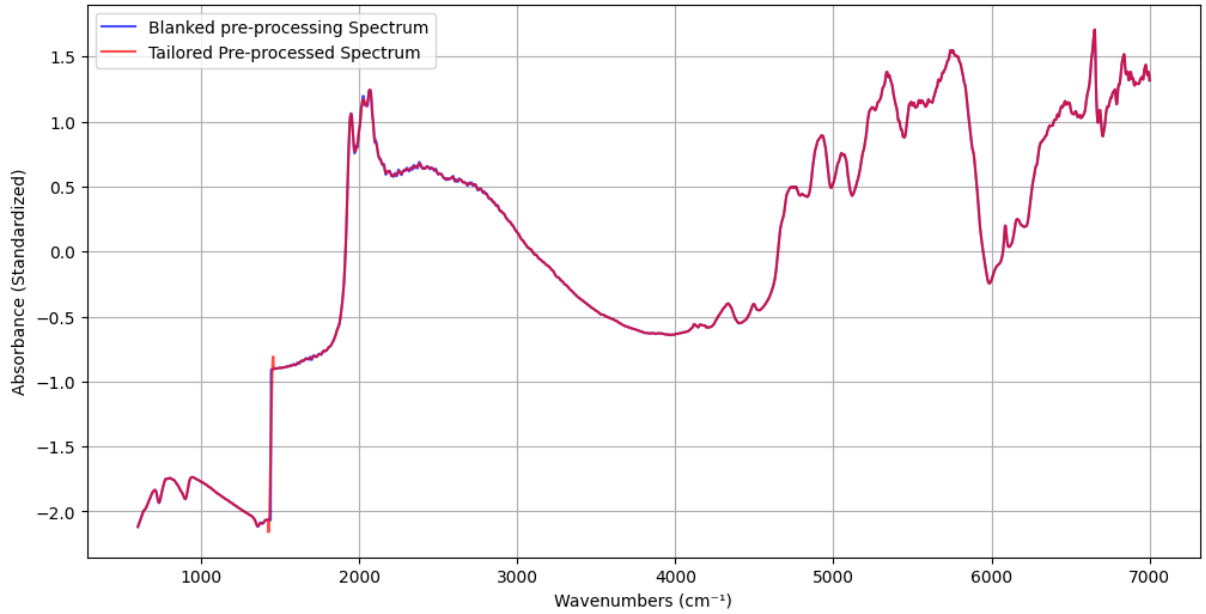

Figure 3-2: Comparison of blanket pre-processing spectrum (blue) and tailored pre-processed spectrum (red) for a selected soil sample across the full spectral range (600 - 7000 cm^{-1}).

Figure 3-3 highlights the effects of tailored pre-processing on the combined NIR and MIR spectral data, focusing on the region subjected to Savitzky-Golay filtering (600 - 4000 cm^{-1}). The blanket pre-processing spectrum (blue line) exhibits noticeable noise and baseline variations, which can obscure critical spectral features and compromise the accuracy of predictive models. By contrast, the tailored pre-processed spectrum (red line) shows significantly reduced noise and improved baseline correction. The smoothing effect preserves essential peaks and valleys, particularly in regions prone to high-frequency noise, resulting in a refined and more interpretable spectral signal. These enhancements underline the importance of customized pre-processing in improving data quality for robust soil property predictions.

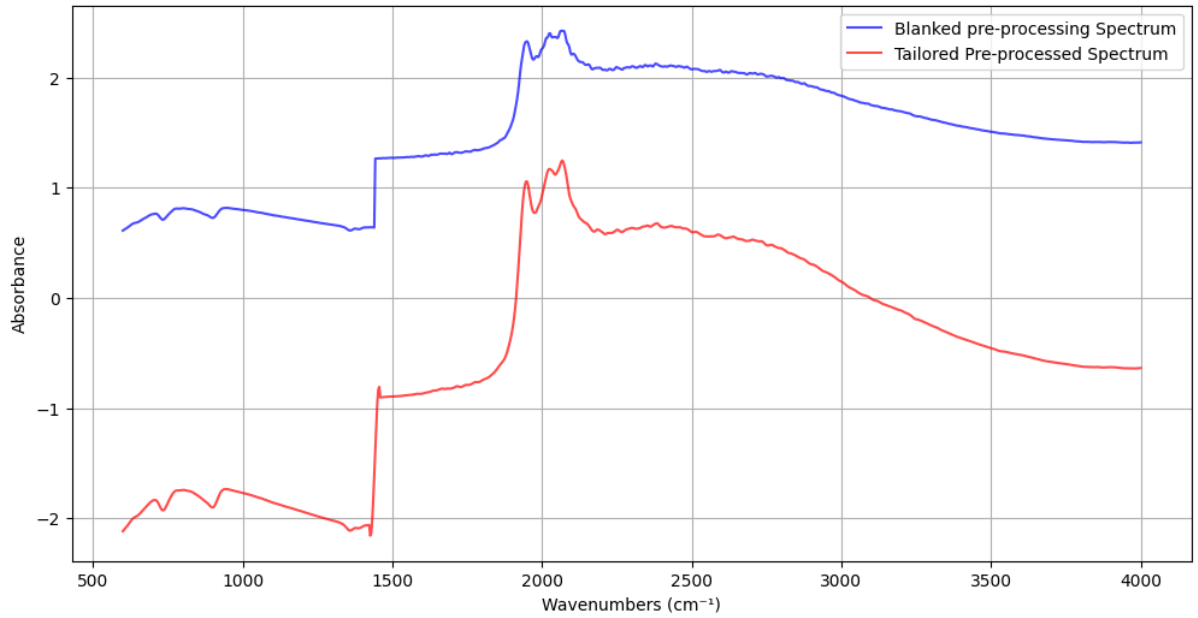

Figure 3-3: Comparison of blanket pre-processing spectrum (blue) and tailored pre-processed spectrum (red) for a selected soil sample, zoomed into the Savitzky-Golay filtered region (600 - 4000 cm^{-1}). The tailored spectrum demonstrates effective noise reduction and smoothing, preserving key spectral features while aligning with the baseline correction of the blanket spectrum.

3.4.2 Model Performance Using NIR, MIR, and Combined Spectral Data

The predictive performance of models using NIR spectroscopy alone was limited, as indicated by relatively lower LCCC values across most soil properties (Table 3-1). Ca exhibited the highest LCCC of 0.84 in the NIR-only model, likely due to NIR's partial sensitivity to Ca-associated structures like $CaCO_3$ (Adeline *et al.*, 2017). However, this performance remained inferior to that achieved with MIR spectroscopy, which provides broader spectral coverage and captures fundamental molecular vibrations relevant to mineral-based soil attributes (Kock et al., 2024; Viscarra Rossel et al., 2016).

Integrating both NIR and MIR raw spectral data significantly improved model accuracy for most soil properties. For example, the pH (KCl) model achieved an LCCC of 0.94 and an RMSE of 0.28, representing a 23.68% increase in LCCC and a 41.67% reduction in RMSE compared to the NIR-only model. Similarly, Ca reached an LCCC of 0.93 and an RMSE of 87.85, highlighting the advantages of combining spectral regions. Mg also showed notable improvements, with an LCCC of 0.89 and an RMSE of 31.06, indicating enhancements of 15.58% in LCCC and 26.55% in RMSE. These results demonstrate that integrating NIR and MIR data enhances predictive capabilities for critical soil fertility parameters by leveraging NIR's sensitivity to organic matter and moisture, alongside MIR's strength in capturing mineral-related information (Padarian *et al.*,

2020). However, the combined raw spectral approach did not improve predictions for all properties. Na remained challenging to predict accurately, with a low LCCC of 0.28 and an RMSE increase of 32.35%, suggesting that unprocessed spectral data may contain noise or overlapping features that hinder accurate modelling for Na. Applying blanket pre-processing techniques demonstrated a moderate level of enhancement for certain soil properties but negatively impacted others. While pH (KCl) predictions remained strong with an LCCC of 0.92 and an RMSE reduction of 37.50%, properties like K and P experienced decreases in LCCC of 18.06% and 5.80%, respectively. These results highlight that while blanket pre-processing effectively mitigates baseline shifts and noise, it is insufficient for properties with complex or overlapping spectral signals (Adeline *et al.*, 2017). Tailored pre-processing showed limited improvements over raw spectral models. For instance, pH (KCl) and Ca maintained high LCCC values of 0.93, similar to the raw spectral results, while Mg and P showed minor improvements with LCCC increases of 16.88% and 6.94%, respectively. However, these gains were not substantial enough to demonstrate a significant advantage over using raw combined spectra. For K and Na, tailored pre-processing offered negligible improvements or even declines in predictive accuracy.

Overall, the results support the initial hypothesis that integrating NIR and MIR spectral data enhances predictive performance, as the combined approach consistently improved accuracy for key soil properties. The integration of NIR and MIR spectral data in its raw form consistently provided significant reductions in RMSE and improvements in LCCC for key soil properties pH (KCl), Ca, and Mg. While raw spectral data provided strong predictive performance, the integration of MIR and NIR spectra proved beneficial for improving accuracy across most soil properties. While tailored pre-processing provided slight enhancements for select properties like P, its effectiveness varied, highlighting the need for further refinement to optimize predictions for all soil attributes. These findings underscore the importance of leveraging the complementary strengths of NIR and MIR spectroscopy to enhance soil property predictions. Future research should focus on refining tailored pre-processing approaches to target properties like Na and K, which continue to show suboptimal performance across all methods. This could involve isolating and enhancing specific spectral bands associated with these properties, particularly when dealing with spectral complexity (Zhang *et al.*, 2020).

Table 3-1: Soil spectral inference models statistical performance calculated from an independent validation dataset.

Target	RMSE (mg kg^{-1})	R^2	MAE	RPIQ	LCCC	Bias	CV
NIR Only							
pH (KCl)	0.48	0.61	0.40	2.19	0.76	0.08	11.35
Ca	119.65	0.70	91.42	2.24	0.84	29.91	43.08
K	45.48	0.54	35.56	2.06	0.72	9.14	35.45
Mg	42.27	0.61	29.64	1.51	0.77	6.45	45.64
P (Bray -1)	14.70	0.50	9.65	1.23	0.69	-0.52	65.33
Na	6.92	-0.17	3.98	1.01	0.44	1.30	52.99
MIR Only (Kock et al.. 2024)							
pH (KCl)	0.29	0.86	-	3.62	-	- 0.03	5.61
Ca	90.32	0.84	-	3.14	-	- 4.5	20.2
K	55.23	0.37	-	1.74	-	- 5.06	39.44
Mg	32.36	0.76	-	2.13	-	- 2.86	29.96
P (Bray -1)	15.39	0.52	-	1.23	-	- 2.28	65.09
Na	6.93	0.06	-	1.16	-	- 0.8	67.25
NIR+MIR Raw Spectra							
pH (KCl)	0.28	0.89	0.21	4.37	0.94	0.02	14.64
Ca	87.85	0.87	63.60	3.63	0.93	2.11	50.62
K	47.55	0.57	33.51	1.91	0.73	-1.76	37.65
Mg	31.06	0.80	20.77	2.16	0.89	-5.26	53.34
P (Bray -1)	15.91	0.55	10.05	1.07	0.74	2.16	71.88
Na	9.16	0.13	4.39	0.76	0.28	-1.30	39.40
NIR+MIR Blanket Pre-processing							
pH (KCl)	0.30	0.85	0.23	3.50	0.92	-0.01	13.29
Ca	95.83	0.86	64.34	3.65	0.92	-8.49	48.65
K	52.36	0.53	39.08	2.35	0.69	-3.43	36.08
Mg	33.97	0.77	23.43	2.03	0.87	-4.89	52.17
P (Bray -1)	15.06	0.55	9.69	1.26	0.71	-0.85	65.37
Na	6.81	-0.55	4.39	1.17	0.45	0.97	61.02
NIR+MIR Tailored Pre-processing							
pH (KCl)	0.29	0.87	0.22	3.81	0.93	0.00	14.21
Ca	89.46	0.87	63.50	3.21	0.93	-6.54	52.44
K	44.66	0.55	31.93	2.18	0.72	2.81	36.19
Mg	30.09	0.82	18.14	1.96	0.90	-0.09	55.40
P (Bray -1)	12.60	0.64	9.49	1.59	0.77	0.90	62.39
Na	8.50	-0.03	4.37	0.82	0.35	0.10	54.03

RMSE = Root Mean Squared Error, R^2 = Coefficient of Determination, MAE = Mean Absolute Error, RPIQ = Ratio of Performance to Interquartile Range, LCCC = Lin's Concordance Correlation Coefficient, CV = Coefficient of Variation

3.4.3 Advantages of Tailored Pre-Processing in Enhancing Soil Property Prediction

The diverse pre-processing techniques applied, ranging from raw spectral data to tailored adjustments, revealed varying impacts on model performance across different soil properties. While tailored pre-processing is generally expected to enhance predictive accuracy by refining spectral signals, the observed improvements over raw spectral data were relatively minor for selected properties. This limited enhancement can be explained by several factors, as also reflected in the scatter plots of predicted versus observed values (Figure 3-4, Figure 3-5, and Figure 3-6). One key factor is that the raw spectral data already provided sufficient predictive capability for certain properties, particularly those with strong spectral signals in the NIR and MIR regions, for instance Ca. Tailored pre-processing did little to further improve these properties because the relevant spectral features were already well-captured in the raw data. This is evident in the scatter plots (Figure 3-4), where points for pH (KCl) and Ca tightly clustered around the 1:1 line, indicating high agreement between predicted and observed values, even without tailored adjustments.

For properties like K and Na, tailored pre-processing had minimal impact on performance, as seen in Figure 3-6. The scatter plots for these properties show continued scattering around the 1:1 line, similar to what was observed in the raw data (Figure 3-4) and blanket pre-processing data (Figure 3-5). This suggests that spectral interferences from overlapping absorption bands or noise, particularly for these challenging properties, limited the effectiveness of tailored pre-processing in this study. The spectral regions associated with K and Na are particularly susceptible to interference from silicates, sulfates, and organic matter (Stenberg et al., 2010). This interference can obscure the spectral signals specific to these elements, complicating accurate prediction. For instance, silicates often dominate the spectral regions relevant to K and Na, masking their unique absorption features (Stenberg et al., 2010). Similarly, organic matter contributes overlapping absorption bands, which further reduces the distinctiveness of the spectral signals for these soil properties (Adeline et al., 2017; Stenberg et al., 2010). Moreover, sulfates, as well as other soil components, add to the spectral complexity, emphasizing the need for advanced pre-processing techniques to isolate and amplify the relevant signals for K and Na (Zhang *et al.*, 2020). Additionally, the relatively small and region-specific dataset used in this study might have limited the potential of tailored pre-processing to fully enhance predictive accuracy. Small datasets often lack the variability needed to identify robust and distinct spectral patterns, which are crucial for the effectiveness of tailored methods (Tsakiridis *et al.*, 2019). Larger and more diverse datasets, encompassing broader soil properties and environmental conditions, have been shown to improve the identification of meaningful spectral features and enhance the performance of tailored approaches (Padarian *et al.*, 2020). By leveraging such datasets, tailored

pre-processing could potentially achieve greater effectiveness, particularly for complex soil properties The scatter plots reinforce this, as the clustering of points for challenging properties like Na remained dispersed across all pre-processing methods, suggesting that the dataset lacked sufficient variability to fully optimize the tailored approach.

The performance of tailored pre-processing also depends on the modelling algorithm used. While Cubist models are effective for many applications, more advanced algorithms, for instance convolutional neural networks (CNN) or gradient boosting machines, could better exploit the refined spectral signals from tailored pre-processing. The scatter plots suggest that while tailored pre-processing slightly improved clustering for some properties (e.g., P (Bray-1)), the underlying model might not have been capable of fully leveraging these enhancements.

In summary, while tailored pre-processing has potential for improving soil property predictions, its benefits were limited in this study, as shown by the scatter plot patterns and statistical results. This underscores the need for targeted strategies, including advanced algorithms, larger datasets, and refined pre-processing techniques, to maximize its utility for challenging properties like K and Na. Future research should aim to address these limitations to further enhance the predictive accuracy of soil spectroscopy models.

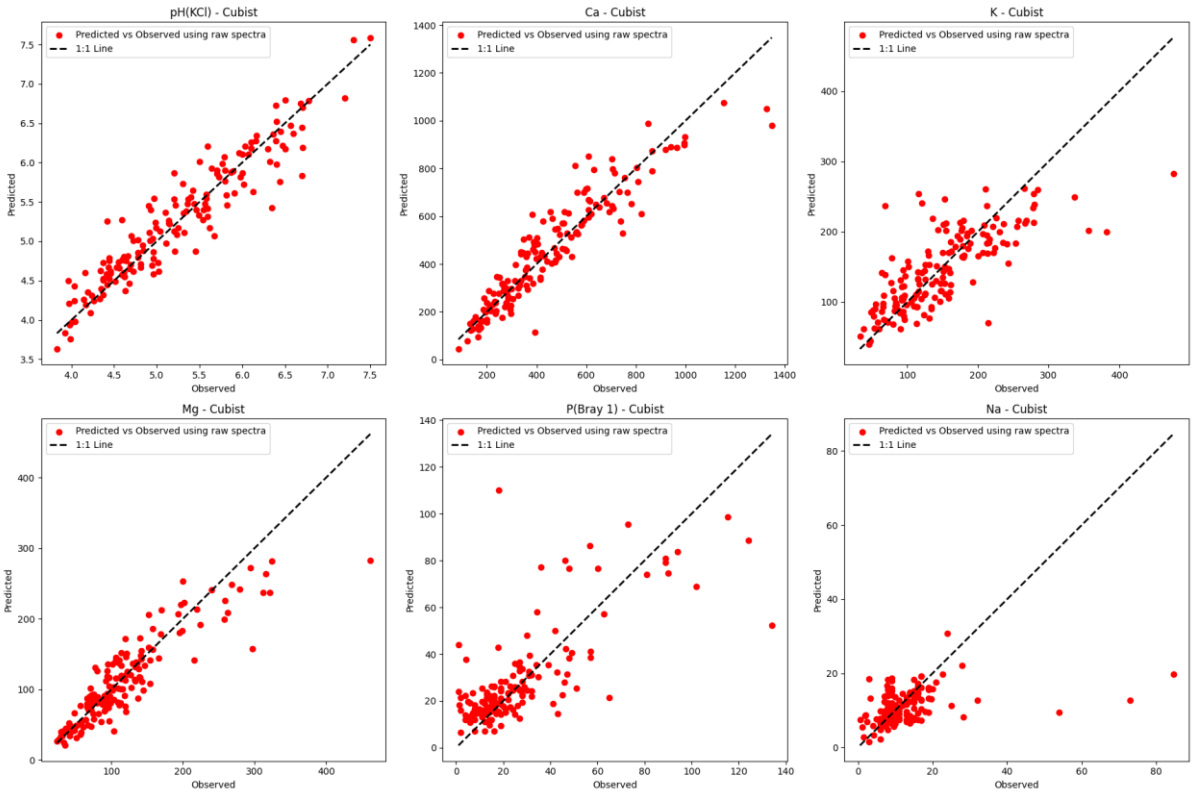

Figure 3-4: Scatter plots of predicted versus observed values for six key soil properties (pH, Ca, K, Mg, P and Na) using the combined NIR and MIR spectral data with no pre-processing. The 1:1 line represents perfect agreement between predictions and observations.

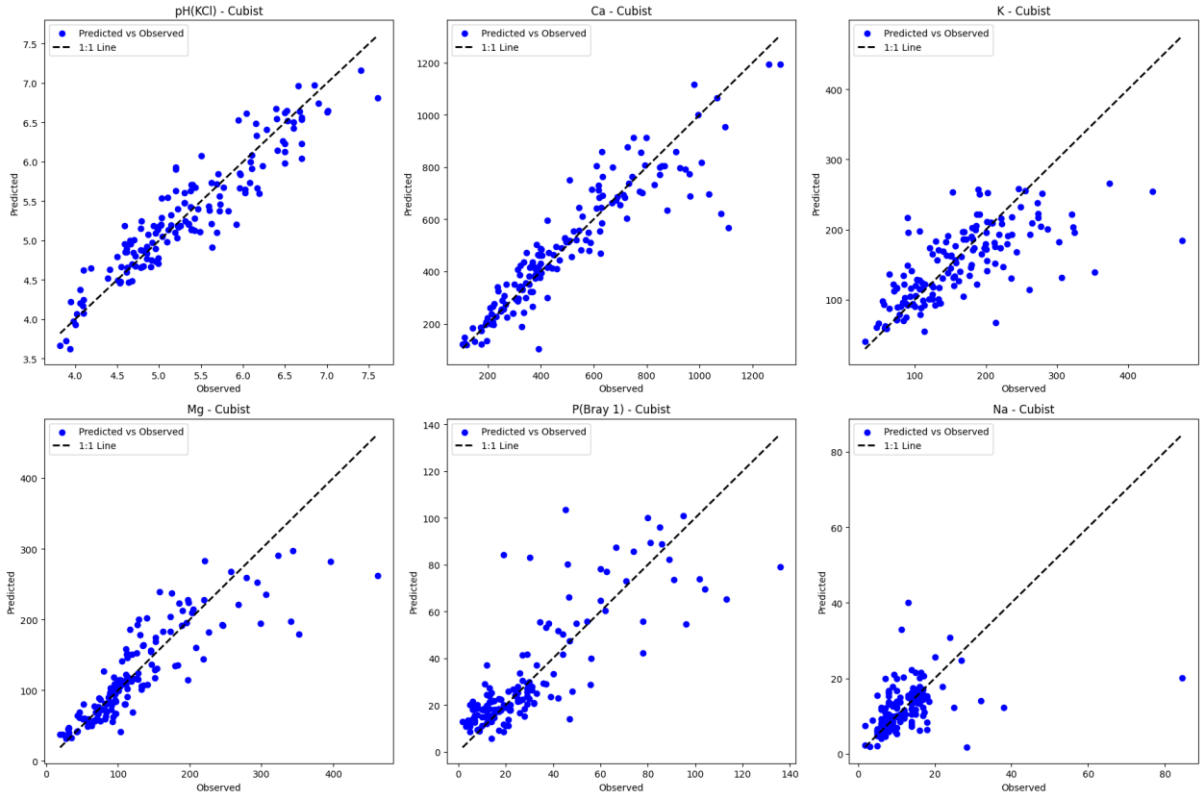

Figure 3-5: Scatter plots of predicted versus observed values for six key soil properties (pH, Ca, K, Mg, P and Na) using the combined NIR and MIR spectral data with blanket pre-processing. The 1:1 line represents perfect agreement between predictions and observations.

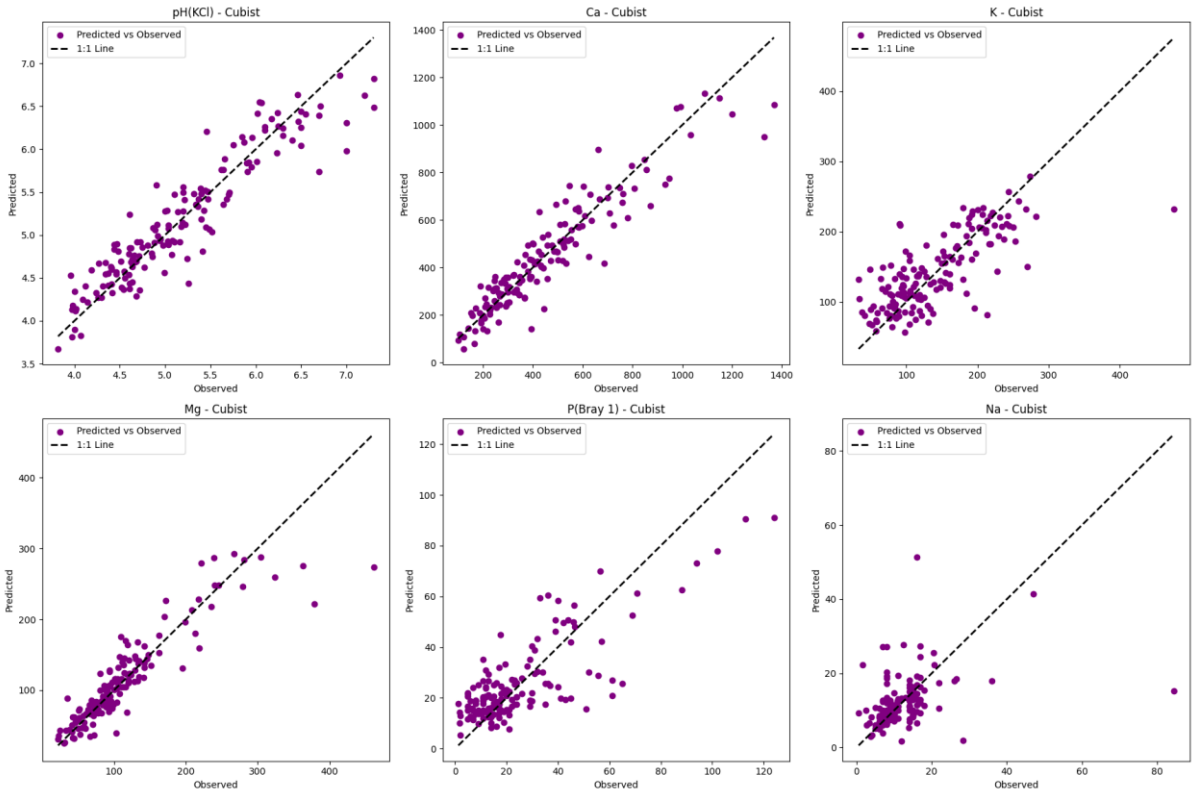

Figure 3-6: Scatter plots of predicted versus observed values for six key soil properties (pH, Ca, K, Mg, P and Na) using the combined NIR and MIR spectral data with tailored pre-processing. The 1:1 line represents perfect agreement between predictions and observations.

3.4.4 Assessment of Predictive Model Accuracy Relative to Laboratory Standards

The predictive accuracy of the developed models varies across different soil properties when compared to laboratory standards, as demonstrated by both the current results and previous findings (Kock et al., 2024). For soil pH (KCl), the model achieved an LCCC of 0.94 and an RMSE of 0.28, reflecting a high degree of accuracy. For soil pH (KCl), the model achieved an LCCC of 0.94 and an RMSE of 0.28, indicating strong predictive performance for field applications where rapid assessments are necessary. However, since laboratory precision typically requires an RMSE of ≤ 0.1, this model should be used as a supplementary tool rather than a replacement for lab-based measurements in regulatory contexts. The models for exchangeable Ca and Mg also demonstrated strong predictive capabilities. For Ca, the LCCC of 0.93 and RMSE of 87.85 suggest strong predictive capability, making the model suitable for practical use in field applications. While it aligns reasonably well with laboratory precision thresholds (± 7.7% CV), further refinements may be required for high-precision regulatory applications (Kock et al., 2024). Similarly, for Mg, the LCCC of 0.89 and RMSE of 31.06 indicate strong predictive accuracy, though further validation is necessary to ensure full compatibility with laboratory standards. These

results, comparable to those achieved in a previous study (R^2 of 0.76 for Mg), suggest that the models can serve as useful tools in precision agriculture, where rapid estimations are prioritized over laboratory-grade precision (Kock et al., 2024).

In contrast, the models for P and K demonstrated limited predictive accuracy. While tailored pre-processing improved performance slightly for P, the LCCC values and RMSE did not meet laboratory standards, as evidenced by results comparable to previous findings (Kock et al., 2024), where R^2 values were ≤ 0.55. The models for K also underperformed, failing to meet the CV threshold of < 18% commonly expected in laboratory analyses. Thus, both properties require continued reliance on laboratory methods for reliable quantification, limiting the applicability of these models for precise soil fertility management. The weakest performance was observed for Na, where the model's LCCC of 0.28 and a substantial RMSE increase indicated significant predictive limitations. These results align with our previous reports on poor model accuracy for Na, with R^2 values below 0.2 (Kock et al., 2024). The inability to accurately predict Na underscores the need for further model refinement or reliance on laboratory analyses for this property. In conclusion, the models for pH (KCl), Ca, and Mg exhibited sufficient accuracy for field-based applications, serving as complementary tools to laboratory methods in scenarios where rapid assessments are critical, though they do not fully replace laboratory precision analyses. However, the predictive accuracy for P, K, and Na remains insufficient for standalone use, emphasizing the continued importance of laboratory analyses for these properties. Future improvements should include techniques that focus on expanding calibration datasets while, refining pre-processing techniques, and also integrating advanced algorithms to enhance the predictive performance for these challenging soil properties.

3.4.5 Study Limitations and Future Research

Combining NIR and MIR data can lead to redundancy in the prediction models, as both spectral regions may capture similar information for certain soil properties. This overlap can introduce noise rather than improving accuracy. Effective fusion strategies are required to maximize useful information while minimizing redundancy. Studies suggest that combining spectra can enhance specific predictions, but techniques like feature selection are necessary to avoid redundant data processing (Knox *et al.*, 2015; Viscarra Rossel *et al.*, 2016).

Soil heterogeneity and environmental conditions, particularly soil moisture, impact spectral signals differently in each region. For instance, moisture absorption affects NIR signals more strongly than MIR, introducing variability in predictions (Stenberg *et al*., 2010; Soriano-Disla *et al*., 2014). Adaptive pre-processing techniques are being researched to manage these variations and standardize models across different soil types (Guerrero *et al.*, 2010). The optimal method for

merging NIR and MIR spectra remains debated, as each spectral region has strengths for different properties. For example, NIR is often better for organic matter, while MIR excels at detecting minerals due to the spectral regions ability to capture more information related to the chemical bonds that make up minerals (Nocita *et al*., 2015). Ongoing research into fusion techniques, for example selective pre-processing for each region, aims to maximize the complementary strengths of both regions (Greschuk *et al.*, 2022).

More complex models like deep learning are effective at handling spectral fusion but often lack interpretability, limiting their real-world utility. For soil scientists and agronomists, interpretable models allow better understanding of which spectral features correlate with soil properties. Research into interpretable machine learning could make fused spectral models more practically applicable (Ng *et al*., 2019). Using advanced techniques like tailored pre-processing, as indicated by our results, can significantly enhance model performance over a generalized (blanket) approach. Studies show that specific pre-processing methods, like baseline correction and smoothing, can enhance features important to soil properties and reduce noise (Stenberg *et al*., 2010; Minasny *et al*., 2011). Further research on customization for soil-specific models could improve prediction accuracy. Large-scale validation across diverse soil types and environments is crucial for building robust models. Many studies note that spectral models trained on one dataset may not generalize well, underscoring the need for more extensive benchmarking and validation (Viscarra Rossel *et al.*, 2016). Future studies could focus on enhancing the generalizability of fused spectral models by validating them across a range of soils. These areas of improvement highlight the need for optimized fusion techniques, adaptive pre-processing, and robust model validation to maximize the practical application of NIR and MIR spectral data in soil science.

3.5 Conclusion

This study demonstrated the significant potential of integrating NIR and MIR spectral data for predicting critical soil properties essential for South African agriculture, including pH (KCl), P (Bray-1), Ca, Mg, K, and Na. By leveraging the complementary strengths of NIR and MIR, the combined spectral approach outperformed models relying solely on a single spectral region, providing robust predictive accuracy for properties like pH (KCl), Ca, and Mg. These models exhibited sufficient precision for practical applications and could serve as viable alternatives to laboratory methods in field-based scenarios. The raw combined spectral data proved particularly effective, underscoring the utility of minimal pre-processing for most soil properties. Tailored pre-processing, while beneficial for properties like P (Bray-1) and Mg, did not universally improve model performance, highlighting the need for property-specific spectral adjustments.

However, the predictive accuracy for Na and K remains inadequate for practical use, aligning with the challenges noted in previous studies. Spectral interference and noise continue to hinder reliable predictions for these properties, emphasizing the importance of refining spectral fusion strategies and exploring advanced machine learning techniques. The study also highlighted the importance of independent validation datasets to ensure the broader applicability of predictive models across diverse soil types and conditions.

This research significantly contributes to the advancement of soil spectroscopy by demonstrating the effectiveness of combining NIR and MIR spectral data. It provides a strong foundation for improving the predictive accuracy of soil property models previously developed for the Western Highveld region. By enabling more accurate and efficient soil property assessments, this integrated approach has the potential to support precision agriculture practices, optimize resource management, and enhance the sustainability of agricultural systems in South Africa. Future efforts should prioritize the development of property-specific spectral pre-processing methods, expand calibration datasets to encompass a broader range of soil types, and adopt advanced machine learning algorithms to further improve prediction accuracy, particularly for challenging soil properties like Na and K.

CHAPTER 4 HARNESSING CONVOLUTIONAL NEURAL NETWORKS FOR PREDICTING SOIL PROPERTIES IN SOUTH AFRICA USING COMBINED NIR AND MIR SPECTRAL DATA

4.1 Abstract

Soil spectroscopy is a transformative analytical approach offering rapid, non-destructive insights into soil properties, crucial for sustainable agricultural practices. However, the heterogeneity of soils, particularly in regions like South Africa's Western Highveld, complicates accurate predictions using soil spectral inference models. Addressing this gap, this study investigated the integration of convolutional neural networks (CNN) with combined near-infrared (NIR) and mid-infrared (MIR) spectral data to enhance the prediction of key soil properties, including pH, phosphorus (P), and exchangeable cations (Ca, Mg, K, and Na). The methodology involved a rigorous pre-processing pipeline, including outlier removal and spectral data fusion, followed by tailored CNN model development optimized for soil spectral analysis. The study utilizes a dataset of 772 samples, partitioned using conditioned Latin Hypercube Sampling, ensuring robust model training, validation, and independent testing. Key findings revealed that CNN excel in capturing subtle spectral patterns, achieving strong predictive performance for properties like pH (R^2 = 0.75, RMSE = 0.40), Ca (R^2 = 0.84, RMSE = 95.40), and Mg (R^2 = 0.74, RMSE = 32.23). However, challenges persisted in predicting Na, P, and K due to spectral overlaps and inherent property variability. This research contributes to a novel application of CNN tailored for soil spectroscopy in South Africa, offering a scalable and efficient alternative to previous soil spectral inference methods. The integration of deep learning models with combined spectral data demonstrated the potential to revolutionize soil property prediction, advancing precision agriculture and sustainable land management practices. The study underscores the importance of model refinement and hybrid approaches for tackling complex spectral interactions, paving the way for improved soil analysis frameworks globally.

4.2 Introduction

Soil spectral inference, a non-invasive analytical technique, has revolutionized the field of pedology by providing rapid and cost-effective methods for characterizing a wide range of soil properties (Shepherd and Walsh, 2002; Van Vuuren et al., 2006; Wadoux et al., 2021). This approach, based on the analysis of the interaction of electromagnetic radiation with soil components, offers critical insights into the physical, chemical, and biological attributes of soils (Wadoux *et al.*, 2021). These insights are essential to understand soil health, guide sustainable

land management practices, and advance precision agriculture, which requires detailed knowledge of soil variability to optimize resource use and maximize crop productivity (Du and Zhou, 2009).

Despite its transformative potential, traditional soil analysis methods remain widely used. However, these methods are time-consuming, expensive, and rely on specialized laboratory equipment, making them impractical for large-scale or real-time agricultural decision making (Angelopoulou *et al.*, 2020). Furthermore, environmental concerns about chemical waste generated during such analyses highlight the need for more sustainable alternatives (Riebe et al., 2019; Stenberg Viscarra Rossel, et al., 2010; Wadoux et al., 2021). While soil spectroscopy addresses some of these limitations by providing a rapid and non-destructive means of analysis, the complexity and heterogeneity of soils (particularly in regions like the Western Highveld of South Africa) pose significant challenges for the accurate prediction of soil properties (Kock et al., 2024; Shepherd and Walsh, 2002). This variability, influenced by various geologies and land use histories, complicates the development of generalized predictive models (Kawamura *et al.*, 2021; Van Vuuren *et al.*, 2006).

Advancements in deep learning, specifically convolutional neural networks (CNN), offer promising solutions to some of these challenges. CNN are particularly proficient at identifying intricate patterns in high-dimensional data, which makes them ideally suited for interpreting complex spectral outputs generated by NIR and MIR spectroscopy (Ng *et al.*, 2020; Padarian *et al.*, 2019). By leveraging CNN it is possible to significantly improve the accuracy and scalability of soil property predictions, providing actionable insights that support precision agriculture practices (Ng *et al.*, 2019). Such advancements are critical to address the limitations of existing soil analysis methods and to enable real-time, data-driven decision-making in agriculture.

This study aimed to connect advanced deep learning techniques and practical soil analysis by exploring the application of CNN to soil spectroscopy, focussing on the diverse and challenging soils of South Africa. By integrating these technologies using soil data from a South African context, this research aimed to address the knowledge gap in precise, scalable, and real-time soil property prediction for South African soils. The findings have the potential to significantly improve already existing MIR soil spectroscopy prediction models created using machine learning techniques, potentially providing farmers with soil analysis tools that are more reliable and robust (Kock et al., 2024).

The objectives of this research were two-fold: first, to evaluate the effectiveness of CNN in predicting key soil properties using NIR and MIR spectroscopy. Second, to address the challenges of soil variability and limited regional data by developing models tailored to the

Western Highveld of South Africa. The methodological framework combined advanced spectral analysis with CNN modelling to optimize predictive accuracy for critical soil attributes, for example pH, P (Bray-1) and extractable cations.

By developing and applying CNN models to tackle the challenges of soil heterogeneity, this research offers a novel and significant contribution to the field of soil spectroscopy in South Africa. This study highlights the transformative potential of integrating more efficient deep learning with spectral analysis and offers a scalable and practical solution for modern agriculture in the country. In doing so, this study not only addresses regional soil spectroscopy challenges, but also advances the global movement toward sustainable agriculture and food security (Ahmadi *et al.*, 2021).

4.3 Material and Methods

4.3.1 Data Collection and Pre-processing

The study used a combined dataset of NIR and MIR spectral data ranging from 400 cm^{-1} to 7 400 cm^{-1} obtained from soil samples within the western Highveld region of South Africa with corresponding soil property values for soil properties pH (KCl), P (Bray-1), Ca, Mg, K and Na cations extracted using the 1M ammonium acetate method. The spectral data, presented in absorbance, was measured in wavenumbers using a Bruker Alpha II with DRIFTS module (Bruker OPTIK GmbH, Ettlingen, Germany, 2019) for MIR (400 - 4 000 cm^{-1}) and NIR (4 000 - 7 400 cm^{-1}) with a handheld FT-NIR Neospectra NIR spectral scanner from Si-ware instruments(Si-ware Systems, 2023). Using the bind function in R version 4.2.0 (R Core Team, 2023), the NIR spectral data was integrated into the Kock et al. (2024) spectral library, which already includes MIR spectral data. Before modelling, the soil spectral library, comprising both laboratory-measured soil properties and spectral data, underwent an initial outlier removal process using Mahalanobis distance, with a threshold of three standard deviations applied to absorbance values (Wadoux *et al.*, 2021). Samples outside this range were excluded, resulting in a refined set of 772 samples. The raw NIR and MIR spectral data was combined using spectra concatenation to create one dataset ranging between 400 - 7 4 00 cm^{-1} (Wadoux *et al.*, 2021). Chapter 3 of this thesis demonstrated that raw combined spectral data yielded the highest prediction model accuracy. Based on these findings, this approach was adopted for this chapter.

4.3.2 Spectral Data Visualization

To visualize the spectral profile of each soil sample (Figure 4-1), we plotted spectral absorbance data were plotted against wave numbers using Python's matplotlib and seaborn libraries (Hassan

Sial *et al.*, 2021). These tools allowed for detailed and customizable plotting, highlighting spectral variation across samples, which is essential for interpreting soil property data. This representation provided a clear view of the data range, alignment, and potential anomalies within spectral profiles. The ability to visualize these patterns aided in identifying outliers or irregularities, ensuring that the data is suitable for machine learning analysis and that the model training was based on reliable and well-aligned spectral profiles.

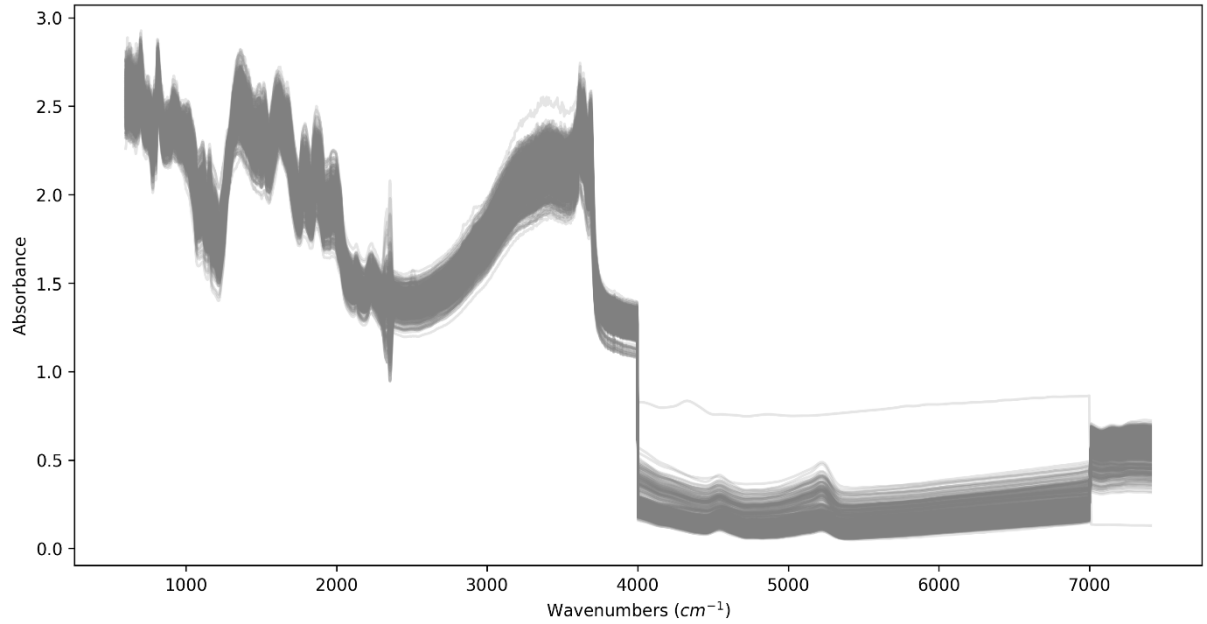

Figure 4-1: Combined raw NIR and MIR spectral data of soil samples across the spectral ranges (400 - 7 400 cm^{-1}). The plot illustrates the absorbance variability and spectral features captured from the soil samples, with notable distinctions in spectral regions corresponding to different wavenumbers, highlighting the complex interactions of soil components with electromagnetic radiation.

4.3.3 Data Splitting Strategy

The data set was first partitioned using conditioned Latin Hypercube Sampling (cLHS) to obtain diverse training and validation samples (Minasny and Mcbratney, 2006). Soil property values were scaled to their specific ranges using Latin Hypercube sampling, which ensured that the selected samples covered the full range of soil properties. Eighty percent of the data set was allocated to training, and the remaining 20% was assigned to independent test set. The training set was further divided into calibration and internal validation parts in a 90 to 10% ratio, respectively (Xu *et al.*, 2018). This segmentation supports model refinement and continuous performance tracking during its training phase, guaranteeing that the model is well adjusted before the final validation stage (Ng *et al.*, 2019; Padarian *et al.*, 2020). This multilayer split method aided in reducing overfitting,

allowing the model to better generalize to new, unseen data by mimicking real-world variability throughout the training process.

4.3.4 Correlation Analysis of Soil Properties

To investigate the interrelationships between the properties of the target soil, a correlation analysis was performed. Pearson's correlation coefficients were calculated for all pairwise combinations of pH, P, Ca, K, Mg and Na using the training data set. The resulting correlation matrix was visualized using a heatmap generated with the Seaborn library in Python (Hassan Sial *et al.*, 2021). This analysis provides valuable information on the relationships between soil properties, which can inform the interpretation of the performance of the CNN model and guide further model development. For example, strong correlations between certain soil properties may indicate that CNN can leverage these relationships to improve prediction accuracy, potentially by identifying shared spectral characteristics or patterns. This aligns with the concept of "transfer learning" in machine learning, where knowledge gained from one task can be applied to related tasks (Liu *et al.*, 2018). On the other hand, high correlations could lead to multicollinearity, potentially affecting model stability and interpretability. This phenomenon can increase the variance of regression coefficients, making it difficult to isolate the individual effects of correlated predictors (John *et al.*, 2021). Understanding these relationships is crucial for optimizing the model architecture and training process.

4.3.5 Neural Network Model Architecture

The predictive model used a one-dimensional CNN, optimized to handle spectral data by identifying and exploiting subtle spectral features unique to each soil property. This approach builds on previous work that highlighted CNN as effective tools for extracting hierarchical characteristics from soil spectra for improved prediction accuracy (Ng *et al.*, 2019; Padarian *et al.*, 2019).

The architecture began with a convolutional layer, designed with 10 filters of varying sizes, each acting as a feature detector across the spectral data. These filters applied a kernel across the input, effectively capturing local dependencies and patterns in the spectral signals. Padding was applied to preserve the input-output dimensionality, preventing loss of boundary information, and ensuring consistent feature representation across layers. This configuration, tailored for spectral data, outperforms fully connected layers, as it reduces the number of parameters and captures localized patterns without excessive computational burden (Ng *et al.*, 2019; Padarian *et al.*, 2020).

The model incorporated the Rectified Linear Unit (ReLU) as its activation function, transforming each feature map non-linearly and introducing crucial non-linear capabilities to the model. ReLU is widely used in deep learning due to its simplicity and computational efficiency, as it enables rapid convergence by preventing the vanishing gradient problem common in other activation functions (Kawamura *et al.*, 2021; Ng *et al.*, 2019; Padarian *et al.*, 2020). By allowing the network to learn more complex representations, ReLU helps to capture the unique spectral features associated with different soil properties, which is essential given the subtle variations in soil spectra (Ng *et al.*, 2019; Padarian *et al.*, 2020).

To reduce the computational load and avoid overfitting, a MaxPooling1D layer was introduced after the convolutional layers (Ng *et al.*, 2019; Padarian *et al.*, 2020). Max-clustering down sampled the feature maps by capturing the most prominent features within each pooling window, reducing the spatial dimensions of the data while retaining essential information. This selective down sampling helps avoid model overfitting by discarding irrelevant data while emphasizing dominant spectral features, an approach proven to be effective in reducing model complexity without sacrificing predictive performance in soil spectral applications (Kawamura *et al.*, 2021; Ng *et al.*, 2019; Padarian *et al.*, 2020).

The final portion of the network consisted of flattening- and fully connected (dense) layers, which processed the extracted features to generate the final output (Ng *et al.*, 2019; Padarian *et al.*, 2020). The flattening layer transformed the pooled feature map into a one-dimensional vector, which was then passed through two dense layers. These layers facilitated feature combination and abstraction, allowing the model to make nuanced predictions for each soil property by drawing on cumulative knowledge from all prior layers. The output layer produced a single prediction for each soil property of interest, aligning with the goal of deriving accurate property-specific predictions from spectral data (Ng *et al.*, 2019; Padarian *et al.*, 2020).

This CNN architecture was selected for its ability to handle spectral data efficiently, extracting meaningful spectral features that traditional machine learning models might overlook (Ng *et al.*, 2019; Padarian *et al.*, 2020). By using these layers in combination, the model could achieve a balance of computational efficiency and predictive power, aligning with recent advances in CNN applications for soil spectroscopy (Albinet *et al.*, 2022; Ng *et al.*, 2019; Padarian *et al.*, 2020).

4.3.6 Model Optimization and Hyperparameter Tuning

To optimize the performance of the CNN, a hyperparameter tuning was performed using Optuna, a Python-based library suited for efficiently exploring a wide range of hyperparameter options. This process was essential to refine the CNN model, as selecting appropriate hyperparameters

is critical to balancing model accuracy, complexity, and computational efficiency, particularly in spectral data applications (Akiba *et al.*, 2019). The tuning focused on key parameters, including the number of convolutional filters, kernel size, pool size, number of dense units, dropout rate, learning rate, and batch size. Each of these hyperparameters directly affects the model's ability to generalize while managing the computational load. During optimization, Optuna generated a variety of CNN configurations by sampling values from the predefined hyperparameter ranges (Akiba *et al.*, 2019). Each configuration was trained in a calibration set and evaluated 10% internal validation set to achieve the lowest possible MSE. Mean squared error was chosen as it emphasizes minimizing prediction errors, which makes it particularly useful for ensuring accuracy in soil property predictions (Ng *et al.*, 2019; Padarian *et al.*, 2020). Through repeated iterations, Optuna was able to identify the most effective combination of hyperparameters for our data, providing a tailored model that achieved a balance of predictive accuracy and efficiency. Using an adaptive tuning approach allowed the model to reach its full potential, maximizing CNN's ability to extract meaningful features from spectral data without overfitting or underutilizing resources. The final hyperparameters identified by Optuna were then applied to the model, resulting in an optimized setup to predict soil properties from spectral measurements.

4.3.7 Model Training and Independent Dataset Evaluation

The CNN models were meticulously configured to ensure optimal performance in predicting soil properties from spectral data. The MSE was selected as the loss function due to its effectiveness in penalizing larger errors more significantly, thereby encouraging the model to focus on minimizing substantial discrepancies between predicted and actual values (Ng *et al.*, 2019; Padarian *et al.*, 2020). Alongside MSE, MAE was used as a metric for accuracy, offering a straightforward interpretation of the average magnitude of prediction errors without considering their direction (Ng *et al.*, 2019; Padarian *et al.*, 2020).

Training was carried out over 200 epochs using Adam Optimizer, an algorithm renowned for its computational efficiency and ability to handle sparse gradients, which is particularly advantageous when dealing with high-dimensional spectral data (Khaire and Dhanalakshmi, 2020). The optimizer adjusts the learning rate adaptively, facilitating faster convergence and improved performance compared to traditional stochastic gradient descent methods (Ng *et al.*, 2019; Padarian *et al.*, 2020). Throughout the process, the model performance was rigorously evaluated using calibration, internal validation, and independent test datasets. This stratified evaluation approach is critical for assessing the model's generalization capabilities and ensuring its robustness when applied to new, unseen data (Ng *et al.*, 2019; Padarian *et al.*, 2020). The following statistical metrics were calculated to provide a comprehensive evaluation of the model.

- Coefficient of Determination (R^2) in Equation 4-1 quantifies the proportion of variance in the response variable that is predictable from the predictor variables. An R^2 value closer to 1 indicates a strong predictive relationship. This can also be expressed as 1 − SSE/SST where SSE is the sum of the squared error and SST is the total sum of squares (Wadoux *et al.*, 2021) .

Equation 4-1

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Obs_i - pred_i)^2}{\sum_{i=1}^{n}(Obs_i - \overline{Obs})^2}$$

- Root Mean Squared Error (RMSE) in Equation 4-2, provides an absolute measure of the differences between response and predictor values, expressed in the same units as the target variable. It is sensitive to large errors, making it a reliable indicator of model performance (Kuang *et al.*, 2015; Wadoux *et al.*, 2021)

Equation 4-2

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Obs_i - pred_i)^2}$$

- Bias or mean error (ME) measures the systematic error in the predictions as shown in Equation 4-3. A non-zero bias indicates consistent overestimation or underestimation by the model, which is crucial to identify for corrective measures (Wadoux *et al.*, 2021).

Equation 4-3

$$ME = \frac{1}{n}\sum_{i=1}^{n}(obs_i - pred_i)^2$$

- Ratio of Performance to Interquartile Distance (RPIQ) in Equation 4-4 assesses predictive performance relative to the variability in the data set. Higher RPIQ values signify better model performance, especially important in heterogeneous soil datasets (Wadoux *et al.*, 2021).

Equation 4-4

$$RPIQ = \frac{(Q_3(Obs) - Q_1(Obs))}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Obs_i - pred_i)^2}}$$

Lin's Concordance Correlation Coefficient (LCCC) in Equation 4-5 evaluates both precision and accuracy by measuring how far the observed data deviate from the line of perfect concordance. It is particularly useful for assessing the agreement between observed and predicted values in soil spectroscopy (Kawamura *et al*., 2021).

Equation 4-5

$$LCCC = \frac{2r\sigma_{pred}\sigma_{Obs}}{{\sigma_{Obs}}^2 + {\sigma_{pred}}^2 + (\mu_{Obs} - \mu_{pred})^2}$$

By employing these metrics, the predictive accuracy and reliability of the model were thoroughly assessed, ensuring that it met the necessary standards for practical application in soil property prediction. The use of multiple datasets for evaluation aligns with best practices in machine learning, mitigating the risk of overfitting and enhancing the model's ability to generalize (Wadoux *et al.*, 2021). Upon completion of training, the final model was saved, facilitating its deployment for future predictions on new spectral data without the need for retraining. This step is essential for practical applications, enabling efficient and consistent soil property assessments across different data sets and study areas (Padarian *et al.*, 2020). Performance is evaluated based on three distinct datasets: calibration, validation, and test sets. The calibration set was used to train the CNN models, while the internal validation set guided model selection by helping identify the best-performing architecture and tuning the hyperparameters. The independent test set provides an unbiased assessment of the models' ability to generalize to new, unseen data (Ng et al., 2019).(Ng *et al.*, 2019)

4.4 Results and Discussion

4.4.1 Model Performance Evaluation

4.4.1.1 Calibration of CNN models

The CNN models demonstrated exceptional predictive capacity during calibration, achieving consistently high R^2 values (< 0.99), low RMSE, and near-perfect LCCC values (> 0.93) for all soil properties when predicting on the calibration dataset, as summarized in Table 4-1. Specifically, pH (R^2 = 0.99, RMSE = 0.07, LCCC = 0.99) and K (R^2 = 0.98, RMSE = 10.54, LCCC = 0.99) showed outstanding calibration performance. These results indicate that the models effectively learned the relationships between the spectral data and these soil properties during training, achieving nearly flawless predictions.

The scatter plots in Figure 4-2 provide further visual evidence of the calibration performance. For pH, the calibration data points align closely along the line of perfect prediction or 1:1 (red dashed line), reflecting the high predictive accuracy. Similarly, for K, Mg, and Ca, the calibration predictions show strong clustering around the 1:1 line, reinforcing the metrics of R^2 and LCCC. For properties for instance Ca (R^2 = 0.95, RMSE = 54.62, LCCC = 0.97) and Mg (R^2 = 0.89, RMSE = 23.08, LCCC = 0.94), the relatively higher RMSE values reflect the greater variability in the dataset for these properties. Calcium, for instance, exhibited the highest RMSE, likely due to its wide concentration range in the calibration samples. Despite this variability, the high LCCC values for Ca and Mg suggest excellent agreement between predicted and observed values, as seen in the tight clustering of calibration points in Figure 4-2. These results demonstrate the CNN models' ability to accurately capture variations in soil properties during the training phase.

4.4.1.2 Validation of CNN Models

The validation phase assessed the generalizability of the CNN models to “unseen” data. Although the R^2, RMSE, and LCCC metrics decreased compared to calibration, the performance remained strong for key properties for instance pH, Ca, and Mg. For pH, the validation results achieved values of R^2 = 0.69, RMSE = 0.46, and LCCC = 0.81, demonstrating that the model could generalize well to “unseen” samples. Figure 4.2 further supports these metrics, where pH validation points, though slightly more dispersed, still align closely with the ideal prediction line. For Ca and Mg, the validation results were similarly robust, achieving R^2 values of 0.78 and 0.67, respectively, with corresponding LCCC scores of 0.87 and 0.78. However, for properties like Na (R^2 = 0.18, RMSE = 10.36, LCCC = 0.38) and P (R^2 = 0.50, RMSE = 17.23, LCCC = 0.65), the models struggled to generalize effectively, as reflected by the increased scatter and deviation

from the prediction line in Figure 4-2. For K, the moderate performance (R^2 = 0.49, RMSE = 76.32, LCCC = 0.49) also highlights the challenges of predicting properties with more complex spectral interactions. These results suggest the need for further optimization in the pre-processing and modelling strategies to better account for the spectral variability of these properties.

4.4.1.3 Test of CNN Models

The independent test set provides the most reliable assessment of the CNN models' predictive capabilities in real-world conditions. Encouragingly, the models maintained strong performance for pH (R^2 = 0.75, RMSE = 0.40, LCCC = 0.86), Ca (R^2 = 0.84, RMSE = 95.40, LCCC = 0.91), and Mg (R^2 = 0.74, RMSE = 32.23, LCCC = 0.85), as shown in Table 4-1. The scatter plots in Figure 4-2 illustrate this robustness, where test set predictions for these properties exhibit tight clustering along the 1:1 line, indicating strong agreement between observed and predicted values. In contrast, the models exhibited limited predictive performance for Na, P, and K, with R^2 values below 0.40 and lower LCCC scores (e.g., Na = 0.60, P = 0.63, K = 0.52). As shown in Figure 4-2, the test set predictions for these properties are more widely dispersed, deviating significantly from the ideal prediction line. This lower performance likely reflects challenges related to the spectral overlap of these properties with other soil constituents, as well as the inherent variability in their concentrations within the dataset. Improving predictions for these elements may require more sophisticated pre-processing or feature engineering techniques to better isolate their spectral signatures.

The results achieved in this study for CNN-based predictive performance in soil spectroscopy are consistent with findings from similar studies. Ng et al. (2019) demonstrated that CNN models, particularly when using fused Vis-NIR and MIR spectral data, can achieve high accuracy for properties like pH, Ca, and Mg, with R^2 values ranging from 0.95 to 0.98. This highlights the potential of deep learning approaches to outperform traditional models for example partial least squares regression (PLSR) in soil spectroscopy applications. Similarly, Haghi et al. (2021) showed that CNN models outperformed other regression methods, including PLSR and support vector regression (SVR), with R^2 values exceeding 0.80 for exchangeable Mg and K, further supporting the use of CNN for soil property prediction.

Despite these successes, challenges remain in predicting properties like Na, P, and K due to spectral overlaps and variability in their concentrations. These issues have been widely reported in other studies. For instance, Islam et al. (2003) observed weaker predictions for Na and K (R^2 < 0.60), highlighting the inherent difficulty in isolating their spectral signatures. Similarly, Lu et al. (2020) used laser-induced breakdown spectroscopy combined with multivariate regression techniques, achieving strong performance for Mg (R^2 ~ 0.90), but lower values for P and Na,

underscoring the limitations of spectral-based methods for these elements. Data fusion approaches have been shown to improve predictions for challenging soil properties. Kandpal et al. (2022) demonstrated significant improvements in predicting Mg, Na, and Ca by combining MIR and XRF spectral data using CNN and PLS models. Their results yielded LCCC scores comparable to those reported for Ca and Mg in this study. This suggests that integrating multiple spectral data sources could enhance the accuracy of soil property predictions. The findings of this study align with existing literature, confirming strong CNN model performance for properties like pH, Ca, and Mg, while also reflecting common challenges in predicting Na, P, and K due to their spectral complexity. Future improvements may require advanced pre-processing techniques or data fusion methods to better isolate the spectral signatures of these elements and further enhance predictive performance.

Table 4-1: Statistical performance metrics of the CNN models for predicting soil properties across calibration, validation, and independent test datasets. Metrics include R^2, RMSE, bias, RPIQ and LCCC. The table highlights the models' predictive accuracy, robustness, and generalization capabilities for each soil property, with pH, Ca, and Mg showing the highest performance across all datasets.

Property	**R^2**	**RMSE**	**Bias**	**RPIQ**	**LCCC**	**R^2**	**RMSE**	**Bias**	**RPIQ**	**LCCC**	**R^2**	**RMSE**	**Bias**	**RPIQ**	**LCCC**
	Calibration					**Validation**					**Independent test set**				
pH (KCl)	0.99	0.07	-0.01	17.55	0.99	0.69	0.46	-0.05	2.50	0.81	0.75	0.40	-0.06	2.49	0.86
P (Bray-1) (mg kg^{-1})	0.97	4.43	-1.13	4.58	0.98	0.50	17.23	0.34	1.22	0.65	0.40	15.61	0.05	1.15	0.63
Ca (mg kg^{-1})	0.95	54.62	10.36	5.94	0.97	0.78	105.70	10.07	2.62	0.87	0.84	95.40	6.13	3.50	0.91
K (mg kg^{-1})	0.98	10.54	-3.19	9.92	0.99	0.26	76.32	-3.21	1.32	0.49	0.31	69.26	-4.21	1.46	0.52
Mg (mg kg^{-1})	0.89	23.08	-2.10	2.79	0.94	0.67	40.63	-10.96	1.62	0.78	0.74	32.23	-3.91	2.02	0.85
Na (mg kg^{-1})	0.89	2.66	0.29	2.63	0.93	0.18	10.36	-0.50	0.87	0.38	0.37	4.73	0.52	1.48	0.60

R^2; coefficient of determination, RMSE; root mean squared error, RPIQ; ratio of performance to interquartile range, LCCC; Lin's Concordance Correlation Coefficient.

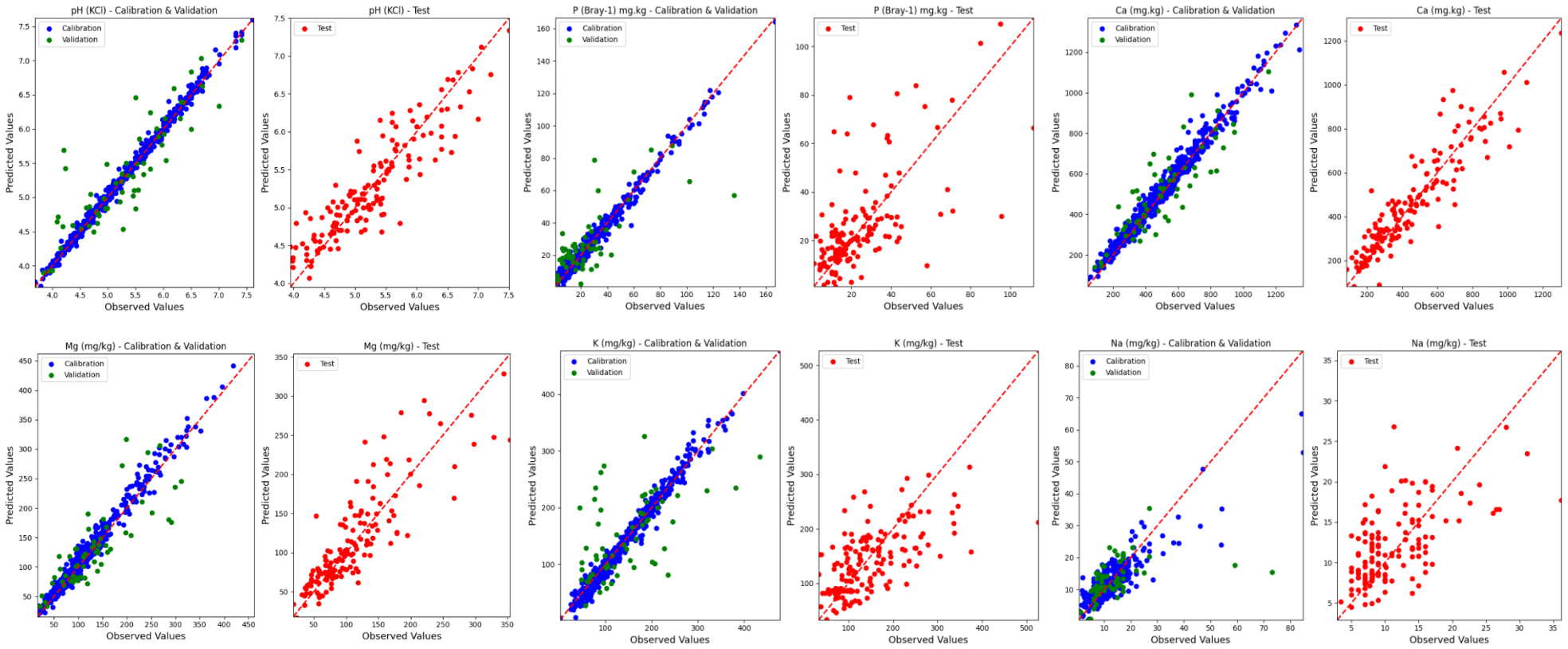

Figure 4-2: Scatter plots showing observed versus predicted soil property values for pH, P, Ca, K, Mg, and Na during calibration, validation, and independent test phases. Points are color-coded to represent datasets: calibration (blue), validation (green), and test (red). The red dashed line represents the 1:1 ideal prediction line. The plots illustrate the model's predictive accuracy and agreement, with tighter clustering around the line indicating stronger model performance for properties like pH, Ca, and Mg, while more dispersion is observed for Na and P.

4.4.2 CNN Architectures

The CNN architecture employed in this study, as depicted in Figure 4-3, leverages one-dimensional convolutional layers specifically tailored for soil spectroscopy data to predict various soil properties effectively. Each property (e.g., pH, Ca, K) was modelled using its unique CNN architecture, beginning with a tailored input layer followed by convolutional, pooling, and fully connected layers. For instance, the architecture predicting pH used an initial convolutional layer with a kernel size of 7 × 1 × 36 and ReLU activation, optimized to extract critical spectral features while retaining sequential and spatial data characteristics. This modular design ensures that each model addresses property-specific spectral patterns. One-dimensional CNN have proven particularly effective in capturing non-linear and hierarchical relationships in spectral data without requiring extensive pre-processing steps for example normalization or smoothing (Ng *et al.*, 2019). Compared to traditional regression models like PLSR and Cubist, the CNN models in this study achieved improved predictive accuracy across several soil properties. ReLU activation further enhanced the model's efficiency by mitigating vanishing gradient issues and enabling the detection of subtle spectral variations. The architecture also included pooling layers to reduce computational complexity, ensuring robust feature extraction without overfitting. This framework aligns with findings that CNN models benefit significantly from spectral fusion methods and require less pre-processing while still maintaining superior performance compared to tree-based models (Padarian *et al.*, 2020). The use of property-specific CNN architectures in this study reflects the nuanced spectral features associated with different soil attributes. For example, high variability in spectral signals for elements like Na required more sophisticated feature extraction, highlighting the flexibility of CNN architectures in addressing property-specific challenges. Additionally, this tailored approach effectively manages computational resources, making it scalable for large-scale applications. Future optimizations, which employs transfer learning or multi-channel architectures, could enhance generalizability and performance further. This study confirms that CNN are a transformative tool in soil spectroscopy, offering a scalable and robust solution for precise soil property predictions across diverse agricultural settings.

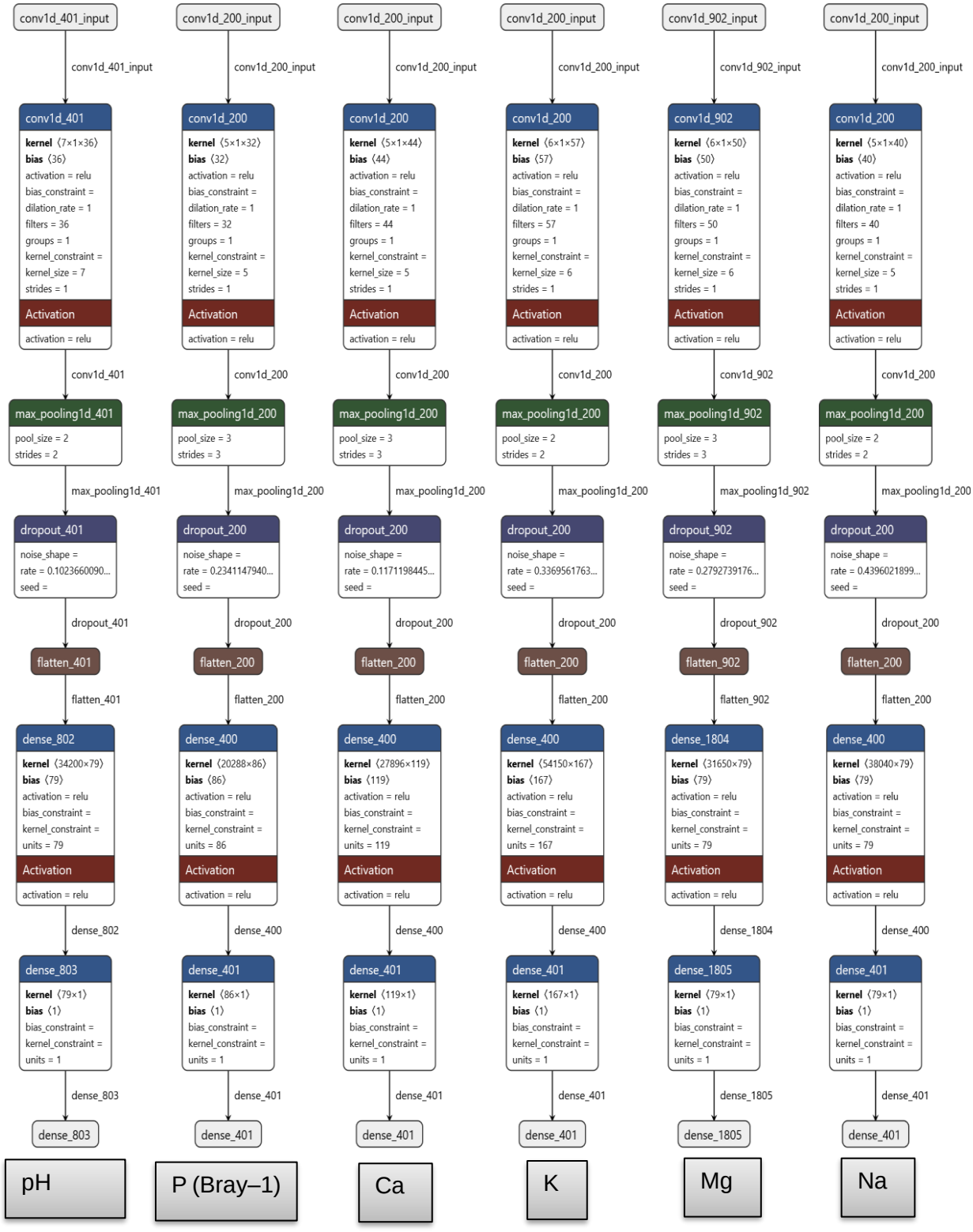

Figure 4-3: Complete architecture of the CNN models used for predicting individual soil properties (pH, P, Ca, K, Mg, and Na). Each model features separate convolutional layers (conv1d) with tailored kernel sizes and filters, followed by max-pooling, dropout, flattening, and dense layers. This design ensures the extraction of relevant spectral features unique to each soil property while preventing overfitting through dropout layers and pooling operations. The final dense layer outputs property-specific predictions, reflecting the individualized approach for each target variable.

4.4.3 Comparison with Prior Methods

The CNN models demonstrated inferior performance compared to Cubist models from Chapter 3 across most soil properties. For pH (KCl), the CNN models underperformed compared to the Cubist-Combined-Raw model, achieving an R^2 of 0.89, RMSE of 0.28, RPIQ of 4.37, and LCCC of 0.94 on the independent test set, compared to the CNN model's R^2 of 0.75, RMSE of 0.40, RPIQ of 2.49 and LCCC of 0.86. While the CNN-pH model showed lower RPIQ for pH, their predictive robustness, reflected in comparable LCCC values, highlights their ability to generalize well to unseen data. This superior performance of Cubist models for pH is consistent with previous studies that found Cubist excels in leveraging complex non-linear relationships in soil spectroscopy data (Ng *et al.*, 2019). For Ca, the CNN model closely aligns with those of the Cubist models and achieved strong predictive accuracy with a R^2 of 0.84, RMSE of 95.40, RPIQ of 3.50, and LCCC of 0.91 compared to the Cubist-Ca model with a R^2 of 0.87, RMSE of 87.85, RPIQ of 3.63, and LCCC of 0.93. This is reflective of both model's ability to capture the spectral signatures of Ca effectively, aligning with findings that CNN models can perform comparably to Cubist in highly structured datasets (Haghi *et al.*, 2021). The Mg CNN exhibited similar performance compared to Cubist, achieving an R^2 of 0.74, RMSE of 32.23, RPIQ of 2.02, and LCCC of 0.85. In comparison, the Cubist models achieved slightly higher accuracy, with an R^2 of 0.82, RMSE of 30.09, RPIQ of 1.96 and LCCC of 0.90. The Cubist models for P far outperformed the CNN model on all statistical metrics with a R^2 of 0.64, RMSE of 12.60, RPIQ of 1.59, and LCCC of 0.77 compared to the CNN-P with a R^2 of 0.40, RMSE of 15.61, RPIQ of 1.15, and LCCC of 0.63. This underperformance by CNN indicates challenges in extracting features for P from the spectral data, which is known to involve complex and weak spectral interactions (Haghi *et al*., 2021). The performance gap for K was substantial. Potassium CNN model achieved an R^2 of 0.31, RMSE of 69.26, RPIQ of 1.46 and LCCC of 0.52, while the Cubist models achieved R^2 of 0.57, RMSE of 47.55, RPIQ of 1.91 and LCCC of 0.73. Sodium on the other hand performed better using CNN, with metric values of R^2 of 0.37, RMSE of 4.73, RPIQ of 1.48 and LCCC of 0.60 compared to the best performing Cubist model with a R^2 of -0.55, RMSE of 6.81, RPIQ of 1.17 and LCCC of 0.45. This suggests that CNN architectures may be better suited for capturing weak and noisy spectral features, as highlighted in other soil spectroscopy studies (Nyawasha *et al.*, 2024).

While Cubist models excelled in most metrics, CNN demonstrated reasonable performance for key properties like Ca and Mg and outperformed for Na. The flexibility, scalability, and potential for optimization underscore CNN's transformative potential for high-dimensional soil spectral data analysis. Future work should focus on refining CNN architectures to close the performance gap with traditional machine learning models like Cubist. This comparative analysis underscores the

importance of selecting modelling techniques tailored to the specific challenges of soil spectroscopy datasets.

Using CNN instead of Cubist models can reduce computational costs and time in applications where pre-processing and scalability are critical. CNN typically benefit from their ability to process raw or minimally processed data directly, eliminating the need for extensive pre-processing required by Cubist models. This reduction in pre-processing steps enhances scalability and operational efficiency, particularly in high-dimensional datasets (Ioannou *et al.*, 2016). However, the upfront investment in training CNN is higher due to their complexity and reliance on specialized hardware like graphics processing units (GPUs), which offer significant acceleration in computations. Studies have shown that modern CNN architectures optimized for efficiency can substantially lower training times and computational expenses compared to conventional deep learning models, making them more suitable for applications requiring scalability (Lin *et al.*, 2019). Future research could focus on optimizing CNN architectures to further reduce computational overhead, making them more competitive with simpler methods like Cubist (Wang *et al.*, 2023). Convolutional neural networks offer significant advantages in agricultural applications, particularly in precision agriculture, by enabling rapid, automated processing of large datasets. Kamilaris and Prenafeta-Boldú, (2018) conducted a comprehensive review of CNN applications in agriculture, highlighting their efficiency in handling complex agricultural datasets and the potential long-term cost benefits despite higher initial computational demands. This study emphasizes the transformative role of CNN in scaling precision agricultural methods, which is particularly relevant in contexts like South African agriculture where automation and scalability could be critical for the agricultural sector.

4.4.4 Correlation among Soil Properties

The correlation analysis -Figure 4-4- revealed several key relationships between soil properties, providing insight into both the dataset and model predictions. Strong positive correlations were observed between Ca and Mg ($r = 0.88$) and Ca and K ($r = 0.55$), consistent with their shared geochemical origins and roles in soil mineral phases (Sparks, 2003). These correlations suggest that spectral features associated with these elements likely contributed to the strong predictive performance of the CNN models for these properties. Low correlations between P and Ca ($r = 0.27$) highlight the influence of Ca on the fixation of P in soils (Mengel *et al.*, 2001). Meanwhile, the weak correlation between Na and K ($r = 0.19$) further reflects competitive adsorption behaviours in soil colloids. From a modelling perspective, these correlations offer both opportunities and challenges. On the one hand, the presence of correlated features can help CNN models learn shared spectral patterns, improving predictive accuracy. On the other hand, high

correlations can introduce multi-collinearity, which may reduce model stability and interpretability. Strategies which include regularization or feature selection could mitigate these effects (Dormann *et al.*, 2013).

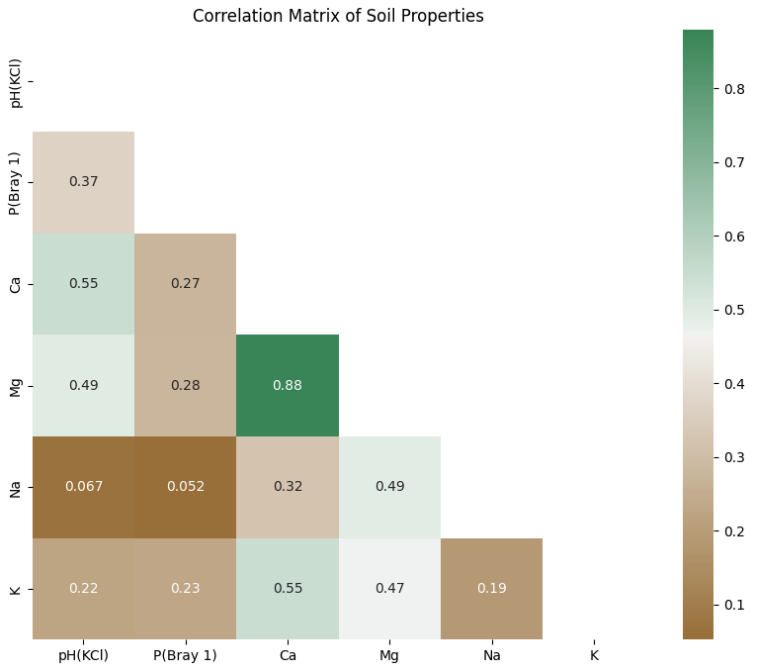

Figure 4-4: Correlation matrix of soil properties (pH, P, Ca, Mg, Na, and K) derived from the dataset. The colour intensity indicates the strength of the correlation, with values ranging from -1 (strong negative correlation) to +1 (strong positive correlation). Notable correlations include a strong positive relationship between Ca and Mg (r = 0.88), and moderate correlations between Ca and P (r = 0.27) and Mg and K (r = 0.0.47). These relationships highlight potential interactions and shared geochemical pathways influencing soil properties.

4.4.5 Conclusion

The study demonstrates a significant advancement in the application of CNN to soil spectroscopy, providing a scalable and innovative solution to challenges in soil property prediction. By integrating CNN architectures with combined NIR and MIR spectral data, the study achieved robust predictive accuracy for key soil properties pH, Ca, and Mg. The approach underscores the ability of CNN to handle high-dimensional spectral data effectively while reducing the need for extensive pre-processing, a notable improvement over other machine learning methods. The novelty of the study lies in its tailored CNN architectures designed to address the unique spectral characteristics of different soil properties, including the challenges posed by soil heterogeneity and complex spectral overlaps. This specificity allowed the models to excel in capturing intricate patterns, particularly for properties with well-defined spectral signatures. The findings provide a practical framework for leveraging CNN in soil spectroscopy for South Africa, aligning with broader goals of precision agriculture and sustainable land management. Despite these achievements,

the study highlights areas for improvement. CNN models underperformed relative to Cubist models for certain properties like K and P, indicating the need for more precise pre-processing techniques or hybrid modelling approaches. Additionally, limited generalizability for properties with complex spectral interactions, for instance Na, suggests opportunities to enhance data diversity and include environmental covariates to improve model robustness. Future research should focus on expanding the South African soil spectral datasets to include diverse soil types, employing hybrid models to combine the strengths of multiple algorithms, and utilizing advanced hyperparameter optimization techniques. The integration of environmental factors, such as soil texture and climate variables, and the application of transfer learning could further enhance the applicability and efficiency of CNN in real-world scenarios. This study contributes to a transformative tool to the field of soil spectroscopy in South Africa and builds on the soil spectral inference development research of previous studies.

4.4.6 Future Work and Potential Improvements

To address the limitations and challenges identified in this study, several areas of future research and model enhancement are proposed. Expanding the dataset to include a broader range of soil samples with diverse physical and chemical properties could significantly improve model generalizability across varying soil types and environments. Including region-specific datasets, particularly from underrepresented soil types, may better address variability in challenging soil properties for example Na, P, and K. Research has demonstrated that larger and more diverse datasets enhance model performance by capturing the variability in soil spectral data more comprehensively (Safanelli *et al.*, 2023). Incorporating advanced spectral pre-processing techniques, such as wavelet transformations or advanced filtering, could help reduce noise and emphasize critical spectral features. Wavelet-based approaches are particularly beneficial for high-dimensional spectral data, as they capture both local and global patterns, enhancing spectral feature extraction and improving prediction accuracy (Vašát *et al.*, 2017). These techniques could complement existing pre-processing methods, refining model inputs further and improving overall model performance. Hybrid modelling approaches that combine CNN with other machine learning techniques, such as Random Forests, Gradient Boosting, or Cubist models, could leverage the strengths of multiple methods. Hybrid models have been shown to improve predictions by integrating the feature extraction capabilities of CNN with the interpretability and robustness of tree-based algorithms, addressing the complexities of soil property interactions (Padarian *et al.*, 2020). These approaches may enhance the prediction of more challenging soil properties by capturing both linear and non-linear relationships in the spectral data. Advanced hyperparameter optimization methods, such as Bayesian optimization or genetic algorithms, could be employed to refine CNN architectures further. These techniques efficiently identify optimal configurations for

complex models, improving performance for soil properties with high spectral variability, for instance Na, P, and K. Optimized hyperparameters could reduce overfitting while improving the generalizability of CNN across different datasets (Yang *et al.*, 2020). The integration of contextual data, such as soil texture, topography, and climatic variables, into the modelling process may improve predictions by accounting for external factors that influence soil properties. Combining spectral data with environmental covariates has been shown to provide a more comprehensive understanding of soil-climate interactions, enabling models to capture complex dependencies more effectively (Stenberg et al., 2010). Leveraging transfer learning techniques, where pre-trained models are fine-tuned on region-specific datasets, could enhance performance in data-limited contexts. Transfer learning allows models to benefit from previously learned spectral patterns, reducing the need for extensive training on new datasets. Additionally, multi-task learning approaches could be explored to predict multiple soil properties simultaneously, capturing interdependencies while reducing computational costs. These approaches have demonstrated their value in enhancing prediction accuracy while improving model efficiency (Ng *et al.*, 2019). By addressing these areas, future studies could significantly enhance the performance, scalability, and practical applicability of CNN models in soil spectroscopy. These improvements would advance the use of CNN in precision agriculture, supporting sustainable soil management practices and contributing to global efforts in agricultural innovation.

CHAPTER 5 SOIL SPECTRAL INFERENCE IN THE WESTERN HIGHVELD, SOUTH AFRICA: EVALUATING SPIKING AS A TOOL FOR IMPROVED MODELING ACCURACY

5.1 Abstract

Global soil spectral libraries, for instance the Open Soil Spectral Library (OSSL), hold significant potential for improving soil property predictions essential to precision agriculture and sustainable land management. However, these libraries often lack sufficient data from underrepresented regions like South Africa, leading to inaccuracies in global predictive models. This study addressed the critical need to improve soil property predictions in the Western Highveld region by employing a "spiking" technique—the strategic integration of local soil spectral data into global datasets. Using mid-infrared (MIR) spectroscopy, we collected soil spectral and property data from the Western Highveld and merged it with the OSSL dataset. Through bootstrapping and machine learning algorithms (Random Forest, Elastic Net, and Cubist), we analysed the impact of varying spiking levels on the accuracy of soil property predictions for calcium (Ca), potassium (K), magnesium (Mg), sodium (Na), and phosphorus (P). The results reveal that spiking significantly enhances prediction accuracy for most soil properties, with a one-fold spiking level yielding the most substantial improvements compared to the OSSL models with no spiking. However, excessive spiking demonstrated diminishing returns and, in some cases, reduced performance due to a combination of overfitting and increased heterogeneity. Locally calibrated models consistently outperformed spiked models, and global spiking did not improve local prediction models, emphasizing the importance of regional specificity. This study contributes to soil spectroscopy by providing a scalable strategy to address data gaps in global spectral libraries while underscoring the limitations of over-reliance on global datasets. The findings offer practical insights for optimizing spectral libraries to support precision agriculture and sustainable land management in data-deficient regions, particularly in contexts of South Africa.

5.2 Introduction

Global challenges such as food security and climate change demand increasingly sustainable precision agricultural and land management practices (Zhang *et al.*, 2021). These practices rely heavily on accurate and efficient soil analysis (Farooqi *et al.*, 2021). Accurate soil property predictions are essential for applications spanning agriculture, forestry, and land use management (Ng et al., 2022a; Vona et al., 2022). This need has driven the development of rapid and non-destructive techniques like mid-infrared (MIR) spectroscopy, which aid in assessing soil

composition and soil-health (Rossel and McBratney, 2008; Wadoux et al., 2021). Advances in remote sensing and machine learning further enhance the power of these techniques.

Large spectral libraries for example the Open Soil Spectral Library (OSSL) created by Safanelli et al. (2023) hold immense potential for soil analysis, but they often suffer from a critical limitation: a lack of comprehensive data from underrepresented regions like South Africa (Kock et al., 2024). This lack of representation can lead to significant inaccuracies when using soil property prediction models created by the OSSL for regions such as the Western Highveld region of South Africa (Kock et al., 2024). Soil property data inaccuracies hinder effective land management decisions and can have negative consequences for example, agricultural productivity and environmental conservation (Fuentes *et al.*, 2021).

To address the limitation of underrepresented regions in global spectral libraries, this study employs a 'spiking' technique, strategically integrating local soil spectral data from the Western Highveld region into the OSSL. By leveraging existing local datasets, we aim to refine the OSSL's predictive models and improve the accuracy of soil property assessments in this region. The primary research question is: "How can the integration of local soil spectral data into the OSSL improve the accuracy of global soil property prediction models and vice versa?" To answer this, we employed a mixed-methods approach involving soil spectral library augmentation, spiking, machine learning, and model validation.

This study made several contributions to the field of soil spectroscopy and its application in sustainable land management. First, it addressed the critical gap in existing global spectral libraries by systematically incorporating local soil and spectral data from the Western Highveld region, South Africa. This approach offers a novel perspective on improving the accuracy and regional relevance of global soil property prediction models, particularly in areas with unique soil characteristics for example the Western Highveld region. Secondly, by employing a 'spiking' technique, this research provided a practical and efficient strategy for enhancing existing spectral libraries without the need to build entirely new databases, which can be resource intensive. Finally, the study utilized advanced machine learning techniques used to create global soil property estimation models to develop and validate improved predictive models, contributing to the growing body of knowledge on the application of soil spectroscopy in the field of soil science. Then, by addressing the limitations of current global models and demonstrating the value of incorporating localized data, this study provides a crucial step towards more accurate and regionally relevant soil property predictions. This supports more effective and sustainable land management practices and the adoption of precision agricultural practices (Ng et al., 2022a).

5.3 Materials and Methods

5.3.1 Data Acquisition

A soil spectral library containing samples from the Western Highveld region of South Africa as well as the accompanying MIR spectral data for each soil sample used by Kock *et al*. (2024) was also used for this study. The MIR spectral measurements for the samples were obtained using Bruker ALPHA II with DRIFTS module with a spectral range of 400 - 4 000 cm^{-1}. The corresponding soil properties pH (KCl), P (Bray-1), and extractable K, Ca, Mg, Na. Spectral and soil property data were also downloaded from the publicly available Open Soil Spectral Library (OSSL) version 1.2 (Safanelli *et al.*, 2023; Sanderman *et al.*, 2023). The MIR spectral data (ossl_mir_L0_v1.2) and the corresponding soil laboratory data (ossl_soillab_L1_v1.2) were downloaded and merged.

5.3.2 Data Pre-processing

5.3.2.1 Local Data

The raw spectral local data was first cleaned by checking for and removing potential spectral outliers that were identified using Principal Component Analysis (PCA) and Mahalanobis distance (Wadoux *et al.*, 2021). Spectra with a Mahalanobis distance greater than 3 from the centre of the PCA scores were considered outliers and subsequently removed from the dataset (Wadoux *et al.*, 2021). To ensure that the distribution of data was not impacted, the Shapiro-Wilk test was run to check for normality (Safanelli *et al.*, 2023). Finally, column names were standardized to match that of the OSSL data format which require spectral MIR data to be expressed in wavenumbers and in absorption units.

5.3.2.2 OSSL Data

Because the vast OSSL spectral library contains thousands of soil sample data of which it also contains soil property data that is of no use in this study, the OSSL MIR spectral and soil attribute data was filtered to retain only the wavenumbers that matched the range of the local western Highveld spectral data and soil properties that will be covered in this study. Relevant soil properties (P, Ca, Mg, K, Na) were extracted from the OSSL soil laboratory data, while for the Western Highveld dataset pH(KCl) data was omitted, as this pH extraction method is not contained in the OSSL dataset, to retain consistency the extractable cations –(Ca, Mg, K and Na)- were converted from cmolc kg^{-1} -which is used in the OSSL dataset- to mg kg^{-1} to mitigate confusion when interpreting results and directly comparing them to that of previous studies by Kock *et al*., (2024).

5.3.2.3 Combined Data

Both the local and OSSL spectral data were pre-processed using Savitzky-Golay smoothing with a second-order polynomial, a window size of 11, and a first derivative. This step helps to reduce noise and enhance spectral features (Safanelli et al., 2023; Savitzky and Golay, 1964; Wadoux et al., 2021).

5.3.3 Dataset Preparation

The local Western Highveld dataset was split into a training set –(681 samples)- for spiking and a separate independent validation set using k-means clustering and soil properties as variables to ensure representative sampling (Metwally *et al.*, 2019). The spiking datasets were constructed to varying sizes – one-fold (x1) (training set used in local models), two-fold (x2), three-fold (x3), five-fold (x5), ten-fold (x10), and two-hundred-fold (x200) of the original local training set. A bootstrapping approach was used, where samples were randomly selected with replacement from the original training set (Ukil *et al.*, 2010). Bootstrapping was used to increase the size of the training datasets while preserving the statistical characteristics of the original data. This method allows for the construction of larger datasets without the need to collect additional samples, making it especially valuable when resources are limited (Ukil *et al.*, 2010). Additionally, bootstrapping introduces variability, which can help models generalize better by exposing them to multiple instances of similar data (Ukil *et al.*, 2010). This ensures that the constructed datasets remain representative of the local soil spectral variability, critical for evaluating the effectiveness of spiking in improving model performance. Unique identifiers were assigned to the synthetic samples generated through bootstrapping and the augmented local datasets were then merged with the corresponding OSSL datasets for each soil property (P, Ca, Mg, K, Na) creating spiked datasets with increasing proportions of local data.

5.3.4 Spectral Data Analysis

The same pre-processing procedures were used as outlined by the OSSL manual, ensuring consistency and comparability. The OSSL operating procedure uses PCA to reduce the dimensionality of the spectral data and capture the most important spectral variations (Safanelli *et al.*, 2023). A cumulative explained variance threshold of 99% was used to select the principal components (Safanelli *et al.*, 2023); PCA score plots were generated to visualize the distribution of the spiked and validation samples in the reduced dimensional space (Safanelli *et al.*, 2023); PCA was applied to all spiked datasets.

5.3.4.1 Model Calibration and Validation

The PCA-transformed datasets for each soil property (P, Ca, Mg, K, Na) and each spiking level - x1, x2, x5, x10, and x200 - were loaded into R using with Rstudio 2023.06.1 Build version 524 (R Core Team, 2023). The target variables (soil properties) were first checked for normality and subsequently log-transformed using $log1p\,()$ to improve model performance (Safanelli *et al.*, 2023). A dummy model was trained for each soil property as a baseline for comparison (Safanelli *et al.*, 2023). Three machine learning algorithms - Random Forest, Elastic Net, and Cubist - were benchmarked using 10-fold cross-validation (Safanelli *et al.*, 2023). The best-performing model algorithm was then selected for subsequent calibration processes. A random search was employed to tune the hyperparameters of the Cubist model, while default search spaces were used for Random Forest and Elastic Net (Safanelli *et al.*, 2023). The best-performing Cubist model was selected for each soil property and spiking level. These models were then trained on the entire spiked training datasets and evaluated on the held-out validation data using performance metrics which, include: RMSE, bias, and R^2 (Safanelli *et al.*, 2023). Additional accuracy metrics (CCC and RPIQ) were also calculated (Safanelli *et al.*, 2023)

5.3.4.2 Outlier Detection and Visualization

Observed versus predicted scatter plots were generated for each soil property and spiking level to visually evaluate the models' predictive performance. These plots included a 1:1 reference line to highlight deviations between observed and predicted values. To further identify potential outliers that may indicate overfitting or anomalous predictions, Z-scores were calculated for both the predicted values in Equation 5-1 and Equation 5-2:

Equation 5-1

$$Z = x - \mu\sigma$$

Equation 5-2

$$Z = \sigma x - \mu$$

Where x represents the data point, μ is the mean of the dataset, and σ is the standard deviation. A Z-threshold of 2 (two standard deviations) was applied to flag extreme outliers. Points with absolute Z-scores greater than 2 were highlighted in the scatter plots. These flagged points represent values that deviated significantly from the distribution and were used to assess potential overfitting in the models (Chandola *et al.*, 2009).

5.3.4.3 Benchmarking Against Global Standards

The methods used in this study, including spectral pre-processing and model creation and validation steps, were obtained from the provided manual of the OSSL GitHub page (Safanelli *et al.*, 2023). This approach ensured consistency with the global soil property inference models and allowed for benchmarking against existing standards.

5.4 Results and Discussion

5.4.1 Global Prediction Model Improvement

In this study, we evaluated the impact of spiking the OSSL dataset with varying amounts of local spectral data from the Western Highveld region of South Africa on the accuracy of predicting key soil properties routinely used for precision agriculture in South Africa. Our results showed that spiking the OSSL with any local data significantly improved model performance compared to the OSSL estimation models -OSSL Only- calibrated with using the OSSL datas*et al*one. For instance, at the one-fold spiking level, the RMSE for Ca decreased dramatically from 1128.45 mg kg^{-1} in the OSSL-only model to 46.09 mg kg^{-1}, and the LCCC increased from 0.04 to 0.70 (see Table 5-1). Similar trends were observed for other soil properties, RMSE decreased for K from 213.28 to 52.85 mg kg^{-1}, for Mg from 206.92 to 45,12 mg kg^{-1} and for Na from 59.55 to 42.63 mg kg^{-1}. For P on the other hand the RMSE increased from 25.23 to 64.82 mg kg^{-1}. Apart from P, it was indicated that even a small addition of local samples can substantially enhance the predictive capability of global models which is consistent with the findings from Barthès et al., 2020; Guerrero et al., 2010; St. Luce et al., 2022; Ng et al., 2022a; Vona et al., 2022; Wetterlind and Stenberg 2010). This improvement aligns with the findings of Barthès *et al*. (2020), who reported that representative spiking improved the quantification of other soil properties like soil in SOC using MIR spectroscopy prediction models. They found that incorporating a modest number of local samples into a global spectral library significantly enhanced model accuracy, particularly for soil properties that are highly variable and region-specific.

Table 5-1: Model performance metrics (RMSE, ME, R^2, LCCC, and RPIQ) for soil properties (Ca, K, Mg, Na, and P) across spiking levels (x1, x2, x5, x10, and x200) using local models and the global model (OSSL ONLY).

Soil Properties	RMSE (mg kg^{-1})	ME	R^2	LCCC	RPIQ
OSSL ONLY (Kock *et al*. 2024)					
Ca	1128.45	- 868.07	0.04	0.04	0.62
K	213.28	- 187.05	0.01	0.02	0.57
Mg	206.92	33.85	0.01	0.01	0.14
Na	59.55	- 27	0	0.01	0.27
P	25.23	19.88	0	0.01	0.25
Local Models (Kock *et al*. 2024)					
Ca	87.23	1.09	0.84	0.91	3.4
K	55.23	- 5.06	0.37	0.53	1.74
Mg	34.98	- 3.01	0.71	0.82	1.89
Na	7.48	- 1.3	0.08	0.16	1.04
P	15.69	2.87	0.55	0.66	1.26
X1					
Ca	46.09	-2.59	0.51	0.70	1.73
K	52.85	-4.79	0.24	0.48	1.41
Mg	45.12	9.29	0.46	0.67	1.54
Na	42.63	0.03	0.14	0.37	1.47
P	64.82	-4.58	0.25	0.45	1.14
X2					
Ca	46.70	-7.05	0.50	0.69	1.71
K	47.90	-3.43	0.29	0.51	1.56
Mg	50.00	6.81	0.40	0.62	1.39
Na	69.30	-10.30	0.01	0.09	0.907
P	68.20	-4.06	0.19	0.39	1.08
X5					
Ca	46.50	-5.46	0.46	0.67	1.71
K	53.10	-5.36	0.23	0.47	1.40
Mg	51.70	3.87	0.37	0.60	1.35
Na	110.00	-28.80	0.01	0.06	0.57
P	74.90	-10.70	0.11	0.30	0.98
X10					
Ca	55.90	-5.74	0.37	0.59	1.42
K	72.00	-4.84	0.12	0.33	1.04
Mg	54.60	8.03	0.31	0.55	1.27
Na	68.60	-13.70	0.07	0.22	0.916
P	78.20	-4.09	0.12	0.34	0.943
X200					
Ca	53.30	-4.76	0.38	0.61	1.49
K	61.00	-0.58	0.09	0.29	1.22
Mg	58.70	8.37	0.29	0.53	1.19
Na	71.30	-7.75	0.02	0.11	0.882
P	73.90	-8.15	0.18	0.41	0.997

(RMSE); root mean squared error, (ME); mean error, (R^2); coefficient of determination, (LCCC); Lin's concordance correlation coefficient and (RPIQ); the ratio of performance to interquartile distance.

5.4.2 Comparison of Spiking Models to Local Prediction Models

When comparing the prediction results of the spiked OSSL models to that of locally calibrated models, the locally calibrated models vastly outperformed any spiked prediction model across all statistical metrics. The local models consistently achieved superior R^2, LCCC and RPIQ values, reflecting their suitability to make better region-specific soil property predictions. For example, locally calibrated models for Ca had R^2 values of 0.84, compared to 0.04 for the OSSL-only model and 0.51 in the one-fold spiked model which was the best performing of all the spiked models (Table 5-1). Similarly, the LCCC for Ca local model was 0.91 compared to 0.69 for the one-fold spiked model and an even lower 0.04 for the OSSL-only model. These findings indicate the critical advantage of locally calibrated models in terms of predictive accuracy for Ca. Potassium followed a similar trend, with local K predicting models having $R^2 = 0.37$, compared to the OSSL-only model with an $R^2 = 0.01$ and the one-fold spiked model with $R^2 = 0.24$. While Mg demonstrated notable improvements with spiking, the local model remained superior, with an R^2 of 0.71, compared to 0.29 for the one-fold spiked model and LCCC of 0.82 for the local model, compared to 0.37 for the one-fold spiked model and 0.01 for the OSSL-only model. The Na local model -while not regarded as accurate- also outperformed the OSSL and spiked models with an $R^2 = 0.08$ and the OSSL-only model with $R^2 = 0.01$. The one-fold spiked model demonstrated significant improvement, with an $R^2 = 0.06$, but it still fell short of the performance achieved by the local model. The locally calibrated model for pP, with an LCCC of 0.66 and R^2 of 0.55, significantly outperformed the OSSL-only model (LCCC = 0.01, $R^2 = 0$) and the best-performing spiked model at one-fold (LCCC = 0.45, $R^2 = 0.25$), highlighting the critical importance of localized calibration for accurate predictions.

5.4.3 Various Spiking Level Performances

Interestingly, our study found that increasing the spiking levels beyond one-fold did not lead to any proportional improvements in model accuracy. In some cases, greater spiking levels resulted in decreased performance as can be seen in Figure 5-1. For example, for Mg, the RMSE increased from 45.12 mg kg^{-1} at one-fold spiking to 58.70 mg kg^{-1} at two-hundred-fold spiking, and the LCCC decreased from 0.67 to 0.53. This suggests the presence of diminishing returns with excessive spiking. Guerrero *et al*. (2010) observed a similar phenomenon and noted that while spiking regional models with samples from target sites effectively improved prediction accuracy. They found that increasing the model size beyond a certain point does not improve predictions, as additional local samples cease to provide further benefits.They emphasized the importance of selecting representative local samples to maximize the benefits of spiking. Using smaller spiking dataset sizes was outside the scope of this study, and we could not observe this

trend, as our main priority was to evaluate a very large soil spectral library. It is therefore, recommended that future research focus on using smaller datasets to evaluate the impact on the performance of global prediction models.

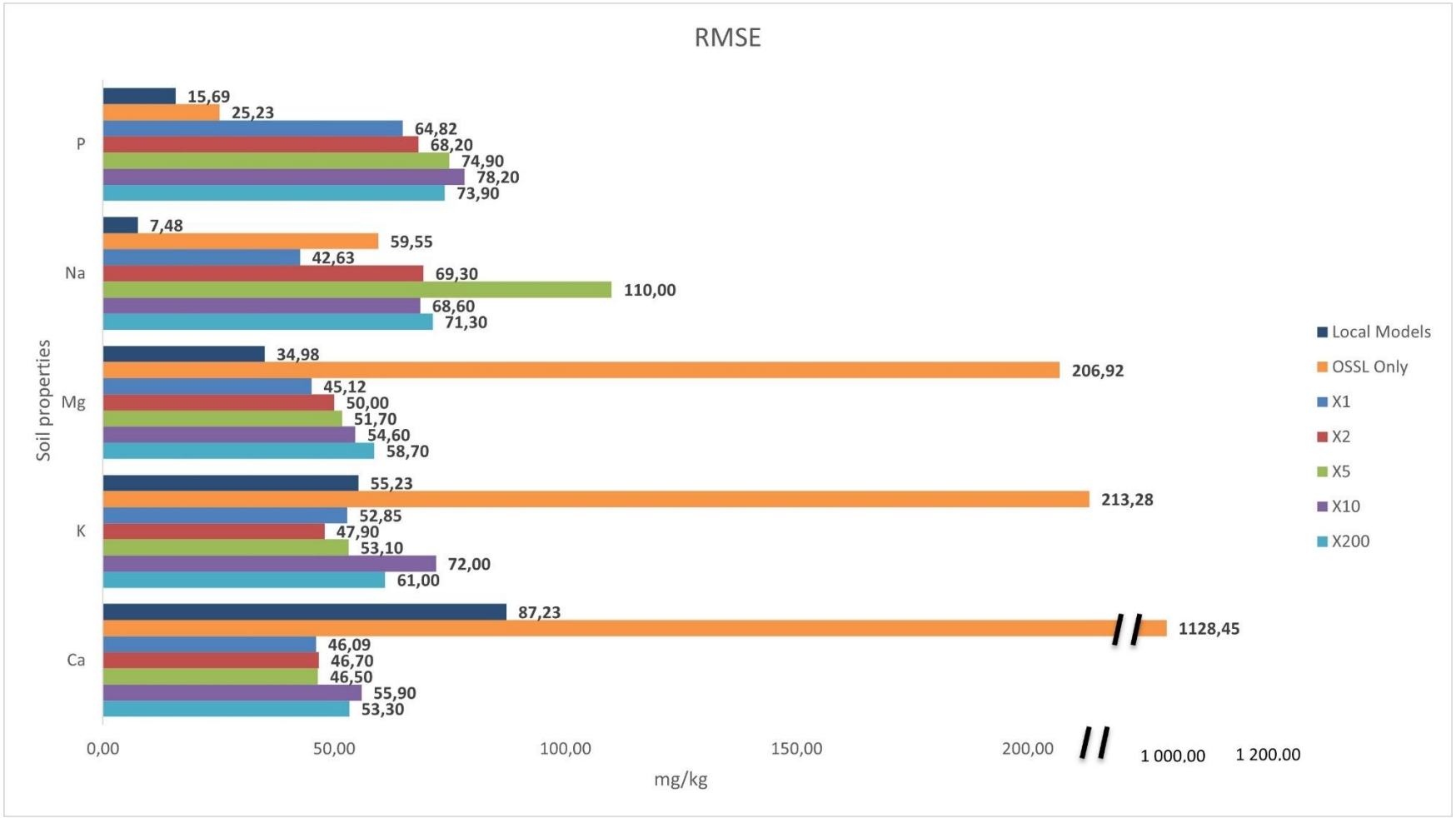

Figure 5-1: Comparison of Root Mean Square Error (RMSE) values for soil property predictions (Ca, K, Mg, Na, and P) across different models. The models include the OSSL-only global model, local models based on South African data, and spiked models at varying levels (X1, X2, X5, X10, X200). The results indicate the significant improvement of local and spiked models over the OSSL-only model, with the local model generally outperforming spiked models. However, diminishing returns and in some cases, reduced performance, are observed at higher spiking levels. RMSE values are reported in mg kg^{-1}.

5.4.4 Balance Between Local and Global Data

The observed decline in model performance at higher spiking levels may stem from the overrepresentation of local data, which can introduce a bias that reduces the model's generalizability. This shows the need to balance the precision provided by localized data with the versatility of broader and diverse datasets like that of the OSSL. Wetterlind and Stenberg (2010) compared small, local calibrations with national libraries enhanced by spiking with local samples. They found that while localized calibrations offered high predictive accuracy for specific regions, spiking national libraries with a carefully limited number of local samples could achieve similar results while preserving the model's applicability across broader contexts (Wetterlind and Stenberg 2010). Our findings align with these observations, suggesting that a careful balance between incorporating sufficient local data to capture regional variability and maintaining the heterogeneity of global datasets is essential for model robustness. This balance ensures that models remain accurate for local applications without compromising their effectiveness across diverse soil types and geographical areas. Ng *et al*. (2022) also examined strategies to enhance the prediction of local soil properties using regional spectral libraries and concluded that spiking with an optimal number of local samples boosted model performance without diminishing generalizability (Ng et al., 2022a). Further research supports this balanced approach, indicating that adding a moderate number of local samples can improve accuracy by capturing regional soil variability, especially for soil properties like SOC and pH, which are highly sensitive to local environmental conditions (Barthès *et al*., 202). However, excessive spiking, especially beyond an optimal threshold, can lead to model overfitting, where the model is finely tuned to the local dataset but loses the ability to perform well on other soils. This overfitting risk shows the importance of selecting an appropriate spiking strategy that integrates diverse data sources to capture both specific and general soil characteristics (Hong *et al*., 2018). Thus, our study highlights that while local spiking is invaluable for refining soil property predictions in specific regions, an optimal balance between local and global data is necessary to develop robust models that can serve a wide range of applications without compromising predictive accuracy.

5.4.5 Spiking Effectiveness on Different Soil Properties

The impact of spiking varied significantly across the soil properties studied, as evidenced by the scatterplots for different spiking levels and soil properties (Figure 5-2-Figure 5-6). For Ca and K, lower spiking levels (one-fold and two-fold) demonstrated substantial improvements in model performance, narrowing the gap between spiked models and local models. As seen in Figure 5-2 and Figure 5-3, the predicted values for Ca and K at one-fold and two-fold spiking levels align more closely with the observed values, with a noticeable reduction in prediction outliers. The

improved clustering around the 1:1 line suggests that these properties are relatively straightforward to predict with moderate levels of local data enrichment. This may be due to their dependence on soil parent material and climatic factors, which are partially captured by global datasets (Vona *et al.*, 2022; Ng et al., 2022a). However, at higher spiking levels (e.g., ten-fold and two-hundred-fold), the scatterplots reveal an increasing number of extreme prediction outliers, indicating potential overfitting or redundancy in the spiked data. Sodium and P exhibited less improvement with spiking, even at higher levels. As shown in Figure 5-5 and Figure 5-6, the predicted values for Na and P remain widely scattered from the observed values, with many points deviating significantly from the 1:1 line. This underperformance might be from the unique distribution and behaviour of these elements in soils, which are highly susceptible to localized factors such as irrigation practices, salinity, and historical fertilizer use (Peng *et al*., 2013; Wetterlind and Stenberg 2010). For Na, even at the one-fold spiking level, there are substantial deviations, and as spiking increases to ten-fold and two-hundred-fold, with the number of extreme outliers grows considerably. Similarly, for P, the spread of predicted values remains large, with multiple outliers evident across all spiking levels. These results suggest that Na and P require more intensive local data collection to capture their complex spatial patterns and achieve accurate predictions. Magnesium displayed an intermediate response to spiking. As seen in Figure 5-4, lower spiking levels (one-fold and two-fold) showed notable improvement in the alignment between predicted and observed values, with relatively few outliers. However, at higher spiking levels (e.g., two-hundred-fold), the scatterplots reveal an increasing number of extreme predictions, suggesting diminishing returns from excessive spiking. This trend aligns with previous research indicating that the inclusion of redundant data can lead to model overfitting and reduced generalizability (Vona *et al.*, 2022).

The presence of prediction outliers across all soil properties, particularly at higher spiking levels, underscores the need for a strategic and property-specific approach to spiking. For properties Ca and K, lower levels of spiking appear sufficient to improve model accuracy, while for Na and P, more extensive local data collection is necessary to address their unique variability and ensure reliable predictions (Barthès *et al.*, 2020). Additionally, as highlighted in the scatterplots, the increased presence of outliers at higher spiking levels suggests that excessive spiking can introduce noise and reduce model robustness.

In summary, the effectiveness of spiking is highly dependent on the soil property in question. Moderate spiking levels (one-fold and two-fold) are effective for properties like Ca and K, but achieving comparable accuracy for Na and P requires more localized data and tailored strategies. These findings emphasize the importance of balancing local and global data contributions to

optimize predictive models and minimize outliers, aligning with adaptive strategies proposed in previous research.

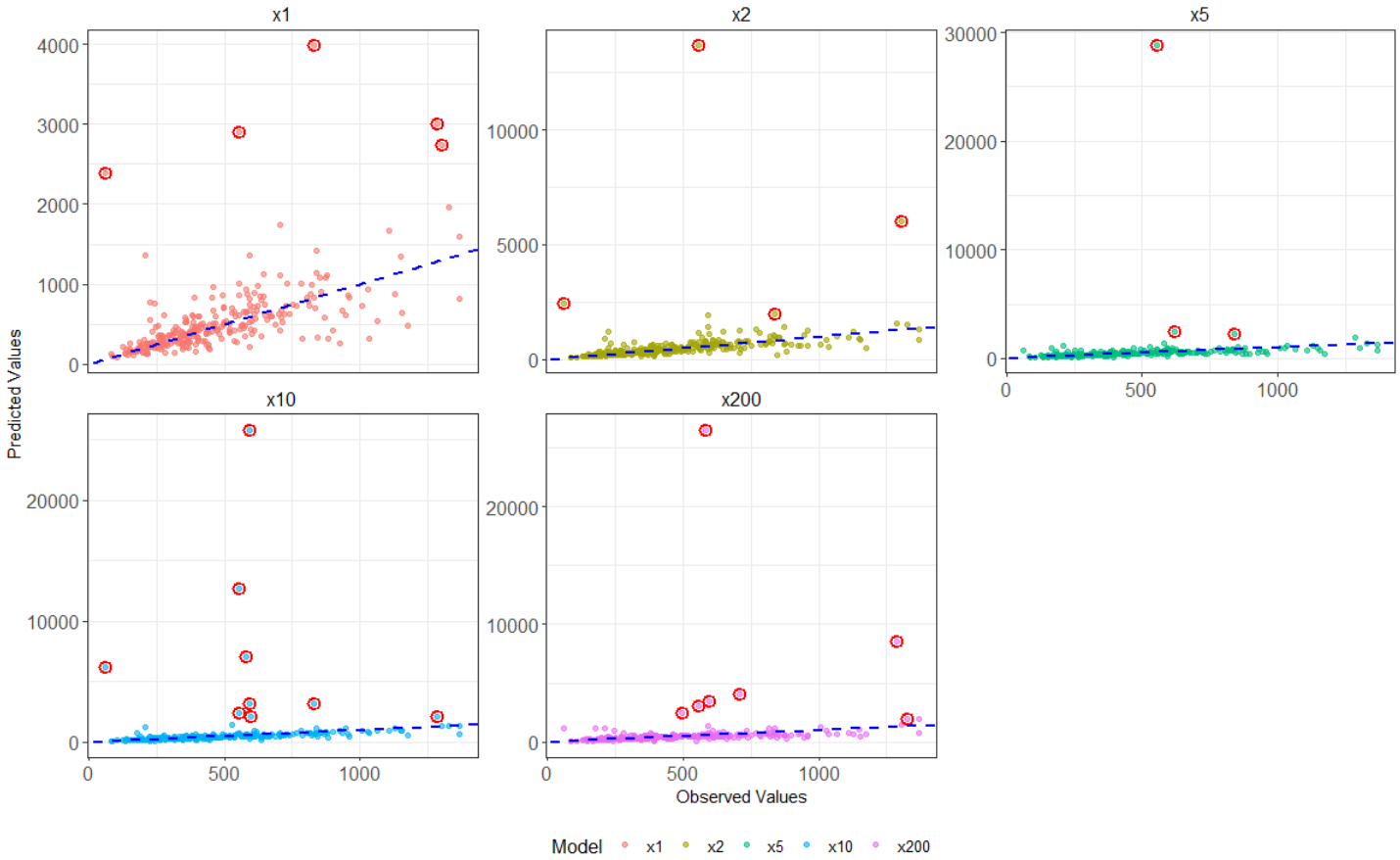

Figure 5-2: Scatter plots of observed versus predicted values for Ca across five spiking levels (x1, x2, x5, x10, and x200). Each point represents a data sample, with the 1:1 dashed blue line indicating perfect predictions. Outliers identified using Z-scores (|Z| > 2) are highlighted in red to illustrate extreme deviations and potential overfitting in the models.

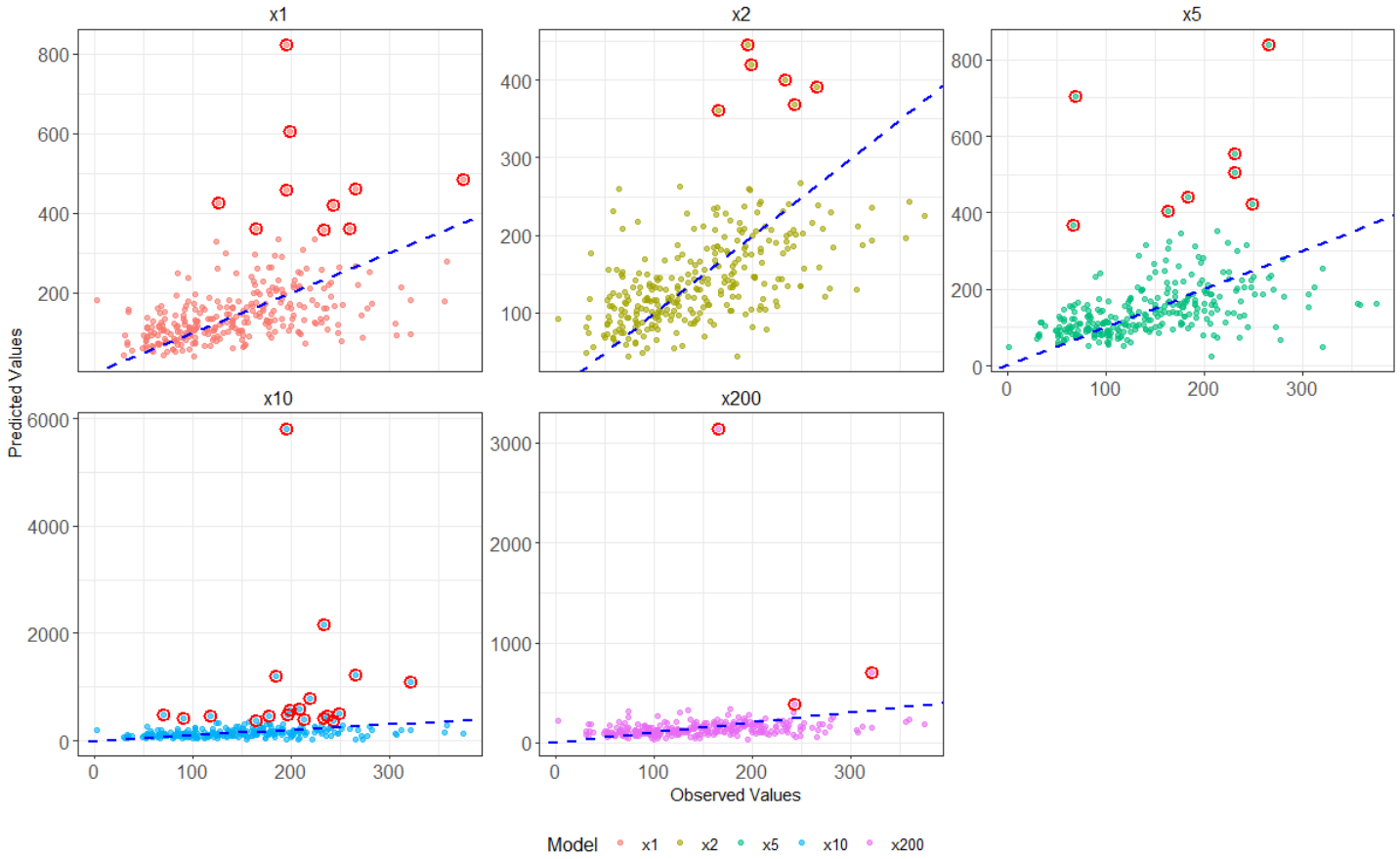

Figure 5-3: Scatter plots of observed versus predicted values for K across five spiking levels (x1, x2, x5, x10, and x200). Each point represents a data sample, with the 1:1 dashed blue line indicating perfect predictions. Outliers identified using Z-scores (|Z| > 2) are highlighted in red to illustrate extreme deviations and potential overfitting in the models.

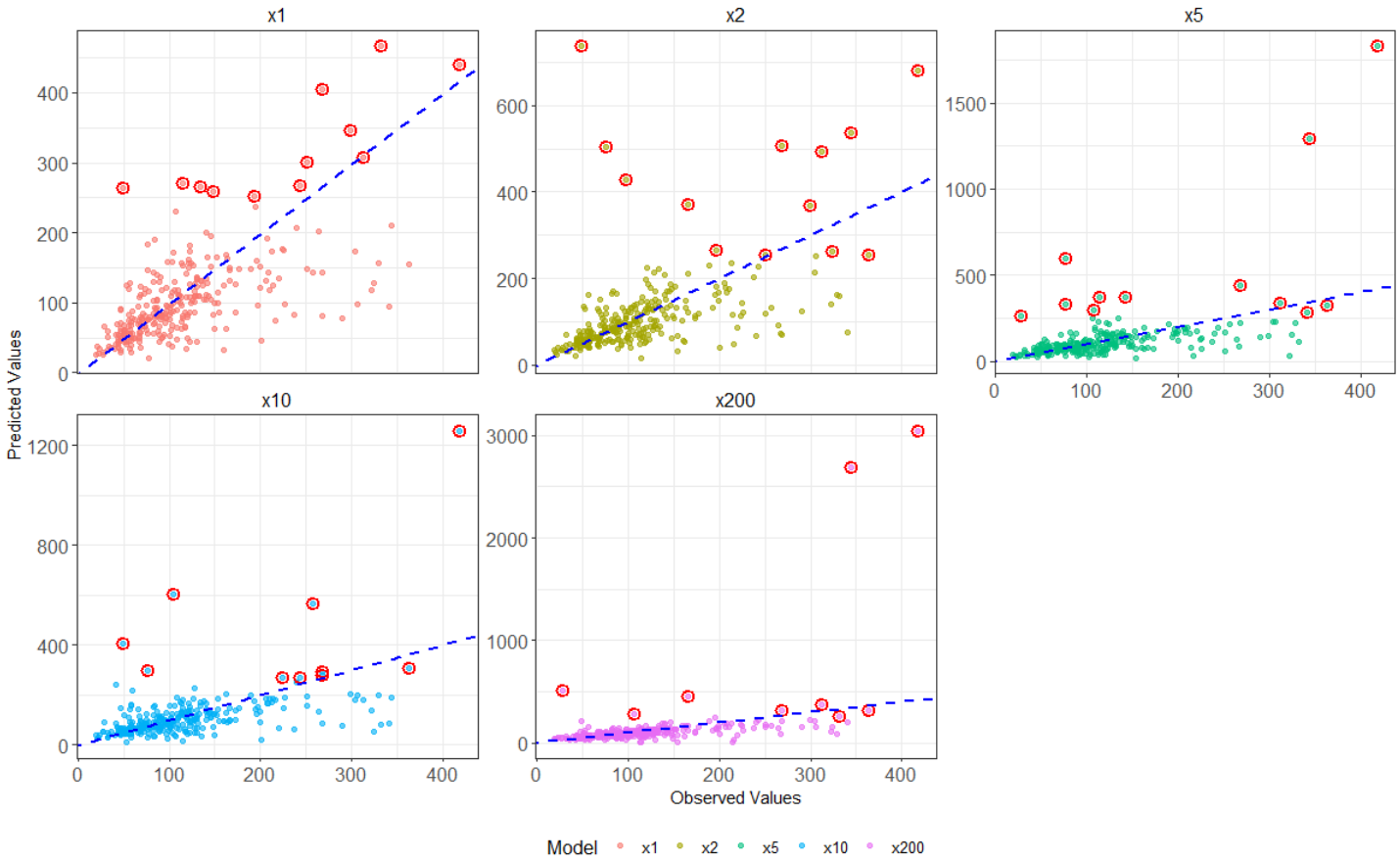

Figure 5-4: Scatter plots of observed versus predicted values for Mg across five spiking levels (x1, x2, x5, x10, and x200). Each point represents a data sample, with the 1:1 dashed blue line indicating perfect predictions. Outliers identified using Z-scores (|Z| > 2) are highlighted in red to illustrate extreme deviations and potential overfitting in the models.

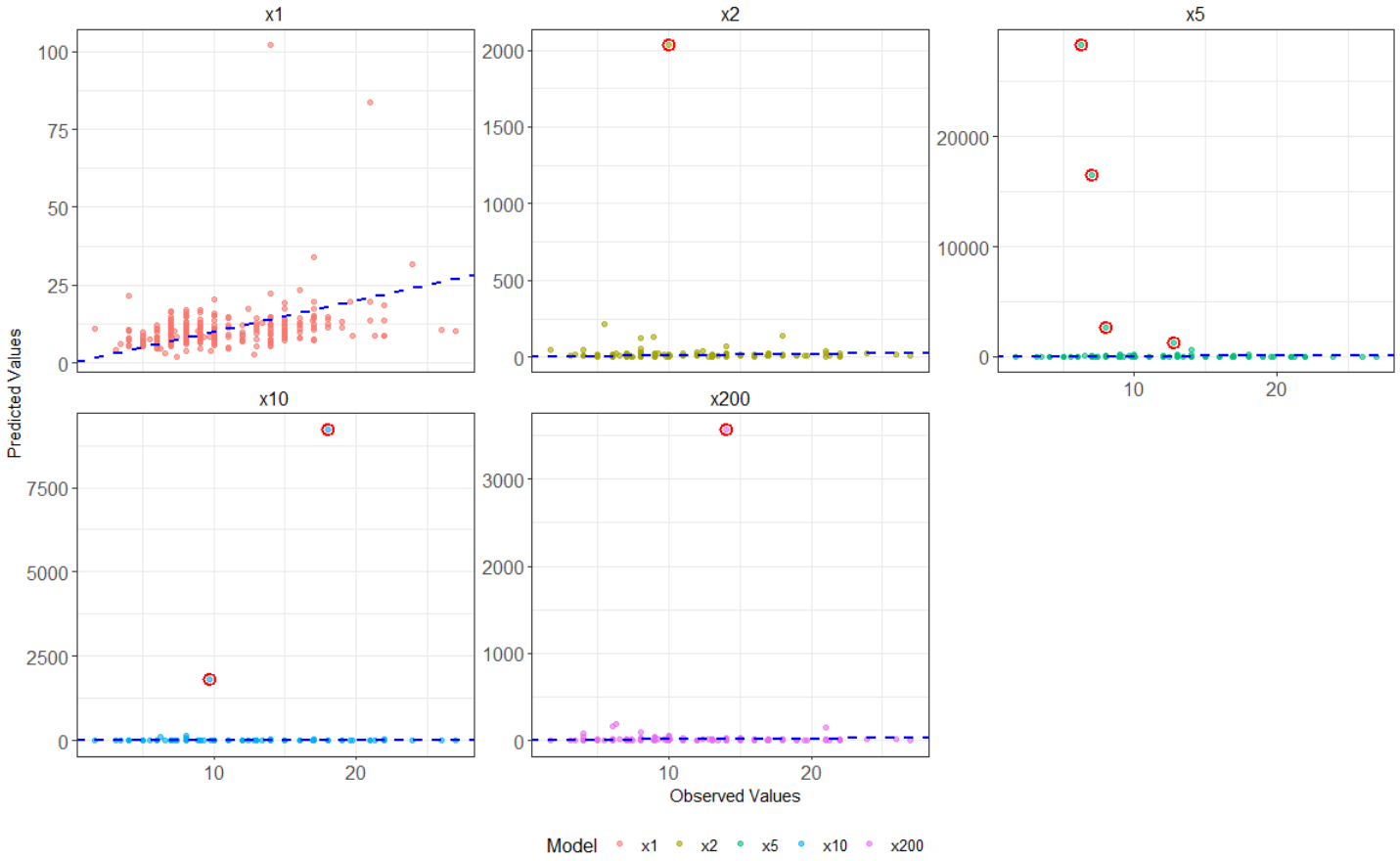

Figure 5-5: Scatter plots of observed versus predicted values for Na across five spiking levels (x1, x2, x5, x10, and x200). Each point represents a data sample, with the 1:1 dashed blue line indicating perfect predictions. Outliers identified using Z-scores (|Z| > 2) are highlighted in red to illustrate extreme deviations and potential overfitting in the models.

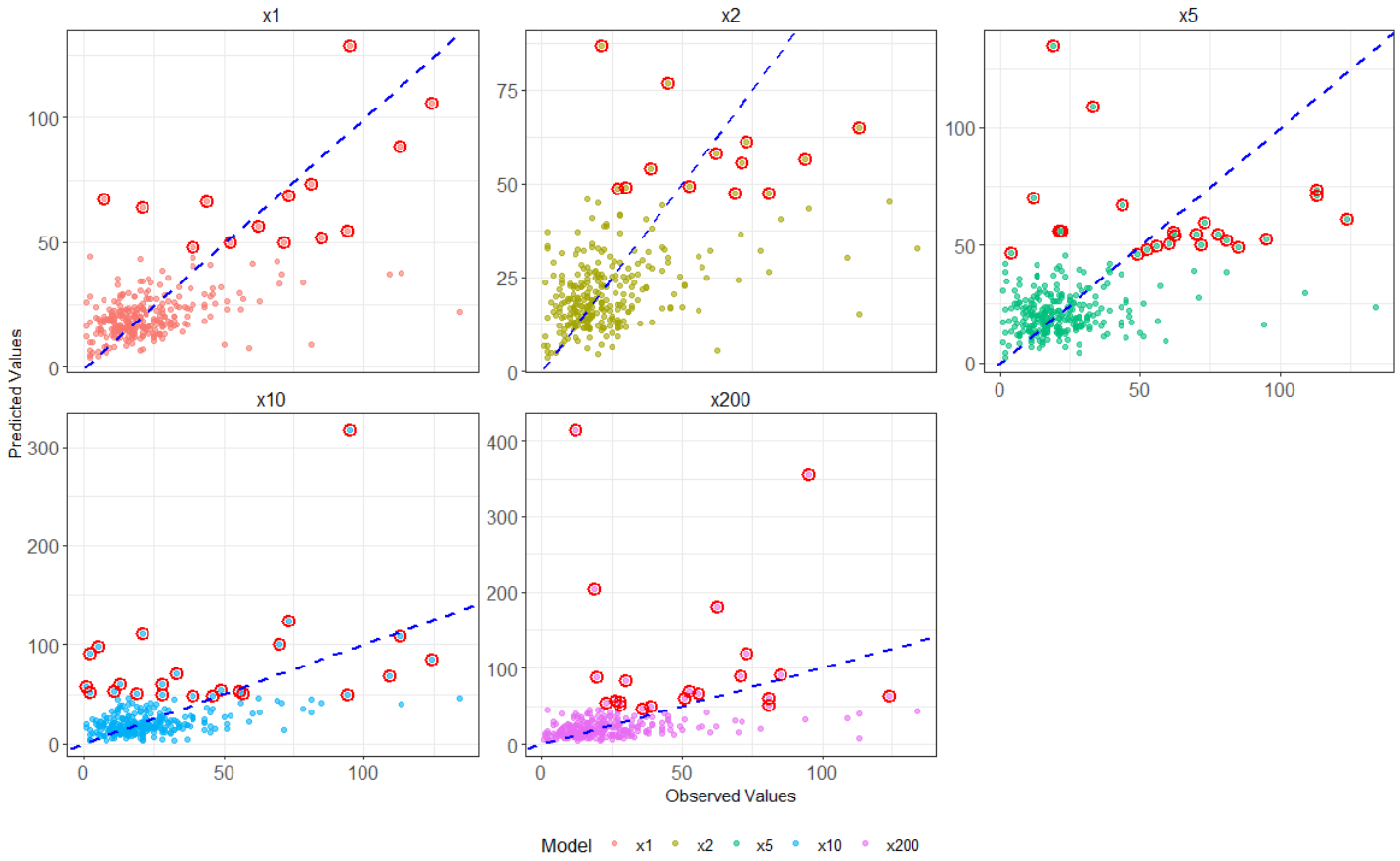

Figure 5-6: Scatter plots of observed versus predicted values for P across five spiking levels (x1, x2, x5, x10, and x200). Each point represents a data sample, with the 1:1 dashed blue line indicating perfect predictions. Outliers identified using Z-scores (|Z| > 2) are highlighted in red to illustrate extreme deviations and potential overfitting in the models.

5.4.6 Implications for Spectral Library Development of South Africa

The findings of our study have significant implications for the development and utilization of global spectral libraries, such as the OSSL, within the South African context. South Africa's diverse soil types and unique environmental conditions, especially in underrepresented regions like the Western Highveld, necessitate the inclusion of localized data to enhance the accuracy of soil property prediction models. While incorporating local samples through spiking has proven effective in improving the accuracy of global models, making them more relevant for precision agriculture in South Africa, they still fall short of the accuracy achieved by locally calibrated models. However, this approach can be particularly beneficial because these global models include a wider range of soil properties, many of which lack corresponding local calibration models.

However, the benefits of spiking must be balanced against potential risks like overfitting, which can reduce a model's generalizability to other regions within South Africa or globally. This risk is particularly relevant in a country as ecologically diverse as South Africa, where overemphasis on data from one region could bias predictions for other areas. Strategic sample selection is crucial to enhance accuracy without introducing bias or limiting the model’s broader applicability (Barthès *et al.*, 2020; Guerrero *et al.*, 2010). These results can add to the further development of the regional Soil Spectral Library for South Africa, which shows that local models should be created first and then contributed to a larger spectral library.

Spiking strategies could also consider the spatial variability and heterogeneity of soils across different South African regions, which includes different climatic and vegetation zones. The research indicates that soil properties with high spatial variability, Na and Mg, benefited the most from localized spiking strategies, where overrepresentation of local data sometimes reduced model performance and increase overfitting (Dudek et al., 2021; Wetterlind and Stenberg 2010).

In addition to strategic spiking, contributing local soil spectral data to global repositories like the OSSL is essential. Sharing high-quality, well-annotated datasets from South Africa enhances the predictive accuracy of global models and promotes inclusivity in spectral library development, addressing representation gaps for underrepresented regions. This collaborative approach can improve the adaptability of future models to diverse ecological and agricultural contexts worldwide (Li *et al.*, 2020; Yi *et al.*, 2013).

Fostering a culture of data sharing encourages the development of next-generation predictive tools by enabling machine learning algorithms to access richer, more diverse datasets. This can lead to breakthroughs in understanding complex soil behaviours and contribute to global efforts

in precision agriculture and sustainable land management (Safanelli *et al.*, 2023). By contributing data to global repositories, South Africa can play a pivotal role in advancing soil spectroscopy research and ensuring that future models are both accurate and globally applicable.

5.5 Conclusion

This research highlights the significance of adding regional soil spectral data to global spectral libraries through a method known as spiking, aimed at addressing the lack of representation for areas like South Africa in global datasets. By carefully assessing the impact of spiking on the accuracy of soil property models, this study offers a clear strategy for enhancing the application of soil spectroscopy in precision agriculture and sustainable land management. Remarkable improvements in OSSL model precision were seen for elements Ca, K, Mg, and Na, while the difficulties in forecasting P underscore the challenges related to its spatial variability and distribution in the soils of the Western Highveld region of South Africa. The superior predictive performance for Ca, Mg, and K can be attributed to their stronger spectral features and well-defined absorption patterns in the mid-infrared (MIR) region, whereas P remains challenging due to its weak spectral response, complex interactions with soil minerals, and high spatial variability. The results make a substantial contribution to soil spectroscopy by proving that spiking is a potent means to boost the relevance and functionality of global spectral libraries. This study builds on previous research by highlighting the necessity of balancing regional specificity with global diversity in spectral datasets. It verifies that elements with significant spatial variability, like Ca and Mg, gain from the inclusion of localized data. However, it also notes diminishing advantages at increased spiking levels, underscoring the importance of careful sample selection to prevent overfitting and maintain model generalizability. Focusing on the Western Highveld region of South Africa, just one of many underrepresented areas in global spectral collections, the study tackles a crucial deficiency and sets the stage for applying these methods to other ecologically varied and/or data-deficient regions worldwide. While the study presents strong evidence supporting the advantages of spiking, it acknowledges certain limitations that need further research, for example exploring the use smaller spiking increments to determine the least number of samples needed. This study also showed that spiked global prediction models cannot replace well calibrated local models and emphasises the need for local soil spectral data sets. While these models will likely show similar outcomes to global spiking results, there is a need to investigate spiking on a regional South African spectral library using field scale spectral data. Future studies should examine the relationship between spiking levels and particular soil properties to improve guidelines for spectral library enhancement and model refinement. A takeaway from this work is the importance of adding localized soil spectral data to global repositories like the OSSL. Enhancing representation from underrepresented areas holds great potential for significantly improving global model

performance, increasing adaptability to various contexts, and promoting fair progress in soil science. Collaborative initiatives to share high-quality, well-documented datasets can fortify global spectral libraries, facilitating the creation of predictive tools that are both precise and widely applicable.

CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

This PhD study represents a comprehensive exploration into the integration of soil spectroscopy and advanced computational techniques to improve soil spectral inference predictions in South Africa, and specifically the Western Highveld region. The main aim of this study was to develop and evaluate accurate, region specific soil property prediction models for the Western Highveld region of South Africa, by leveraging the combined strengths of NIR and MIR spectroscopy and investigating different pre-processing techniques, ML and deep learning methods on model performance. The three main objectives of this thesis wereto:

1. Create and assess the effectiveness of combining different spectral regions in improving the prediction accuracy of key soil properties, compared to using single spectral regions (NIR or MIR alone).
2. Create and compare the performance of various machine-learning algorithms like ML and neural networks in predicting soil properties using the combined pre-processed NIR and MIR spectral data.
3. Investigate the effectiveness of "spiking" a global soil spectral library with soil spectral data from the Western Highveld region of South Africa in improving the prediction accuracy of local prediction and global soil property models.

Through three interconnected objectives, significant strides have been made in enhancing the accuracy, efficiency, and applicability of existing prediction models. Collectively, these contributions provide a robust foundation for advancing soil spectroscopy in underrepresented regions, such as South Africa, while addressing the need for local representation of soil spectral features in large global spectral libraries.

The first objective was met in Chapter 3. This study underscored the transformative potential of combining NIR and MIR spectral data to predict critical soil properties used in fertilizer recommendations. By leveraging the complementary strengths of these spectral regions, this research demonstrated the value of minimal pre-processing for many soil attributes, while highlighting the need for tailored adjustments to address the complexities of specific properties like Na and K. This integrated spectral approach achieved significant improvements in predictive accuracy, emphasizing its utility for South African soils and supporting the adoption of a multi-spectral input approach for soil spectral inference models.

The second objective was addressed in Chapter 4, which introduced convolutional neural networks (CNN) as an innovative solution to manage the high-dimensional complexity of soil spectral data. By tailoring CNN architectures to address the challenges posed by soil heterogeneity and spectral overlaps, this research achieved robust predictions for properties pH (KCl), Ca, and Mg. The findings validated CNN as a scalable and efficient tool for soil spectroscopy while identifying opportunities for further refinement, particularly through hybrid modelling and expanded spectral libraries. This work demonstrated the power of deep learning to unlock insights from spectral data, aligning with global efforts toward accurate and robust prediction models.

The third objective was addressed in Chapter 5, which tackled the underrepresentation of South African soils in global spectral libraries by employing spiking techniques to integrate regional data into global models. The research revealed significant improvements in predictive accuracy for key soil properties in a global dataset, demonstrating the efficacy of spiking in enhancing the relevance of global datasets for local contexts. This study also provided critical insights into the trade-offs associated with spiking, including the risks of overfitting at high levels of local data integration. These findings underscore the importance of balancing regional specificity with global diversity in spectral prediction models. However, the spiking of a global dataset with local data did not improve the local prediction models, emphasising the need for local data collection and modelling to utilize soil spectroscopy.

6.2 Contributions and Implications

Together, these objectives contribute significantly to the field of soil spectroscopy by addressing key methodological and data challenges. The integration of combined spectral data, advanced deep learning techniques, and spiked spectral libraries offers a multidimensional framework for improving soil property predictions in South Africa. These advancements have huge practical implications for precision agriculture, enabling more accurate soil assessments, informed resource management, and sustainable land use practices. Furthermore, this research highlights the necessity of developing localized models and spectral datasets to enhance global predictive tools.

South Africa has made significant progress in adopting soil spectroscopy methods, particularly MIR and NIR, for predicting soil properties. However, global soil spectral models often fall short of capturing the unique and diverse soil profiles of the region, which include significant variability in soil texture, organic matter content, mineral composition, and nutrient availability across different landscapes. Locally calibrated models have consistently outperformed global models,

particularly for properties pH, Ca, and Mg. These results highlight the critical need for regionalized approaches to soil spectral modelling tailored to South African conditions.

Currently, soil spectroscopy models in South Africa show promising reliability for properties like pH, Ca and Mg, particularly when sufficient local calibration data is available. However, challenges persist with properties such as K and Na, which are often influenced by weak spectral signals, high spatial variability, and complex interactions with other soil components, making accurate predictions more difficult. While MIR and NIR spectroscopy models are increasingly effective, their usability depends on the specific soil property being analysed and the degree of data heterogeneity. This indicates that further refinement and targeted strategies are needed to improve model performance for challenging properties.

Moving forward, expanding local spectral libraries is paramount. The development of a comprehensive South African Soil Spectral Library would provide the foundation for more accurate prediction models by ensuring that regional soil variability is well-represented. This effort requires systematic collection of spectral and laboratory-measured data across diverse soil types. Additionally, integrating MIR and NIR data with complementary technologies, such as pXRF or LIBS, offers an opportunity to enhance predictions through multi-model approaches. Advancements in machine learning, including deep learning models like CNN, can further improve predictive robustness and enable multi-task learning for multi-variate analysis.

Spiking remains a critical technique to address the disconnection between global spectral libraries and local soil variability. This involves incorporating a minimal yet effective number of local samples into global models to significantly improve predictive accuracy. Strategies for example dynamic spiking, where spiking levels are adjusted based on the spectral characteristics of specific soil properties, could maximize the benefits of this approach while mitigating overfitting risks. Practical field-ready solutions, developed through collaborations between researchers, agronomists, and technology developers, are essential to ensure that these advances translate into accessible tools for farmers. Additionally, government and industry support for spectral library development, research, and capacity building will be pivotal in scaling these technologies.

The implications of advancing soil spectroscopy in South Africa are profound. Enhanced models tailored to regional conditions have the potential to revolutionize precision agriculture, offering cost-effective, accurate, and scalable solutions for soil analysis. This progress would not only support sustainable farming practices but also contribute to food security and environmental conservation. With continued investment in research and infrastructure, soil spectroscopy can become a cornerstone of South Africa's agricultural strategies, setting a precedent for other regions with similarly underrepresented soils.

BIBLIOGRAPHY

Adeline, K.R.M., Gomez, C., Gorretta, N. & Roger, J.M. 2017. Predictive ability of soil properties to spectral degradation from laboratory Vis-NIR spectroscopy data. *Geoderma*. 288:143-153.

Ahmadi, A., Emami, M., Daccache, A. & He, L. 2021. Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: A systematic review and meta-analysis. *Agronomy*. 11(3).

Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In: (KDD '19). *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery. pp. 2623-2631.

Ambushe, A.A., Du Plessis, A. & McCrindle, R.I. 2015. Laser-induced breakdown spectroscopy and inductively coupled plasma-mass spectrometry for determination of Cr in soils from Brits District, South Africa. *Bulletin of the Chemical Society of Ethiopia*. 29(3):357-366.

Angelopoulou, T., Balafoutis, A., Zalidis, G. & Bochtis, D. 2020. From laboratory to proximal sensing spectroscopy for soil SOC estimation-A review. *Sustainability (Switzerland)*. 12(2).

Atkins, P. & De Paula, J. 2010. *Physical Chemistry*. 9th ed. New York: W. H. Freeman and Company.

Bai, Z., Xie, M., Hu, B., Luo, D., Wan, C., Shi, Z. 2022. Estimation of Soil Organic Carbon Using Vis-NIR Spectral Data and Spectral Feature Bands Selection in Southern Xinjiang, China. *Sensors*. 22(16).

Barthès, B.G., Kouakoua, E., Coll, P., Clairotte, M., Moulin, P., Chevallier, T. 2020. Improvement in spectral library-based quantification of soil properties using representative spiking and local calibration - The case of soil inSOC prediction by mid-infrared spectroscopy. *Geoderma*. 369.

Bayer, A., Bachmann, M., Müller, A. & Kaufmann, H. 2012. A Comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African Ecosystem. *Applied and Environmental Soil Science*. 2012.

Breiman, L. & Cutler, A. 2018. *Breiman and Cutler's Random Forests for Classification and Regression, CRAN*

Breure, T., Milne, A., Webster, R., Hannam, J., Haefele, S. & Corstanje, R. 2020. *Quantifying the uncertainty in the prediction of soil properties from NIR and MIR soil-spectra using local and regional spectral libraries*.

Bruker OPTIK GmbH. 2019. *OPUS Reference Manual*. Ettlingen, Germany. www.bruker.com.

Buontempo, C., Thépaut, J.-N. & Bergeron, C. 2020. Copernicus Climate Change Service. *IOP Conference Series: Earth and Environmental Science*. 509(1):012005.

Canero, F.M., Rodriguez-Galiano, V. & Aragones, D. 2024. Machine Learning and Feature Selection for soil spectroscopy. An evaluation of Random Forest wrappers to predict soil organic matter, clay, and carbonates. *Heliyon*. 10(9).

Centre, E.C.J.R., Jones, A., Fernández-Ugalde, O. & Scarpa, S. 2020. *LUCAS 2015 topsoil survey - Presentation of dataset and results*. Publications Office.

Chandola, V., Banerjee, A. & Kumar, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41(3).

Chang, C., Laird, D.A., Mausbach, M.J. & Hurburgh, C.R. 2001. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. *Soil Science Society of America Journal*. 65(2):480-490.

Chao, W., Qiao, X., Li, G., Feng, M., Xie, Y., Yang, W. 2021. Hyperspectral Estimation of Soil Organic Matter and Clay Content in Loess Plateau of China. *Agronomy Journal*. 113(3):2506-2523.

Clingensmith, C.M. & Grunwald, S. 2022. Predicting Soil Properties and Interpreting Vis-NIR Models from across Continental United States. *Sensors*. 22(9).

Dangal, S., Sanderman, J., Wills, S. & Ramirez-Lopez, L. 2019. Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library. *Soil Systems*. 3(1):11.

Demattê, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M. V., Dalmolin, R.S.D., do Couto, H.T.Z. 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma*. 354(August):113793.

Dorantes, M.J., Fuentes, B.A. & Miller, D.M. 2022. Calibration Set Optimization and Library Transfer for Soil Carbon Estimation Using Soil Spectroscopy—A Review. *Soil Science Society of America Journal*. 86(4):879-903.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., … Lautenbach, S. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 36(1):27-46.

Du, C. & Zhou, J. 2009. Evaluation of soil fertility using infrared spectroscopy: a review. *Environmental Chemistry Letters*. 7:97-113. https://api.semanticscholar.org/CorpusID:93237773.

Dudek, M., Kabała, C., Łabaz, B., Mituła, P., Bednik, M. & Medyńska-Juraszek, A. 2021. Mid-infrared spectroscopy supports identification of the origin of organic matter in soils. *Land*. 10(2):1-11.

FAO. 2022. *A primer on soil analysis using visible and near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy*. *A primer on soil analysis using visible and near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy*. FAO.

Farooqi, Z., Ayub, M., Nadeem, M., Shabaan, M., Ahmad, Z., Iftikhar, I. 2021. Precision Agriculture to Ensure Sustainable Land Use for the Future. 210-230.

Fuentes, B., Ashworth, A., Ngunjiri, M.W. & Owens, P. 2021. Mapping Soil Properties to Advance the State of Spatial Soil Information for Greater Food Security on US Tribal Lands. 1.

Garrett, L.G., Sanderman, J., Palmer, D.J., Dean, F., Patel, S., Carlin, T. 2022. Mid-infrared spectroscopy for planted forest soil and foliage nutrition predictions, New Zealand case study. *Trees, Forests and People*. 8:100280.

Ge, Y., Thomasson, J.A. & Sui, R. 2011. Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science*. 5(3):229-238.

Godfray, H., Beddington, J., Crute, I., Haddad, L., Lawrence, D., Toulmin, C. 2010. Food Security: The Challenge of Feeding 9 Billion People. *Science*. 327:812-818.

Gomez, C., Lagacherie, P. & Coulouma, G. 2008. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma*. 148(2):141-148.

Greschuk, L.T., Araújo, M.G. da S., Albarracín, H.S.R., Bellinaso, H., Silvero, N.E.Q., Demattê, J.A.M. 2022. Combining spectral ranges for soil discrimination: A case study in the State of Maranhão - Brazil. *Geoderma Regional*. 29.

Guerrero, C., Zornoza, R., Gómez, I. & Mataix-Beneyto, J. 2010. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma*. 158(1-2):66-77.

Haghi, R.K., Pérez-Fernández, E. & Robertson, A.H.J. 2021. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. *Geoderma*. 396:115071.

Hassan, S.A., Yahya, S.R.S., & Hafeez,, K. A. 2021. Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python. *International Journal of Advanced Trends in Computer Science and Engineering*. 10(1):277-281.

Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Crouch, J. 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*. 11(1).

Hollas, M.J. 2004. *Modern Spectroscopy*. 4th ed. John Wiley & Sons Ltd.

Hong, Y., Chen, Y., Zhang, Y., Liu, Y., Liu, Y., Cheng, H. 2018. Transferability of Vis-NIR models for Soil Organic Carbon Estimation between Two Study Areas by using Spiking. *Soil Science Society of America Journal*. 82(5):1231-1242.

Hosseini, F.S., Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Choi, S.M. & Jamshidi, M. 2023. Spatial prediction of physical and chemical properties of soil using optical satellite imagery: a state-of-the-art hybridization of deep learning algorithm. *Frontiers in Environmental Science*. 11.

Huang, C.-C., Liu, S., Li, R., Sun, F., Zhou, Y. & Yu, G. 2016. Spectroscopic Evidence of the Improvement of Reactive Iron Mineral Content in Red Soil by Long-Term Application of Swine Manure. *Plos One*. 11(1):e0146364.

ICRAF, 2021, W.A., Reference, I.S. & (ISRIC), I.C.

Ioannou, Y.A., Robertson, D., Cipolla, R. &Criminisi, A. 2016. Deep Roots: Improving CNN Efficiency with Hierarchical Filter Groups. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5977-5986.

Islam, K., Singh, B. & McBratney, A. 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Soil Research*. 41(6):1101-1114. https://doi.org/10.1071/SR02137.

L.J. Janik, R.H. Merry, J.O., S. (1998) “Can mid infrared diffuse reflectance analysis replace soil extractions?,” Australian Journal of Experimental Agriculture, 38, pp. 681–696.

Janik, L.J., Forrester, S.T. & Rawson, A. 2009. The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. *Chemometrics and Intelligent Laboratory Systems*. 97(2):179-188.

Jiang, Q., Li, Q., Wang, X., Wu, Y., Yang, X. & Liu, F. 2017. Estimation of soil SOC and total nitrogen in different soil layers using VNIR spectroscopy: Effects of spiking on model applicability. *Geoderma*. 293:54-63.

John, K., Afu, S.M., Isong, I.A., Aki, E.E., Kebonye, N.M., Penížek, V. 2021. Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics. *International Journal of Environmental Science and Technology*. 18(11):3327-3342.

Jović, B., Ćirić, V., Kovačević, M., Šeremešić, S. & Kordić, B. 2019. Empirical equation for preliminary assessment of soil texture. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. 206:506-511.

Kamilaris, A. and Prenafeta-Boldú, F.X. (2018) “A review of the use of convolutional neural networks in agriculture,” Journal of Agricultural Science. Cambridge University Press, pp. 312–322. Available at: https://doi.org/10.1017/S0021859618000436.

Kandpal, L.M., Munnaf, M.A., Cruz, C. & Mouazen, A.M. 2022. Spectra Fusion of Mid-Infrared (MIR) and X-ray Fluorescence (XRF) Spectroscopy for Estimation of Selected Soil Fertility Attributes. *Sensors*. 22(9).

Kawamura, K., Nishigaki, T., Andriamananjara, A., Rakotonindrina, H., Tsujimoto, Y., Razafimbelo, T. 2021. Using a one-dimensional convolutional neural network on visible

and near-infrared spectroscopy to improve soil phosphorus prediction in Madagascar. *Remote Sensing*. 13(8).

Khaire, U.M. & Dhanalakshmi, R. 2020. High-dimensional microarray dataset classification using an improved adam optimizer (iAdam). *Journal of Ambient Intelligence and Humanized Computing*. 11(11):5187-5204.

Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B. & Harris, W.G. 2015. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma*. 239-240:229-239.

Koch, J., Chakraborty, S., Li, B., Kucera, J.M., Van Deventer, P., Weindorf, D.C. 2017. Proximal sensor analysis of mine tailings in South Africa: An exploratory study. *Journal of Geochemical Exploration*. 181:45-57.

Kock, A., Ramphisa Nghondzweni, D. & Van Zijl, G. 2024. Development of soil spectroscopy models for the Western Highveld region, South Africa: Why do we need local data? *European Journal of Soil Science*. 75(6).

Kuang, B., Tekin, Y. & Mouazen, A.M. 2015. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil SOC, pH and clay content. *Soil and Tillage Research*. 146(PB):243-252.

Kuhn, M. & Johnson, K. 2013. *Applied Predictive Modeling*. New York, NY: Springer New York.

Li, F., Xu, L., You, T. & Lu, A. 2021. Measurement of potentially toxic elements in the soil through NIR, MIR, and XRF spectral data fusion. *Comput. Electron. Agric.* 187:106257.

Li, H., Jia, S. & Le, Z. 2020. Prediction of soil SOC in a new target area by near-infrared spectroscopy: Comparison of the effects of spiking in different scale soil spectral libraries. *Sensors (Switzerland)*. 20(16):1-14.

Lin, S., Ji, R., Chen, C., Tao, D. & Luo, J. 2019. Holistic CNN Compression via Low-Rank Decomposition with Knowledge Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 41:2889-2905.

Liu, L., Ji, M. & Buchroithner, M. 2018. Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery. *Sensors (Switzerland)*. 18(9).

Lobsey, C., Rossel, R.A. V, Roudier, P. & Hedley, C.B. 2017. Data-mines Information From Spectral Libraries to Improve Local Calibrations. *European Journal of Soil Science*. 68(6):840-852.

López, L.E., Wadoux, A.M., Franceschini, M.H.D., Terra, F. d. S., Marques, K.P.P., … Demattê, J.A.M. 2019. Robust Soil Mapping at the Farm Scale With Vis-NIR Spectroscopy. *European Journal of Soil Science*. 70(2):378-393.

Lu, C., Lv, G., Shi, C., Qiu, D., Jin, F., … Sha, W. 2020. Quantitative analysis of pH value in soil using laser-induced breakdown spectroscopy coupled with a multivariate regression method. *Appl. Opt.* 59(28):8582-8587.

Maia, A.J. 2023. Recent Trends on the Use of Infrared Spectroscopy for Soil Assessment. *Journal of Biomedical Research & Environmental Sciences*. 4(11):1618-1623.

Mammadov, E., Denk, M., Riedel, F., Kaźmierowski, C., Lewińska, K., Gläßer, C. 2022. Determination of Mehlich 3 Extractable Elements With Visible and Near Infrared Spectroscopy in a Mountainous Agricultural Land, the Caucasus Mountains. *Land*. 11(3):363.

Mathadeen, P., Miles, N. & Ad, M. 2013. Infrared Reflectance Spectroscopy for the Rapid Measurement of Agronomically Important Soil Properties. (December 2015):108-113.

McBratney, A., Whelan, B., Ancev, T. & Bouma, J. 2005. Future Directions of Precision Agriculture. *Precision Agriculture*. 6(1):7-23.

McCarty, G.W. & Reeves, J.B. 2006. Comparison of near infrared and mid infrared diffuse reflectance spectroscopy for field-scale measurement of soil fertility parameters. *Soil Science*. 171(2):94-102.

Mengel, K., Kirkby, E.A., Kosegarten, H. &Appel, T. eds. 2001. *Principles of Plant Nutrition*. *Principles of Plant Nutrition*. Dordrecht: Springer Netherlands.

Metwally, M.S., Shaddad, S.M., Liu, M., Yao, R.J., Abdo, A.I., Chen, X. 2019. Soil properties spatial variability and delineation of site-specific management zones based on soil fertility using fuzzy clustering in a hilly field in Jianyang, Sichuan, China. *Sustainability (Switzerland)*. 11(24).

Minasny, B. & Mcbratney, A.B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. 32:1378-1388.

Minasny, B. et al. (2013) Cubist, a regression rule approach for use in calibration of NIR spectra.

Morellos, A., Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R., Mouazen, A.M. 2016. Machine learning based prediction of soil total nitrogen, SOC and moisture content by using Vis-NIR spectroscopy. *Biosystems Engineering*. 152:104-116.

Mouazen, A.M., Kuang, B., De Baerdemaeker, J. & Ramon, H. 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma*. 158(1-2):23-31.

Moura-Bueno, J.M., Dalmolin, R.S., ten Caten, A., Dotto, A.C. & Demattê, J. 2019. Stratification of a local Vis-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil SOC predictions. *Geoderma*.

Naimi, S., Ayoubi, S., Di Raimo, L.A.D.L. & Dematte, J.A.M. 2022. Quantification of some intrinsic soil properties using proximal sensing in arid lands: Application of Vis-NIR, MIR, and pXRF spectroscopy. *Geoderma Regional*. 28.

Nanni, M.R. & Demattê, J.A.M. 2006. Spectral Reflectance Methodology in Comparison to Traditional Soil Analysis. *Soil Science Society of America Journal*. 70(2):393-407.

Nawar, S. & Mouazen, A.M. 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *Catena*. 151:118-129.

Ng, W., Minasny, B. & McBratney, A. 2020. Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. *Science of the Total Environment*. 702.

Ng W., Minasny, B., Jones, E. & McBratney, A. 2022a. To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library. *Geoderma*. 406.

Ng W., Minasny, B., Jeon, S.H. & McBratney, A. 2022b. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. *Soil Security*. 6.

Ng W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., McBratney, A.B. 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*. 352:251-267.

Nocita, M., Stevens, A., van Wesemael, B., Brown, D.J., Shepherd, K.D., Montanarella, L. 2015. Soil spectroscopy: An opportunity to be seized. *Global Change Biology*. 21(1):10-11.

Nocita, M., Kooistra, L., Bachmann, M., Müller, A., Powell, M. & Weel, S. 2011. Predictions of soil surface and topsoil SOC content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. *Geoderma*. 167-168:295-302.

Nyawasha, R.W., Wadoux, A.M.J.-C., Todoroff, P., Chikowo, R., Falconnier, G.N., Cardinael, R. 2024. Multivariate regional deep learning prediction of soil properties from near-infrared, mid-infrared and their combined spectra. *Geoderma Regional*. 37:e00805.

Ogen, Y., Zaluda, J., Francos, N., Goldshleger, N. & Ben-Dor, E. 2019. Cluster-based spectral models for a robust assessment of soil properties. *Geoderma*.

Okparanma, R.N. & Mouazen, A.M. 2013. Visible and Near-Infrared Spectroscopy Analysis of a Polycyclic Aromatic Hydrocarbon in Soils. *The Scientific World Journal*. 2013(1).

O'Rourke, S., Minasny, B., Holden, N. & McBratney, A. 2016a. Synergistic Use of Vis-NIR, MIR, and XRF Spectroscopy for the Determination of Soil Geochemistry. *Soil Science Society of America Journal*. 80:888-899.

Padarian, J., Minasny, B. and McBratney, A.B. (2019) "Using deep learning to predict soil properties from regional spectral data," Geoderma Regional, 16. Available at: https://doi.org/10.1016/j.geodrs.2018.e00198.

Padarian, J., Minasny, B. & McBratney, A.B. 2020. Machine learning and soil sciences: a review aided by machine learning tools. *SOIL*. 6(1):35-52.

Peng, Y., et al. (2013) "Predicting Soil Organic Carbon at Field Scale Using a National Soil Spectral Library," Journal of Near Infrared Spectroscopy, 21(3), pp. 213–222. Available at: https://doi.org/10.1255/jnirs.1053.

R Core Team. 2023. https://www.R-project.org/.

Rachman, L.M. 2020. Using Soil Quality Index Plus to Assess Soil Conditions and Limiting Factors for Dryland Farming. *Sains Tanah - Journal of Soil Science and Agroclimatology*. 17(2):100.

Rathore, M. & Singh, P.N., 2022. Application of Deep Learning to Improve the Accuracy of Soil Nutrient Classification. In: *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*. pp. 1-5.

Reitermanova, Z. (2010) “Data splitting,” in WDS, pp. 31–36.

Riebe, D., Erler, A., Brinkmann, P., Beitz, T., Löhmannsröben, H.G. & Gebbers, R. 2019. Comparison of calibration approaches in laser-induced breakdown spectroscopy for proximal soil sensing in precision agriculture. *Sensors (Switzerland)*. 19(23).

Riedel, F., Denk, M., Müller, I., Barth, N. & Gläßer, C. 2018. Prediction of soil parameters using the spectral range between 350 and 15,000 nm: A case study based on the Permanent Soil Monitoring Program in Saxony, Germany. *Geoderma*. 315(December 2017):188-198.

Rossel, R.A.V. & Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*. 158(1-2):46-54.

Rossel, R.A.V. & McBratney, A.B. 2008. Diffuse Reflectance Spectroscopy as a Tool for Digital Soil Mapping. In: *Digital Soil Mapping with Limited Data*. Dordrecht: Springer Netherlands. pp. 165-172.

Rossel, R. V, Walvoort, D., McBratney, A., Janik, L. & Skjemstad, J. 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*. 131:59-75.

Safanelli, J.L. et al. (2023) “Open Soil Spectral Library (OSSL): Building reproducible soil calibration models through open development and community engagement,” bioRxiv, p. 2023.12.16.572011. Available at: https://doi.org/10.1101/2023.12.16.572011.

Sanderman, J. et al. (2023) “Near-infrared (NIR) soil spectral library using the NeoSpectra Handheld NIR Analyzer by Si-Ware.” Zenodo. Available at: https://doi.org/10.5281/zenodo.7600137.

Savitzky, Abraham. & Golay, M.J.E. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*. 36(8):1627-1639.

Schiedung, M., Bellè, S.-L., Malhotra, A. & Abiven, S. 2022. Organic carbon stocks, quality and prediction in permafrost-affected forest soils in North Canada. *North Canada. CATENA*. 213:106194.

Seboko, K.R., van Tol, J. & Kotze, E. 2023. Predicting soil carbon in granitic soils using Fourier-transform mid-infrared (FT-MIR) spectroscopy: the value of database disaggregation. *South African Journal of Plant and Soil*. 40(1):23-33.

Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S. & Vohland, M. 2019. Strategies for the efficient estimation of soil & at the field scale with Vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations. *Geoderma*. 354.

Seybold, C.A., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Thomas, P. 2019. Application of Mid-Infrared Spectroscopy in Soil Survey. *Soil Science Society of America Journal*. 83(6):1746-1759.

Shao, Y. & He, Y. 2011. Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. *Soil Research*. 49(2):166-172.

Sharififar, A., Singh, K., Jones, E. V, Ginting, F.I. & Minasny, B. 2019. Evaluating a Low-cost Portable<scp>NIR</Scp>spectrometer for the Prediction of Soil Organic and Total Carbon Using Different Calibration Models. *Soil Use and Management*. 35(4):607-616.

Shepherd, K.D. & Walsh, M.G. 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*. 66(3):988-998.

Sila, A. 2016. Multivariate calibration teqniques for Infrared spectroscopic data. *Thesis*. *University of Nairobi*

Si-ware Systems. 2023. *Scanner*. Menlo Park.

Soriano-Disla, J.M., Janik, L., Rossel, R.V.V., Macdonald, L. & Mclaughlin, M. 2014. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. *Applied Spectroscopy Reviews*. 49:139-186.

Sparks, D.L. 2003. Environmental soil chemistry: An overview. *Environmental soil chemistry*. 2:1-42.

St. Luce, M., Ziadi, N. & Viscarra Rossel, R.A. 2022. GLOBAL-LOCAL: A new approach for local predictions of soil SOC content using large soil spectral libraries. *Geoderma*. 425.

Stenberg, B., Rossel, R.A. V, Mouazen, A.M. & Wetterlind, J. 2010. Visible and Near Infrared Spectroscopy in Soil Science. 163-215.

Stevens, A. & Ramirez-Lopez, L. 2013. An introduction to the prospectr package. (August, 16).

Summerauer, L., Baumann, P., Ramirez-Lopez, L., Barthel, M., Bauters, M., Six, J. 2021. The central African soil spectral library: a new soil infrared repository and a geographical prediction analysis. *SOIL*. 7(2):693-715.

Tavares, T.R., Molin, J.P., Nunes, L.C., Wei, M.C.F., Krug, F.J., Mouazen, A.M. 2021. Multi-sensor approach for tropical soil fertility analysis: Comparison of individual and combined performance of vnir, xrf, and libs spectroscopies. *Agronomy*. 11(6):1028

Tilman, D., Balzer, C., Hill, J.D. & Befort, B.L. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*. 108:20260-20264.

Tsakiridis, N.L., Tziolas, N., Theocharis, J.B. & Zalidis, G.C. 2019. A Genetic Algorithm-based Stacking Algorithm for Predicting Soil Organic Matter From Vis-NIR Spectral Data. *European Journal of Soil Science*. 70(3):578-590.

Ukil, A., Bernasconi, J., Braendle, H., Buijs, H. & Bonenfant, S. 2010. Improved calibration of near-infrared spectra by using ensembles of neural network models. *IEEE Sensors Journal*. 10(3):578-584.

Vågen, T.-G., Winowiecki, L.A., Desta, L., Tondoh, E.J., Weullow, E., Sila, A. 2020. *Mid-Infrared Spectra (MIRS) from ICRAF Soil and Plant Spectroscopy Laboratory: Africa Soil Information Service (AfSIS) Phase I 2009-2013* [Dataset] Dataverse. https://doi.org/10.34725/DVN/QXCWP1

Vašát, R., Kodešová, R., Klement, A. & Borůvka, L. 2017. Simple but efficient signal pre-processing in soil SOC spectroscopic estimation. *Geoderma*. 298:46-53.

Vasques, G.M., Grunwald, S. & Sickman, J.O. 2008. Comparison of Multivariate Methods for Inferential Modeling of Soil Carbon Using Visible/Near-Infrared Spectra. *Geoderma*. 146(1-2):14-25.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Ji, W. 2016. A global spectral library to characterize the world's soil. *Earth-Science Reviews*. 155(January):198-230.

Vona, V., Sarjant, S., Tomczyk, B., Vona, M., Kalocsai, R., Centeri, C. 2022. The effect of local samples in the accuracy of mid-infrared (MIR) and X-ray fluorescence (XRF) -based spectral prediction models. *Precision Agriculture*. 23(6):2027-2039.

Van Vuuren, J.A.J., Meyer, J.H. & Claassens, A.S. 2006. Potential use of near infrared reflectance monitoring in precision agriculture. *Communications in Soil Science and Plant Analysis*. 37(15-20):2171-2184.

Wadoux, A., Brus, D. & Heuvelink, G. 2019. Sampling design optimization for soil mapping with random forest. *Geoderma*. 355:113913.

Wadoux, A.M.J.-C., Malone, B., Minasny, B., Fajardo, M. & McBratney, A.B. 2021. *Soil Spectral Inference with R*. *Springer International Publishing AG, Cham.* http://link.springer.com/10.1007/978-3-030-64896-1.

Wang, J., Sun, Z., Qian, Y., Gong, D., Sun, X., … Song, Y. 2023. Maximizing Spatio-Temporal Entropy of Deep 3D CNN for Efficient Video Recognition. *ArXiv*. abs/2303.02693.

Wang, J.-F. et al. (2012) "A review of spatial sampling," Spatial Statistics, 2(1), pp. 1–14. Available at: https://doi.org/10.1016/j.spasta.2012.08.001.

Waruru, B.K., Shepherd, K.D., Ndegwa, G.M., Sila, A. & Kamoni, P.T. 2015. Application of mid-infrared spectroscopy for rapid characterization of key soil properties for engineering land use. *Soils and Foundations*. 55(5):1181-1195.

Wetterlind, J. & Stenberg B. 2010. Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science*. 61(6):823-843.

Xie, H., Zhao, J., Wang, Q., Sui, Y., Wang, J., Liang, C. 2015. Soil Type Recognition as Improved by Genetic Algorithm-Based Variable Selection Using Near Infrared Spectroscopy and Partial Least Squares Discriminant Analysis. *Scientific Reports*. 5(1).

Xu, Y. & Goodacre, R. 2018. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*. 2(3):249-262.

Xu, H., Xu, D., Chen, S., Ma, W. & Shi, Z. 2020. Rapid determination of soil class based on visible-near infrared, mid-infrared spectroscopy and data fusion. *Remote Sensing*. 12(9).

Xu, L., Hu, O., Guo, Y., Zhang, M., Lu, D., She, Y.-B. 2018. Representative splitting cross validation. *Chemometrics and Intelligent Laboratory Systems*. 183:29-35.

Yang, J., Wang, X., Wang, R. & Wang, H. 2020. Combination of Convolutional Neural Networks and Recurrent Neural Networks for predicting soil properties using Vis-NIR spectroscopy. *Geoderma*. 380.

Yen, S.-J. & Lee, Y.-S. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*. 36(3, Part 1):5718-5727.

Yi, P., Knadel, M., Gislum, R., Deng, F., Nørgaard, T., Greve, M.H. 2013. Predicting Soil Organic Carbon at Field Scale Using a National Soil Spectral Library. *Journal of Near Infrared Spectroscopy*. 21(3):213-222.

Zhang, N., Wang, M. & Wang, N. 2002. Precision agriculture - a worldwide overview. *Computers and Electronics in Agriculture*. 36(2):113-132.

Zhang, P., Guo, Z., Ullah, S., Melagraki, G., Afantitis, A. & Lynch, I. 2021. Nanotechnology and artificial intelligence to enable sustainable and precision agriculture. *Nature Plants*. 7(7):864-876.

Zhang, Z., Ding, J., Zhu, C. & Wang, J. 2020. Combination of efficient signal pre-processing and optimal band combination algorithm to predict soil organic matter through visible and near-infrared spectra. *Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy*. 240:118553.