

RESEARCH ARTICLE

Development of soil spectroscopy models for the Western Highveld region, South Africa: Why do we need local data?

Anru-Louis Kock  | Prudence Dimakatso Ramphisa-Nghondzweni  | George Van Zijl 

Unit for Environmental Sciences and Management, North—West University, Potchefstroom, South Africa

Correspondence

Anru-Louis Kock, Unit for Environmental Sciences and Management, North—West University, Potchefstroom, North West, South Africa.

Email: anru.kock@gmail.com

Funding information

National Research Foundation

Abstract

The increasing global demand for sustainable agriculture requires accurate and efficient soil analysis methods. Conventional laboratory techniques are often time-consuming, costly and environmentally damaging. To address this challenge, we developed and validated locally calibrated mid-infrared (MIR) spectroscopy models for predicting key soil properties pH, phosphorus (P) and exchangeable cations in soil samples from South Africa's Western Highveld region, using a dataset of 979 soil samples and machine learning algorithms Cubist, partial least squares regression (PLSR) and random forest (RF). A subset of spectra was also submitted to the newly developed Open Soil Spectral Library's (OSSL) prediction models to determine whether global prediction models could be used for local soil property prediction. Accurate predictions for pH, calcium (Ca) and magnesium (Mg), with coefficient of determination (R^2) values exceeding 0.76 were obtained with the local calibration algorithms. The predictions for P, potassium (K) and sodium (Na) did not meet the requirements for reliability. Soil spectroscopic prediction models calibrated with local soils outperformed the corresponding global prediction models considered. The OSSL prediction results were inaccurate, with a RPIQ <1, and consistently underpredicted all soil properties. Furthermore, the OSSL collection of prediction models does not include a pH (KCl) model, the routinely used pH measurement method in South Africa. These findings highlight the importance of local calibration for accurate soil property prediction and underscore the need for regional representation in global spectral libraries. This research serves as the first local calibration of MIR spectroscopy models for the Western Highveld region of South Africa and provides a foundation for future local soil property inference model development. It also serves as a potential starting point for a comprehensive South African soil spectral library that can be contributed to global spectral libraries.

KEYWORDS

conditioned Latin hypercube sampling, Cubist, local calibration, mid-infrared, random forest, soil properties, soil spectral inference, spectral library

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *European Journal of Soil Science* published by John Wiley & Sons Ltd on behalf of British Society of Soil Science.

1 | INTRODUCTION

The South African population is projected to increase significantly by 2035 to 70 million (Statistics South Africa, 2022). This growing demographic pressure therefore requires a focus on sustainable farming practices to intensify food production (Tilman et al., 2011). One of the actions to take to achieve this end is precision agriculture, which is defined by the International Society of Precision Agriculture as ‘*a management strategy that gathers, processes and analyses temporal, spatial and individual plant and animal data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production*’ (Van Vuuren et al., 2006). A form of precision agriculture is site-specific crop management (SSCM), in which resource application and agronomic practices are correlated with soil characteristics and crop production as they vary across a field (Schmaltz, 2017).

Soil properties vary between landscapes due to regional variations in parent material, climate, topography, biological activity and the passage of time (Jenny, 1994; Warrick, 2002). Soil variability poses a unique challenge for SSCM in terms of soil assessment and soil condition evaluation and requires the use of analytical methods with numerous parameters representing many locations (Nocita, Stevens, van Wesemael, Aitkenhead, et al., 2015). Conventional soil analysis methods employed in laboratories are hindered by their time-consuming nature, high costs and the use of environmentally hazardous chemicals (Van Vuuren et al., 2006). In this context, soil spectroscopy has emerged globally as an innovative technology that offers rapid and cost-effective measurements of soil properties without the environmental risks associated with traditional methods (Van Vuuren et al., 2006; Wadoux, Malone, et al., 2021). Although there is significant activity devoted to soil spectroscopy research globally, South Africa has been trailing behind and limited to small, localized studies most notably by Van Vuuren et al. (2006), Nocita et al. (2011), Mathadeen et al. (2013) and Seboko et al. (2023). Until now no regional calibration models have existed for routinely used soil properties in SSCM areas, which has hindered the adoption and development of soil spectral technologies for use in precision agriculture in South Africa.

The development of accurate and efficient soil property prediction models along with an expanding soil spectral library will empower farmers and agronomists with the information needed to optimize resource management and enable more sustainable agricultural practices in South Africa. This contribution will improve—on a

Highlights

- First locally calibrated mid-infrared (MIR) prediction models for key soil properties in Western Highveld region of South Africa.
- Soil inference models were created using Cubist, partial least squares regression (PLSR) and random forest (RF).
- Local models outperform global model predictions, highlighting regional soil variability.
- The study provides foundation for future soil spectral library development in South Africa.

small scale—the global understanding of soil spectral variability, where existing global soil spectroscopy models often underperform due to the absence of the unique soil characteristics of South Africa. This highlights the importance of regional representation in global soil spectroscopy models. The use of soil spectroscopy in this study also addressed one of the 10 challenges of Pedometrics, which is developing more efficient methods for measuring soil properties (Wadoux, Heuvelink, et al., 2021).

Soil spectroscopic models typically use visible, near and mid-infrared (MIR) electromagnetic radiation and have been used to predict important soil properties such as pH, total carbon, organic carbon, organic matter, extractable phosphorus (P), extractable potassium (K), total nitrogen (N), exchangeable cations, cation exchange capacity, electrical conductivity, extractable microelements, trace and major element analysis, gypsum, electrical conductivity, soluble sulphate and chloride analysis, bulk density, coarse fragments, particle size distribution, porosity, hydraulic conductivity, aggregate stability, moisture content, microbial biomass, soil respiration, enzyme activity, microbial identification, heavy metal elements and other soil pollutants (McCarty & Reeves, 2006; Peng et al., 2016; Safanelli et al., 2023; Soriano-Disla et al., 2014). Mid-infrared spectroscopy (4000–400 cm^{−1}) produces better predictions for soil properties than visible and near-infrared spectroscopy due to the intense vibrational fundamentals of molecular bonds dominating the MIR regions compared to weaker vibrational signals caused by overtones and combinations in the NIR region (McCarty & Reeves, 2006). However, to be able to apply soil spectroscopy for accurate soil property analysis, calibration of the prediction model for each soil property is a prerequisite (Waruru et al., 2015). This calibration is based on a robust soil spectral library, that is, a conventionally analysed soil property database that is correlated with spectroscopic scans of the same soil samples (Waruru et al., 2015). Machine learning algorithms like

Cubist, partial least squares regression (PLSR) and random forest (RF) are used and were employed in this study. An example of a global initiative is the development of the Soil Spectroscopy for Global Good project, which created the Open Soil Spectral Library (OSSL), which also boasts of an estimation service that utilizes calibrated prediction models from globally available open-source Vis-NIR, NIR and MIR spectral data (Safanelli et al., 2023).

This study addressed this critical need in a South African context by creating a local soil spectral library and developing locally calibrated soil spectroscopy models for key soil properties essential in SSCM. The objectives were to: (i) create a local soil spectral library which contains SSCM soil data by combining soil chemical analysis data with the spectral measurements data into one large database; (ii) develop novel MIR calibration models using a soil spectral library containing locally representative soil samples for key soil properties in the Western Highveld region using Cubist, PLSR and RF; (iii) use the best performing local model and compare it to the globally available OSSL models using local soil spectral data. The hypotheses are that: (i) accurate MIR prediction models will be created for certain soil properties and that some soil properties will not be predicted accurately and (ii) locally calibrated MIR spectroscopy models will achieve higher predictive accuracy for pH

and exchangeable cations (Ca, Mg, K, Na) compared to the global OSSL models in the Western Highveld region. This study represents a pioneering effort to establish locally calibrated MIR soil spectroscopy models for the Western Highveld region, which included a soil spectral library that can be extended to other regions of the country, thereby providing a critical foundation for advancing precision agriculture adoption in South Africa.

2 | MATERIALS AND METHODS

2.1 | Study area

All samples originate within the Western Highveld grain-producing region of South Africa, located in the North-West province (Figure 1). The study area is located at an average height of 1769 m above sea level with a temperature range of -8.6°C to 37.5°C and a precipitation range of 300–700 mm, making it part of the arid to semi-arid regions. The geology consists of several formations. The western parts are covered by the Kalahari group which consists mainly of conglomerate, mudstone, gritstone, siliceous/calcareous sandstone, silcrete, diatomaceous limestone and calcrete, while the eastern parts include the Klipriviersberg group, Malmani subgroup and Rietgat formations, which consist of dolomite, chert, shale,

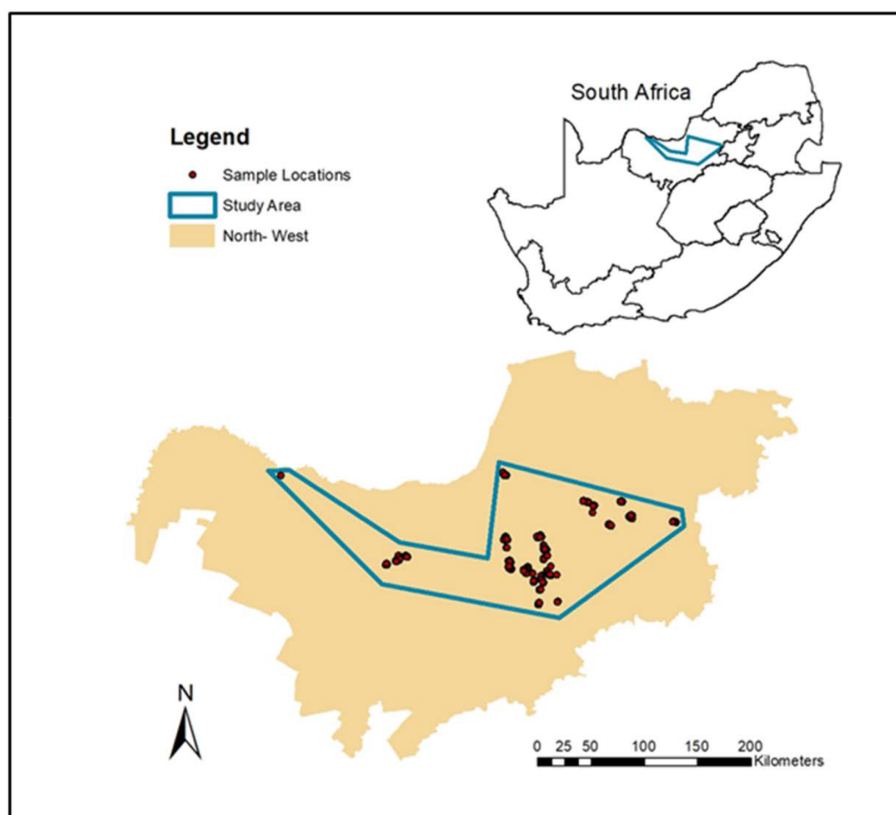


FIGURE 1 Sample locations within the Western Highveld grain-producing area in North West, South Africa.

limestone, quartzite, basalts andesites, tuff, agglomerate, gneiss and granites (Johnson et al., 2006). On a broad scale the following soils occur: red and yellow redoximorphic, plintic, black or red clays, shallow soils, rock outcrops and alluvial soils (LTSS (Land Type Survey Staff), 2021). According to the vegetation map of South Africa the region is classified as Western Highveld Sandy Grassland. A wide variety of crops are cultivated within the Western Highveld region which include and is not limited to: *Zea mays* L.; *Triticum aestivum*; *Helianthus annuus* L.; *Glycine Max* and *Arachis hypogaea*. The most important crops produced are *Zea mays* L. and *Helianthus annuus* L. within the study area.

2.2 | Methodology

2.2.1 | Sample and soil property data acquisition

Routine soil analysis samples for precision agriculture were collected by two private SSCM service providers, NWK Ltd. (Lichtenburg, North West, South Africa) and GWK Ltd. (Douglas, Northern Cape, South Africa), during July–August 2020 for the 2020–2021 growing season and originated from croplands that are used for the cultivation of annual crops, including *Zea mays* L. and *Helianthus annuus* L. The samples were analysed for the soil properties used for SSCM within this area, including pH in a 1:2.5 KCl solution, plant available P with the Bray-1 method and exchangeable base cations (Ca, Mg, K and sodium [Na]) extracted with 1 M ammonium acetate. The two service providers provided access to the conventional soil analysis results of 4393 (NWK Ltd.) and 175 (GWK Ltd.) soil samples. Due to capacity constraints, 900 and 100 samples were, respectively, selected from each database using conditioned Latin hypercube sampling (cLHS) using pH, P and the sum of base cations as variables (Minasny & Mcbratney, 2006). This method resulted in a soil property dataset that was well representative of the region from which the samples were selected during the soil property dataset reduction.

2.2.2 | Soil sample and data preparation

The soil property data obtained from the laboratories were checked for normality and square root transformed to compensate for non-normality, Cubist and RF does not require data to be normally distributed, but due to the use of PLSR in this study the soil property data were transformed to achieve consistency (Dangal et al., 2019). The prediction values were back transformed prior to

statistical analysis. Each soil sample was oven-dried for 24 h at 60°C and ground using a Retsch Mortar Grinder RM 200 (Retsch GmbH, Haan, Germany, 2021) to obtain a grain size of 100 µm.

2.2.3 | Spectral measurement

The prepared soil samples were then scanned 30 times using a Bruker Alpha II spectrometer with DRIFT module (Bruker OPTIK GmbH, Ettlingen, Germany, 2019) with a spectral range of 3996–398 cm⁻¹ and a resolution of 2 cm⁻¹. The samples were placed inside the measurement cup and levelled to create a uniform surface before being placed inside the instrument. Spectral reflectance of each soil sample was averaged from the 30 spectra, which were scanned automatically using the OPUS Base package software (Bruker OPTIK GmbH, Ettlingen, Germany, 2019). Built-in quality checks along with background measurements on a reference standard were conducted at the start of each scanning session and background measurements were obtained at 1-h intervals during the scanning process, as recommended by the manufacturer, to ensure that the spectral measurements were adjusted for atmospheric—temperature and humidity—changes.

2.2.4 | Spectral library and spectral preprocessing

A spectral library was created for each of the measured soil properties, which include pH, P and exchangeable cations by combining the soil property and spectral datasets using R version 4.2.0 (R Core Team, 2023) with the dplyr package (Wickham et al., 2023). The analyses were performed on a computer with the following specifications: Intel® Core™ i7-12700H CPU @ 3.5GHz, 16 GB RAM and running Windows 11 (R Core Team, 2023). The soil spectral library consisted of 979 samples—21 scans were lost due to sample name mismatching, attributed to human error during sample handling. To ensure this did not affect the normality assumption for PLSR, we performed a normality test on the remaining dataset. The dataset was checked for normality both before and after the removal of the 21 scans, using the Shapiro–Wilk test. No significant deviations from normality were found because of the missing scans, and the data were square root transformed to improve normality where necessary (Dangal et al., 2019). The missing scans had no significant impact on the data distribution or model performance. The spectra were preprocessed by applying a Savitzky–Golay filter using the *prospectr* package with the *savitzkyGolay* function a window size of

11, polynomial order of 2 and first-order derivative were used, which produced the best prediction models according to the statistical performance results (Stevens & Ramirez-Lopez, 2022; Wadoux, Malone, et al., 2021).

2.2.5 | Prediction model calibration

After preprocessing, the soil spectral library was divided into calibration and validation datasets with a 70:30 ratio, respectively, using k-means clustering on the spectra (Malone et al., 2017). Other potential methods for partitioning datasets include random sampling, stratified sampling (Kohavi, 1995) and Kennard–Stone sampling (Kennard & Stone, 1969). K-means clustering was selected for this study because it effectively groups similar spectra and ensures that both calibration and validation sets represent the full spectral variability of the dataset, which is crucial for model generalization. This method minimizes the risk of overfitting by ensuring that the validation set is not disproportionately represented by any single group of spectra, as can sometimes occur with random or stratified sampling (Malone et al., 2017). The machine learning algorithms Cubist, PLSR and RF were then used in R to create the calibration algorithms for the soil properties selected for the calibration dataset, using the *cubist* function in the *Cubist* package (Max Kuhn, 2021). Cubist is a rule-based model which finds relationships between spectral and soil property data and creates ‘if-then’ rules which are helpful for identifying important spectral features associated with certain soil characteristics. For PLSR, the *pls* function from the *pls* package (Mevik & Wehrens, 2015) was used. PLSR is effective when handling highly correlated variable data while also reducing the dimensionality of the spectral data. It functions by identifying the most relevant spectral information for soil property prediction. To prevent overfitting and ensure model accuracy, determining the optimal number of latent factors (components) is crucial. The optimal number of PLSR factors was determined using k-fold cross-validation, a standard technique that partitions the data into k subsets—in this study, $k = 10$. In each fold, one subset is reserved as the validation set, while the remaining subsets are used to train the model. This process is repeated k times, with each subset being used exactly once as the validation data. The root mean square error of cross-validation (RMSECV) is calculated at each step, and the number of PLSR factors that minimizes the RMSECV is selected as optimal (Wadoux, Malone, et al., 2021). In this study, cross-validation was performed on the calibration dataset, and the number of components that minimized RMSECV was chosen. Additionally, to avoid overfitting, the performance of models

with different numbers of components was evaluated by comparing prediction errors on the independent validation set (Mevik & Wehrens, 2015; Wadoux, Malone, et al., 2021). Random forest was implemented using the *randomforest* function from the *randomForest* package (Breiman, 2001). Random forest is an ensemble method known to be accurate and able to handle non-linear relationships and combines multiple decision trees and improves generalization, which is useful when working with diverse soil types and variations in spectral data.

Each of the calibration algorithms was validated with the independent validation dataset with the following statistical parameters as used in the literature: mean error (ME) (Shepherd & Walsh, 2002; Sila et al., 2016; Wadoux, Malone, et al., 2021); root mean square error (RMSE) (Dangal et al., 2019; Shepherd & Walsh, 2002; Wadoux, Malone, et al., 2021); squared Pearson correlation coefficient (r^2) (Dangal et al., 2019; Shepherd & Walsh, 2002; Wadoux, Malone, et al., 2021); coefficient of determination (R^2) calculated as $1 - \text{SSE}/\text{SST}$ where SSE is the sum of the squared error and SST of the total sum of squares (Dangal et al., 2019; Shepherd & Walsh, 2002; Wadoux, Malone, et al., 2021); the ratio of performance to deviation (RPD) (Dangal et al., 2019; Shepherd & Walsh, 2002; Wadoux, Malone, et al., 2021); the ratio of performance to inter-quartile (RPIQ) (Wadoux, Malone, et al., 2021; Waruru et al., 2015); the coefficient of variation (CV) (Cantarella et al., 2006); and the mean absolute percentage error (MAPE) (Albinet et al., 2022). In total 18 prediction models were created with our local dataset, one for each soil property with Cubist, PLSR and RF, respectively. The prediction models were tested against the calibration set as well as the independent validation dataset.

2.2.6 | Comparison with global spectral library

To compare the locally calibrated soil property prediction models with the models calibrated with a global spectral dataset from the OSSSL, the same independent validation dataset used to validate the models calibrated with a local soil spectral library was used as input for the OSSSL estimator. The raw spectral data were separated from the soil property data and uploaded to the estimator and were specified as MIR in the Web application. Each corresponding model for the soil properties used for our local models was selected and prediction values for each soil property used for the local models were also obtained using the OSSSL models. The predictions made using the OSSSL models were then exported and compared with the local model's predictions for the independent validation dataset. The soil and spectral database presented by

the SoilSpec4GG initiative in the OSSL database contains roughly 88,867 georeferenced samples worldwide which were scanned using Vis-NIR, NIR and MIR (Safanelli et al., 2023). The MIR database comprises 46,475 samples globally. From South Africa, the dataset contains 64 samples, 34 samples in the central parts of the country and 30 samples at the eastern part of the country. South Africa's contribution to the global database is sub-par compared to other countries. Using the online OSSL engine with estimation service allowed the upload of the MIR scans of the local independent validation set used for this study to evaluate the performance of the OSSL prediction service for this area (Safanelli et al., 2023). Any uploaded data are stored under the user account and can be activated anytime for obtaining predictions. This enables multiple dataset uploads and reduces the need to reupload the spectral data every time. When uploading the spectral data, a visual plot of the active spectral dataset is presented along with a table of the dataset. The corresponding soil properties were selected, and the prediction models were used to obtain prediction values from the spectral data for each soil property. The predicted soil property values were statistically compared using the same parameters in Table 3, and the prediction results of the OSSL were plotted onto the local prediction model results to compare in Figure 4.

3 | RESULTS AND DISCUSSION

3.1 | Soil property database

The soil property data used in the spectral library based on 979 soil samples closely mirrored the ranges observed

in the initial 4400 samples (Figure 2), indicating that the cLHS method effectively captured the diversity of soil properties in the study area.

The Shapiro-Wilk test confirmed the non-normal distribution of selected soil properties, except for pH, which approached normality but still exhibited a significant deviation ($p < 0.05$). This non-normality is consistent with the broader soil property database and reflects the inherent variability of soil characteristics in the Western Highveld, aligning with the findings of other studies in diverse agricultural regions (Minasny & McBratney, 2016). Descriptive statistics (Table 1, Figure 2) reveal that the selected soil samples represent the typical range of values found in agricultural cropland soils within the grain-producing region. Notably, the soil pH values exhibited a wide range (4.72–8.40), reflecting the diverse parent materials (e.g., basalt, granite and sandstone) and management practices (e.g., liming, fertilization) across the study area (Brady & Weil, 2017). This variability underscores the importance of developing locally calibrated models, as soil pH, along with other critical factors such as soil organic matter content and soil texture, plays a key role in influencing nutrient availability and crop growth (Brady & Weil, 2017; Lal, 2006). The skewed distribution of P concentrations (median 10.5 mg.kg^{-1}) suggests that a significant proportion of soils in the region may be deficient in this essential nutrient. This aligns with previous findings highlighting the widespread P deficiency in South African soils (Barnard & du Preez, 2004). There is a potential use for soil spectroscopy to identify areas requiring P fertilization, enabling targeted nutrient management strategies for improved crop yields (Dhawale et al., 2013). The exchangeable cation concentrations (Ca, Mg, K, Na) also vary considerably,

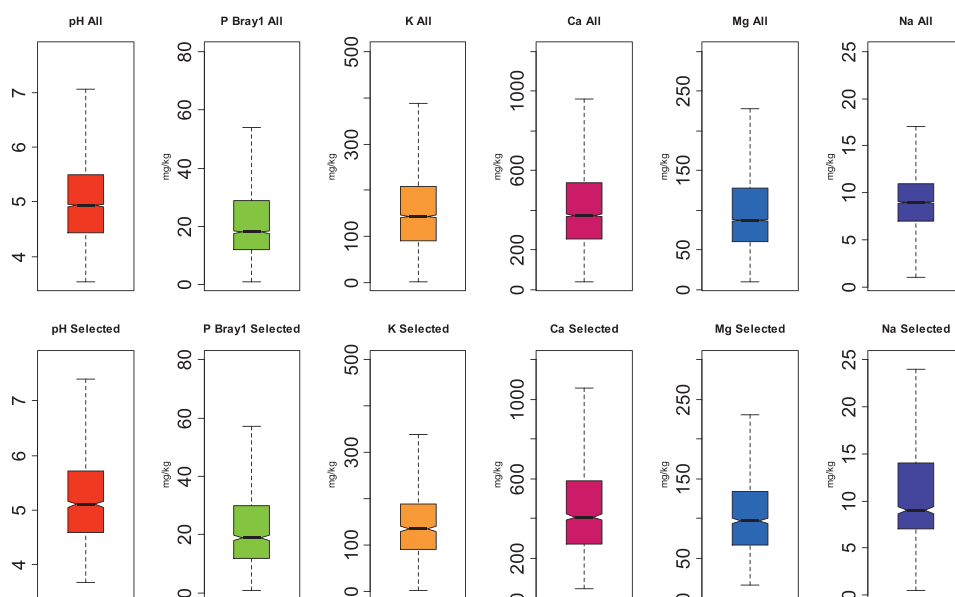


FIGURE 2 Comparison of soil property distributions (pH, P, K, Ca, Mg, Na) in the initial dataset ('All') and the subset selected for model calibration ('Selected') using conditioned Latin hypercube sampling (cLHS). This method ensures that the selected samples are representative of the entire dataset's distribution for each soil property, facilitating the development of robust and accurate prediction models.

reflecting the complex interplay of factors such as parent material, weathering and agricultural practices. Understanding these variations through accurate soil analysis is crucial for optimizing fertilizer recommendations and promoting balanced nutrient supply for crop production.

3.2 | Soil mid-infrared spectra

The MIR spectra of the 979 soil samples exhibited consistent patterns, with strong spectral clusters formed in the ranges of 3700–3000, 2400–2100 and 2000–400 cm^{-1} (Figure 3). These clusters likely represent the fundamental vibrations of molecular bonds associated with common soil constituents such as clay minerals (3700–

3600 cm^{-1}), organic matter (2920–2850 cm^{-1}) and carbonates (1450–1410 cm^{-1}) (McCarty & Reeves, 2006). Minor variations in spectral signatures were observed within the 2400–2300 cm^{-1} range, characterized by deep troughs that intensified during the scanning process. These anomalies were attributed to CO_2 vapour interference, a known phenomenon affecting MIR spectroscopy due to the strong absorption of CO_2 in this spectral region (Nocita, Stevens, van Wesemael, Brown, et al., 2015). To mitigate this effect, the raw spectral data within the 2400–2300 cm^{-1} range were clipped during preprocessing, ensuring that subsequent analyses were based on accurate spectral information. The consistent spectral patterns observed across the samples, despite minor variations, suggest that the soils of the study area

TABLE 1 Descriptive statistics for all soil properties used in the soil spectral database.

Soil properties	μ	Median	σ	Min	Max	Range	<i>n</i>
pH (KCl)	5.19	5.11	0.79	3.68	7.62	3.94	979
Bray-1 P (mg.kg^{-1})	25.2	19	22.53	1	256	255	979
Exchangeable K (mg.kg^{-1})	146.91	134	74.47	2	526	524	979
Exchangeable Ca (mg.kg^{-1})	462.48	407	254.27	45	2243	2198	979
Exchangeable Mg (mg.kg^{-1})	113.69	97	72.82	16	786	770	979
Exchangeable Na (mg.kg^{-1})	11.15	9	7.38	1	85	84	979

Abbreviations: Max, maximum; Min, minimum; *n*, number of samples; μ , mean; σ , standard deviation.

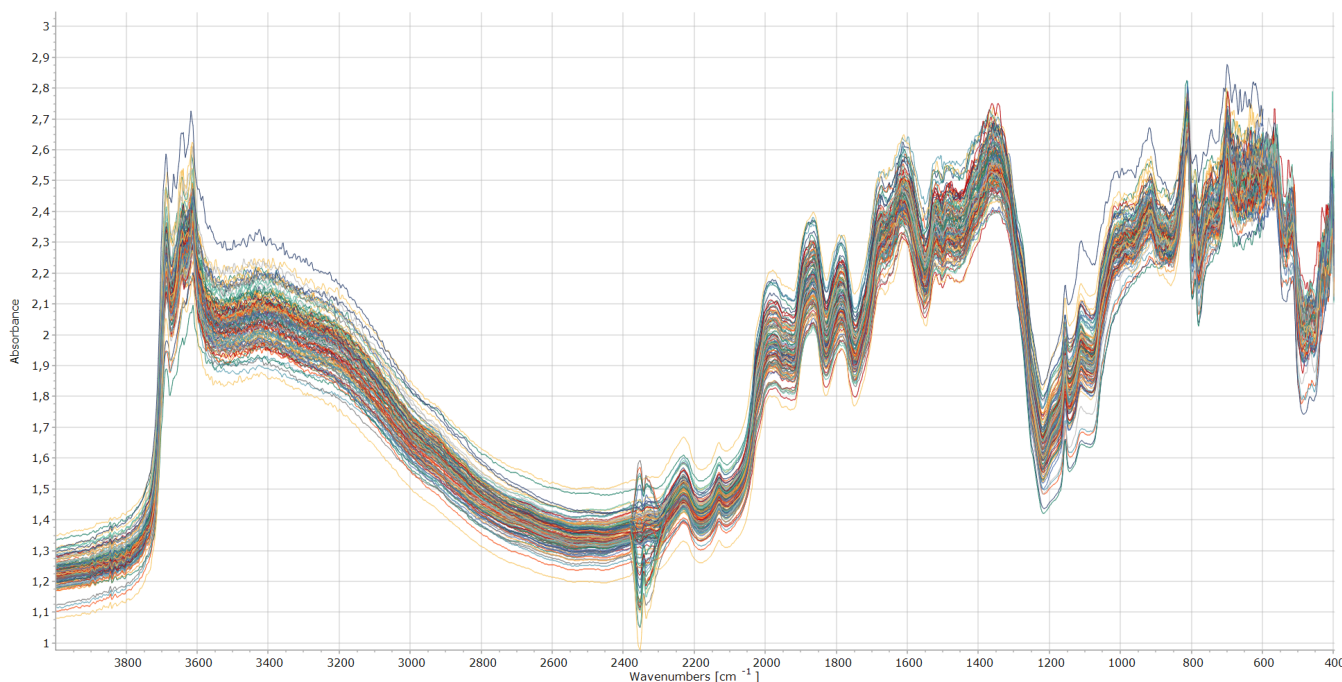


FIGURE 3 Mid-infrared spectral signatures of soil samples collected from the Western Highveld region of South Africa scanned using the Bruker Alpha II Drifts. The spectra exhibit consistent patterns with distinct absorption peaks and troughs, reflecting the diverse mineralogical and organic matter composition of the soils.

share similar mineralogical and organic matter composition. This homogeneity can facilitate the development of robust calibration models for predicting soil properties from MIR spectra. However, the presence of CO₂ interference highlights the importance of careful preprocessing and quality control measures to ensure the accuracy and reliability of soil spectroscopic analysis, as emphasized by previous studies (Minasny & McBratney, 2016).

3.3 | Local prediction model performance with calibration dataset

The statistical measures for each of the 18 calibration algorithms for both the calibration and validation datasets are shown in Table 2. The statistical performance parameter results for the calibration dataset show that most soil properties performed well when evaluated with this dataset. Random forest models outperformed the Cubist and PLSR prediction models for each soil property. Most of the soil properties gave good prediction results for this dataset as did pH for all three algorithms, with the highest R^2 (0.95) and RPIQ (6.03) and lowest RMSE (0.19) using RF. The potassium algorithm performed relatively well with RF ($R^2 = 0.9$, RPIQ = 2.4 and RMSE = 6.95) but to a lesser extent with Cubist and PLSR. Calcium calibrated well with Cubist, RF and PLSR, and RF ($R^2 = 0.95$, RPIQ = 5.61 and RMSE = 57.57) produced superior performance. Potassium ($R^2 = 0.19$, RPIQ = 4.12 and RMSE = 24.05), Na ($R^2 = 0.85$, RPIQ = 2.4 and RMSE = 2.91) and P produced good results using RF, whereas Mg had good predictions for Cubist, RF and PLSR models respectively; again RF ($R^2 = 0.95$, RPIQ = 4 and RMSE = 17.25) performed the best.

3.4 | Prediction model performance with independent validation dataset

Whereas the results for the predictions made using the calibration dataset may indicate successful calibrations, it is more important to evaluate prediction models using an independent dataset (Kuhn & Johnson, 2013), as this indicates the usefulness of the calibration algorithms. Validation using an independent dataset was much worse than on the calibration dataset. Validation results (Table 2) showed that PLSR gave the best results for pH. According to the Soil and Plant Analysis Council (2000), the necessary precision to which pH is read for routine soil analysis is a unit of 0.1. The best-performing prediction model for pH had an RMSE of 0.27. While our models demonstrated satisfactory performance compared to other soil spectroscopy prediction models, caution should be exercised when applying these models to soils

outside the region of calibration. This is consistent with the general understanding that spectral calibrations can be site-specific due to variations in soil properties and environmental factors (Bellon-Maurel et al., 2010; Dhawale et al., 2021; Minasny & McBratney, 2016). Cubist and PLSR models for pH in this study showed similar results than corresponding predictions made with PLSR by Wijewardane et al. (2018) and with MIR ($R^2 > 0.80$ and RMSE < 0.5). Dangal et al. (2019) showed that a Cubist model can produce improved pH predictions ($R^2 = 0.89$ and RMSE = 0.34). The soil pH scatter plots (Figure 2), which compare the predicted versus the actual measured values for each soil property onto a 1:1 line reinforce our findings. The PLSR scatter plot had a higher density of points along the 1:1 line compared to the Cubist and RF plots for pH.

The calibration models for Bray-1 P achieved the best performance with the RF algorithm ($R^2 = 0.58$, RPIQ = 1.08, RMSE = 15.78). However, this performance fell short of the standard laboratory precision (CV = 5%) established by the Soil and Plant Analysis Council (2000). While the accuracy of our models did not meet this criterion, it aligns with findings from other studies on MIR spectroscopy for Bray-1 P prediction. For instance, Soriano-Disla et al. (2014) reported R^2 values of 0.5–0.7, and Wijewardane et al. (2018) observed a similar R^2 of 0.14 and RMSE of 45.17. The challenges in accurately predicting Bray-1 P may be attributed to spectral interferences from silicates and sulphates, which share absorption bands with P in the 1100–1000 cm⁻¹ region (Smith, 1999). Overall, while the current MIR calibration models for Bray-1 P provide reasonable estimates, their accuracy could be improved by addressing the challenges associated with spectral interference and refining the models with larger and more representative datasets. This suggests that future studies could potentially improve prediction accuracy by prioritizing the identification and use of secondary P absorption bands or by incorporating spectral preprocessing techniques designed to mitigate these interferences.

Evaluation of the independent validation dataset indicated that the Cubist model performed best for predicting exchangeable Ca, achieving an R^2 of 0.84, RPIQ of 3.14 and RMSE of 90.32 mg.kg⁻¹. Although this RMSE exceeded the standard laboratory precision (CV = ±7.7%) for exchangeable Ca (Soil and Plant Analysis Council, 2000), the overall accuracy was comparable to other studies employing MIR spectroscopy for exchangeable Ca prediction. For example, Janik et al. (2009) achieved R^2 values of 0.87 and 0.89 using PLS and PLS combined with neural networks (PLS-NN), respectively. Similarly, Ji et al. (2016) reported R^2 values of 0.81 for organic soils and 0.73 for mineral soils. These findings,

along with ours, highlight the suitability of MIR spectroscopy for Ca prediction, likely due to the strong representation of soil organic carbon (SOC) and texture-related absorptions in this spectral region, which are known to be correlated with exchangeable Ca concentrations (Viscarra Rossel et al., 2006). Furthermore, the relationship between soil pH and Ca availability is well-established, with Ca often applied as lime to raise soil pH (Vona et al., 2022). Thus, the accurate prediction of pH using MIR spectroscopy in this study likely contributed to the robust performance of the Ca models. However, our calcium scatter plots (Figure 4) reveal that all three models (Cubist, PLSR and RF) struggled to accurately predict [Ca] exceeding 800 mg.kg^{-1} . This suggests a potential limitation in the current models for high-Ca soils, which may be due to saturation effects or non-linear relationships between spectral features and Ca content at elevated concentrations. This highlights the need for further model refinement, which could be addressed by incorporating additional spectral variables, applying more advanced data transformations or utilizing more sophisticated machine learning algorithms. In addition to the Cubist, RF and PLSR algorithms used in this study, other potential methods such as Support Vector Machines (SVM), Gradient Boosting Machines (GBM) and Extreme Gradient Boosting (XGBoost) could offer improved performance for soil property predictions (Deiss et al., 2020; Liu et al., 2017; Zhao et al., 2022). Implementing these advanced algorithms has been shown to increase prediction accuracy by better capturing complex relationships in high-dimensional spectral data (Deiss et al., 2020; Liu et al., 2017; Zhao et al., 2022).

Exchangeable K prediction proved challenging, with PLSR yielding the most accurate results among the three algorithms ($R^2 = 0.38$, RMSE = 52.83 mg.kg^{-1} , RPIQ = 1.82). Although all models exhibited high accuracy along the 1:1 line in the scatter plots, their precision was lacking, as evidenced by the statistical performance metrics. These findings align with those of Albinet et al. (2022), who also reported better exchangeable K predictions using PLSR ($R^2 = 0.64$) and further improved accuracy ($R^2 = 0.79$) by employing a convolutional neural network (CNN). While definitive accuracy requirements for exchangeable K prediction are not established, some soil analysis handbooks recommend a CV below 18% (Haluschak, 2006). Our models did not achieve this level of precision. However, considering the optimal exchangeable K range for maize production in the Highveld region ($55\text{--}120 \text{ mg.kg}^{-1}$) and the critical level ($<40 \text{ mg.kg}^{-1}$) (FERTASA, 2016), the current PLSR model may still be useful for distinguishing between soils with sufficient and deficient K levels for crop growth. The limited accuracy of exchangeable K prediction in this study may be

attributed to several factors, including: spectral complexity—where the complex spectral signature of exchangeable K, which overlaps with other soil components, may make accurate quantification difficult; sample size and diversity—the relatively small sample size and limited variability in exchangeable K concentrations in our dataset may have hindered model development—and model selection, where PLSR, while effective for capturing linear relationships, may not be the optimal algorithm for complex soil properties like exchangeable K. Exploring alternative approaches, such as CNN as demonstrated by Albinet et al. (2022) could yield improved results. Overall, while the current MIR spectroscopy models for exchangeable K prediction show potential for practical application, further research is needed to enhance their accuracy and precision. This may involve incorporating larger and more diverse datasets, exploring alternative calibration algorithms and utilizing advanced spectral preprocessing techniques to isolate the spectral signature of exchangeable K from interfering compounds.

Exchangeable Mg prediction was most accurate using the Cubist model, achieving an R^2 of 0.76, RMSE of 32.36 mg.kg^{-1} and RPIQ of 2.13. While all models exhibited high accuracy and precision for Mg concentrations below 150 mg.kg^{-1} , prediction accuracy declined for higher concentrations. This suggests that model refinement may be needed to accurately assess soils with elevated Mg levels. Despite this limitation, the Cubist model's performance for Mg within the optimal range for maize production in Highveld soils ($24\text{--}48 \text{ mg.kg}^{-1}$, with a critical level $<19 \text{ mg.kg}^{-1}$) (FERTASA, 2016) indicates its potential utility for informing Mg fertilization decisions. Soils with predicted Mg concentrations above 150 mg.kg^{-1} can likely be considered sufficient, while those below this threshold may require further investigation or supplementation. The accuracy of our Mg prediction model slightly exceeded that of previous studies using MIR spectral data, such as Soriano-Disla et al. (2014) ($R^2 = 0.74$) and Thomasson et al. (2001) ($R^2 = 0.70$). However, Shepherd and Walsh (2002) achieved a higher R^2 of 0.81 using NIR spectroscopy and multivariate adaptive regression splines (MARS), suggesting that exploring alternative spectral regions and modelling techniques could further enhance Mg prediction accuracy.

Exchangeable Na prediction was most accurately modelled using the RF algorithm, but the overall performance was relatively low ($R^2 = 0.21$, RPIQ = 1.15, RMSE = 4.49 mg.kg^{-1}). This accuracy falls short of the acceptable standard (5%–10% CV) for determining exchangeable cation concentrations using ammonium acetate extraction, as established by the Soil and Plant Analysis Council (2000). Previous studies using MIR and

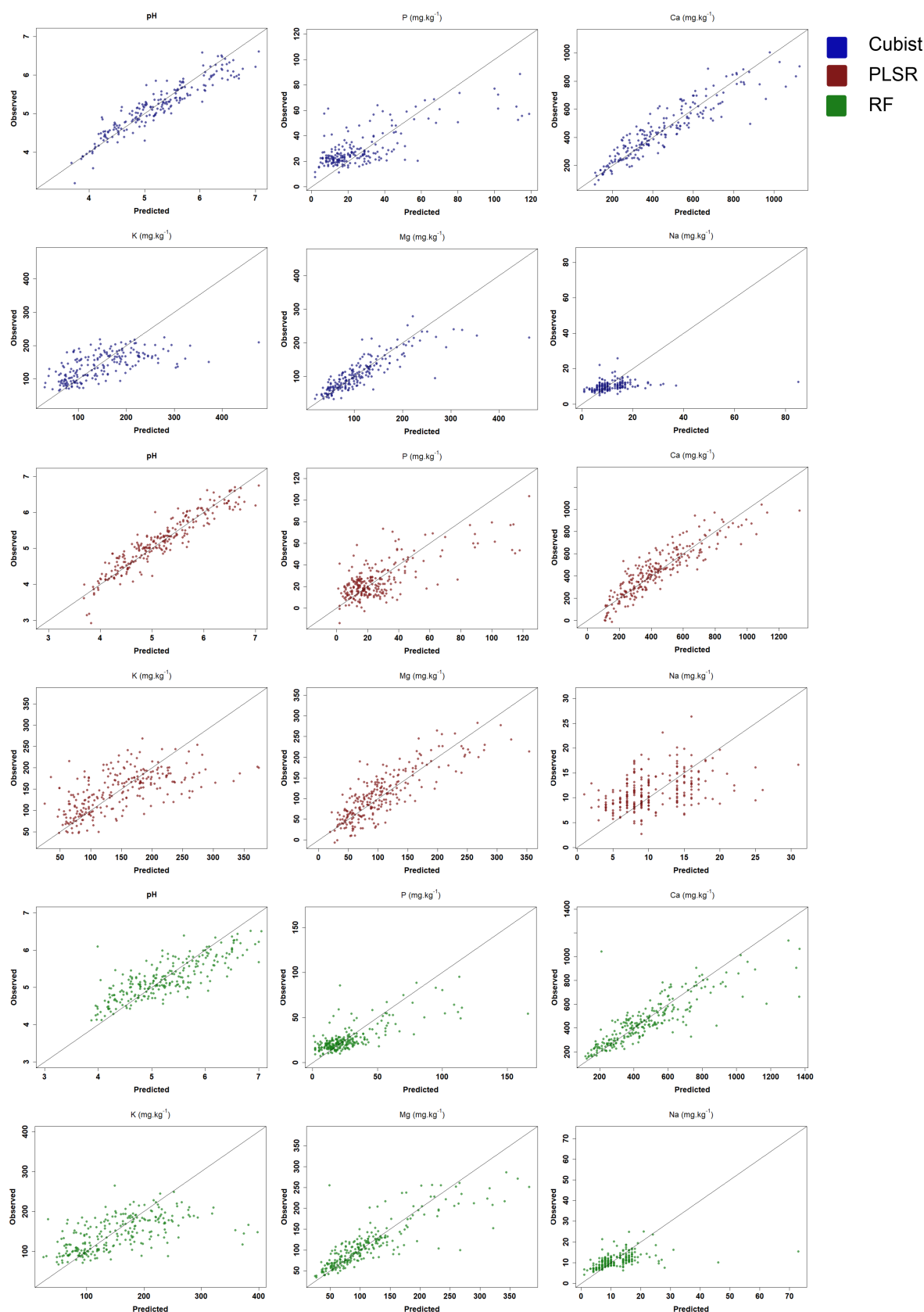


FIGURE 4 Comparison of observed versus predicted values for soil properties (pH, P, K, Ca, Mg, Na) using three different calibration models (Cubist, PLSR and RF) on the independent validation dataset. Each row represents a different calibration model, and each column represents a different soil property. The black line indicates the 1:1 line where observed values equal predicted values.

TABLE 2 Soil property prediction models performance results tested against the calibration and validation datasets respectively.

Property	Model	Calibration					Validation										
		ME	RMSE	r ²	R ²	RPD	RPIQ	CV	MAPE	ME	RMSE	r ²	R ²	RPD	RPIQ	CV	MAPE
pH	Cubist	−0.01	0.25	0.9	0.9	3.23	4.65	4.82	3.38	−0.03	0.29	0.87	0.86	2.7	3.62	5.61	4.03
	PLSR	0	0.27	0.89	0.89	2.99	4.26	5.18	3.94	0.03	0.27	0.87	0.87	2.78	4.11	5.22	4.17
	RF	0	0.19	0.96	0.95	4.31	6.03	3.67	2.7	0	0.45	0.65	0.64	1.67	2.49	8.59	6.67
P	Cubist	−1.66	11.56	0.8	0.74	1.96	1.56	49.54	34.58	−2.28	15.39	0.56	0.52	1.44	1.23	65.09	43.27
	PLSR	0	14.13	0.6	0.6	1.59	1.33	57.1	70.54	−0.95	15.82	0.52	0.52	1.45	1.14	61.53	68.67
	RF	0.11	6.95	0.94	0.9	3.2	2.4	28.14	19.44	−0.91	15.78	0.55	0.54	1.47	1.08	60.14	37.34
Ca	Cubist	−4.67	99.07	0.86	0.86	2.65	3.26	21.47	13.53	−4.5	90.32	0.85	0.84	2.54	3.14	20.2	14.15
	PLSR	0	110.59	0.82	0.82	2.39	2.95	23.53	23.35	10.25	100.82	0.8	0.79	2.19	2.85	22.4	70.29
	RF	0.23	57.57	0.96	0.95	4.5	5.61	12.50	7.97	−14.27	129.06	0.71	0.71	1.85	2.25	28.37	17.81
K	Cubist	−2.92	44.74	0.69	0.65	1.7	2.31	30.94	23.22	−5.06	55.23	0.37	0.37	1.26	1.74	39.44	26.4
	PLSR	0	53.14	0.52	0.52	1.44	1.91	36.04	28.99	−1.14	52.83	0.39	0.38	1.27	1.82	36.64	27.73
	RF	0.87	24.05	0.94	0.9	3.13	4.12	16.48	12.58	−8.88	58.51	0.35	0.34	1.23	1.74	40.76	30.01
Mg	Cubist	−2.18	31.62	0.83	0.82	2.37	2.15	28.12	17.24	−2.86	32.36	0.77	0.76	2.05	2.13	29.96	19.43
	PLSR	0	33.81	0.8	0.8	2.26	2.06	5.18	33.88	3.27	35.04	0.68	0.66	1.71	1.94	32.16	54.55
	RF	−0.16	17.25	0.96	0.95	4.32	4	15.17	9.73	−1.2	36.75	0.71	0.71	1.85	1.77	32.8	19.64
Na	Cubist	−1.01	5.1	0.73	0.53	1.46	1.37	50.19	31.68	−0.8	6.93	0.08	0.06	1.03	1.16	67.25	32.89
	PLSR	0	7.02	0.24	0.24	1.15	1	57.1	37.48	0.88	4.49	0.21	0.11	1.06	1.53	39.61	33.24
	RF	0.12	2.91	0.9	0.85	2.59	2.4	25.97	13.15	−0.42	6.08	0.22	0.21	1.13	1.15	55.57	31.75

Abbreviations: CV, coefficient of variation; MAPE, mean absolute percentage error; ME, mean error; r², coefficient of determination; R², squared correlation coefficient; RMSE, root mean squared error; RPD, ratio of performance to deviation; RPIQ, ratio of performance to interquartile.

NIR spectroscopy have reported higher R^2 values for Na prediction, such as 0.74 (Soriano-Disla et al., 2014) and 0.82 (Safanelli et al., 2023). This discrepancy may be attributed to the limited range of Na concentrations in our dataset, potential spectral interferences from other soil constituents or the inherent limitations of the RF algorithm for predicting this specific property. Future research could improve Na prediction accuracy by incorporating a wider range of Na concentrations in the dataset, exploring alternative machine learning algorithms better suited for Na prediction and implementing advanced spectral preprocessing techniques to isolate and enhance the Na spectral signature.

Results show that there is no one calibration algorithm that is able to accurately apply to all soil properties. Instead, each soil property had varying success using the different methods. RF made better predictions for the soil properties P and Na. The RF model performed worst when predicting the remaining soil properties compared to its performance on P and Na, and better predictions were made using the other machine learning methods. Despite RF demonstrating the best performance in statistical measures on the calibration dataset, the discrepancies observed when applying the model to the validation dataset suggest a tendency towards overfitting, as also noted by Wadoux, Malone, et al. (2021).

TABLE 3 Local calibrated prediction model compared to Open Soil Spectral Library’s prediction models using a local spectral dataset.

Local models						OSSL estimator				
	RMSE	ME	r^2	R^2	RPIQ	RMSE	ME	r^2	R^2	RPIQ
pH (KCl)	0.28	−0.04	0.87	0.92	3.73	No Model Available				
P	15.69	2.87	0.55	0.66	1.26	29.13	−17	0	−0.65	0.68
Ca	87.23	1.09	0.84	0.91	3.4	1490.5	−440.67	0.04	−4.15	0.61
Mg	34.98	−3.01	0.71	0.82	1.89	127.41	−109.94	0	−2.92	0.52
Na	7.48	−1.3	0.08	0.16	1.04	13.67	−11.33	0.02	−2.18	0.57
K	55.23	−5.06	0.37	0.53	1.74	163.2	−146.58	0.02	−4.17	0.61

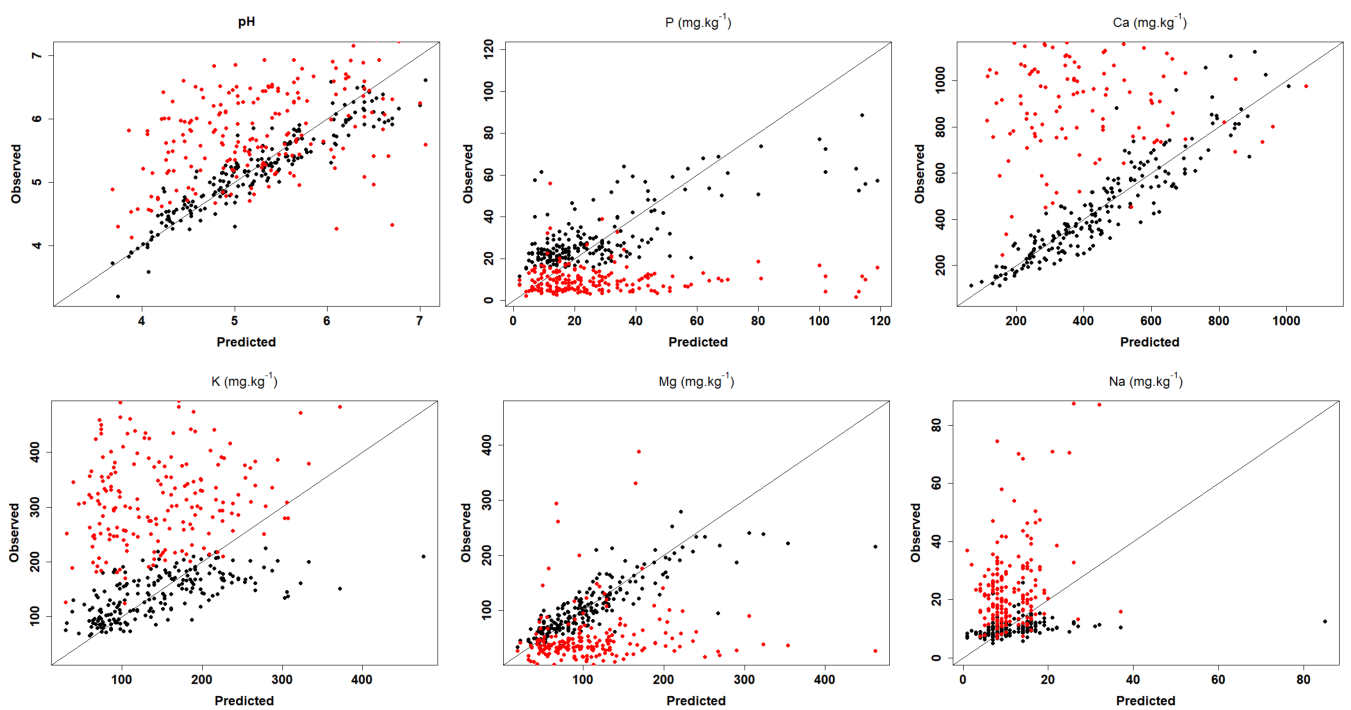


FIGURE 5 The plot illustrates the comparison of observed and predicted soil property values for a set of validation samples using locally calibrated (black points) and OSSL models (red points). The x-axis represents the predicted values, while the y-axis represents the observed values. The 1:1 diagonal line indicates perfect agreement between observed and predicted values. OSSL model predictions for potassium were extremely underpredicted and could not be scaled on the plot.

This study used a relatively small sample size in the soil spectral database to calibrate the respective prediction models for each soil property. This resulted in slightly lower R^2 values than expected when calibrated for pH and Ca. Other studies that used larger sample sizes or large spectral MIR libraries like that of Dangal et al. (2019) produced better R^2 values for their respective prediction models. It is important to note that the calibration dataset calibrated well for all properties, which means that while the validation dataset did not perform well, if the calibration dataset size is increased to represent all possible scenarios of mixtures of soil properties (e.g., all soil pH for different clay contents), then increasing the calibration dataset is a viable option to improve calibration algorithms.

3.5 | Evaluation of global soil inference models with local spectral data

The results confirmed the expected performance of the prediction service with local data. The OSSL flagged all local samples as not represented in the global spectral OSSL MIR database. The representation function flags spectral data from the uploaded samples that are not represented by the spectral feature space of the calibration dataset, which implies that the South African soil spectral data are not well represented in the current OSSL MIR database (Safanelli et al., 2023). The local prediction models outperformed the OSSL prediction models as expected (Table 3 and Figure 5), due to under-representation of local calibration data available in the initial OSSL datasets. All soil properties predicted with the OSSL had RPIQ values <1 and high RMSE, making these predictions unusable. It has been proven that the OSSL works well for spectral data captured in regions sufficiently represented in the initial OSSL dataset used for calibration (Grover & Brungard, 2024). Currently, any spectra collected within South Africa cannot be used using the global model unless sufficient soil data are contributed to the global dataset. Certain soil properties like pH (KCl) are still routinely used commercially in South Africa and are not available in the OSSL estimator; the only available model for pH is 1:1 soil water suspension pH (H_2O) and 1:2 pH ($CaCl_2$). This reinforces the need to create local models that include soil property analysis methods routinely used in South Africa.

4 | CONCLUSION

This study successfully developed MIR calibration models for soil pH, Bray-1 P, exchangeable Ca, K, Mg and Na in the Western Highveld region of South Africa, using Cubist, PLSR and RF algorithms. The results demonstrate

that accurate soil property predictions can be achieved using locally calibrated models, particularly for pH and Ca. However, global models like OSSL currently underperform due to limited representation of South African soil spectral characteristics. The findings highlight the importance of regional soil spectral libraries for accurate soil property predictions in SSCM. Contributing local data to global databases, such as OSSL, can significantly improve their performance for South African users. Furthermore, expanding the scope of local calibration models to include a wider range of soil properties and spectral data, as well as exploring additional machine learning algorithms, can further enhance the accuracy and applicability of soil spectroscopy in South Africa. This research serves as a crucial step towards more sustainable agriculture practices like precision agriculture in South Africa by providing locally relevant, accurate and efficient soil analysis tools. By understanding and accounting for regional soil variability, we can optimize resource management and contribute to food security in the face of growing population demands.

AUTHOR CONTRIBUTIONS

A. Kock: Investigation; methodology; validation; software; formal analysis; data curation; visualization; writing – review and editing; writing – original draft. **P. D. Ramphisa-Nghondzweni:** Writing – review and editing; supervision. **G. M. Van Zijl:** Conceptualization; funding acquisition; writing – review and editing; project administration; supervision; resources; data curation.

ACKNOWLEDGEMENTS

This work is based on the research supported wholly/in part by the National Research Foundation of South Africa (Grant Numbers: 121302). We would like to thank the following people: Martiens Du Plessis at NWK Limited in conjunction with Dailena Pienaar NviroTek Laboratories for providing access to seasonal samples and soil property data and support during this study and Dries Bloem at Geolab for samples and soil property data used in the study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Anru-Louis Kock  <https://orcid.org/0000-0001-5349-9798>

Prudence Dimakatso Ramphisa-Nghondzweni  <https://orcid.org/0000-0002-9637-1716>

George Van Zijl  <https://orcid.org/0000-0001-5003-1081>

REFERENCES

- Albinet, F., Peng, Y., Eguchi, T., Smolders, E., & Dercon, G. (2022). Prediction of exchangeable potassium in soil through mid-infrared spectroscopy and deep learning: From prediction to explainability. *Artificial Intelligence in Agriculture*, 6, 230–241. <https://doi.org/10.1016/j.aiaa.2022.10.001>
- Barnard, R. O., & du Preez, C. C. (2004). Soil fertility in South Africa: The last twenty five years. *South African Journal of Plant and Soil*, 21(5), 301–315. <https://doi.org/10.1080/02571862.2004.10635066>
- Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J. M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends in Analytical Chemistry*, 29(9), 1073–1081. <https://doi.org/10.1016/j.trac.2010.05.006>
- Brady, N., & Weil, R. (2017). *The nature and properties of soils* (15th ed.). Pearson Education.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bruker OPTIK GmbH. (2019). *OPUS Reference Manual*. www.bruker.com
- Cantarella, H., Quaggio, J. A., Van Raij, B., & De Abreu, M. F. (2006). Variability of soil analysis in commercial laboratories: Implications for lime and fertilizer recommendations. *Communications in Soil Science and Plant Analysis*, 37(15–20), 2213–2225. <https://doi.org/10.1080/00103620600817523>
- 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. <https://doi.org/10.3390/soilsystems3010011>
- Deiss, L., Margenot, A. J., Culman, S. W., & Demyan, M. S. (2020). Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma*, 365, 114227.
- Dhawale, N. M., Adamchuk, V. I., Prasher, S. O., & Rossel, R. A. V. (2021). Evaluating the precision and accuracy of proximal soil vis-nir sensors for estimating soil organic matter and texture. *Soil Systems*, 5(3), 48.
- Dhawale, N. M., Adamchuk, V. I., Rossel, R. A. V., Prasher, S. O., Whalen, J. K., & Ismail, A. A. (2013). Predicting extractable soil phosphorus using visible/near-infrared hyperspectral soil reflectance measurements.
- Grover, K., & Brungard, C. (2024). An objective test of the open soil spectral library service. In C. Brungard, A. Wadoux, & S. Salley (Eds.), *Pedometrics* (Vol. 2024, p. 34). Presented at Pedometrics 2024.
- Haluschak, P. (2006). Laboratory methods of soil analysis. *Canada-Manitoba Soil Survey*, 1–134.
- Jenny, H. (1994). *Factors of soil formation: A system of quantitative pedology* (p. 281). Dover.
- 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. <https://doi.org/10.1016/j.chemolab.2009.04.005>
- Ji, W., Adamchuk, V. I., Biswas, A., Dhawale, N. M., Sudarsan, B., Zhang, Y., Viscarra Rossel, R. A., & Shi, Z. (2016). Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. *Biosystems Engineering*, 152, 14–27. <https://doi.org/10.1016/j.biosystemseng.2016.06.005>
- Johnson, M. R., Anhaeusser, C. R., & Thomas, R. J. (Eds.). (2006). *The geology of south africa*. Council for Geoscience. <http://www.gssa.org.za/>
- Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11, 137–148.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and ModelSelection. <http://robotics.stanford.edu/~ronnyk>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
- Lal, R. (2006). Enhancing crop yields in the developing countries through restoration of the soil organic carbon pool in agricultural lands. *Land Degradation and Development*, 17(2), 197–209.
- Liu, L., Ji, M., & Buchroithner, M. (2017). Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra. *Remote Sensing*, 9(12), 1299.
- LTSS (Land Type Survey Staff). (2021). *Land types of South Africa: Digital Map (1:250 000 scale) and soil inventory databases*. Pretoria: Institute for Soil, Climate and Water, Agricultural Research Council. <https://www.arc.agric.za/arc-iscw/Pages/ARC-ISCW-Homepage.aspx>
- Malone, B., Minasny, B., & Mcbratney, A. B. (2017). Progress in Soil Science Using R for Digital Soil Mapping. <http://www.springer.com/series/8746>
- Mathadeen, P., Miles, N., & Ad, M. (2013). *Infrared Reflectance Spectroscopy for the Rapid Measurement of Agronomically Important Soil Properties*. December 2015, 108–113.
- Max Kuhn, M. (2021). Package “cubist” title rule-and instance-based regression modeling.
- 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. <https://doi.org/10.1097/01.ss.0000187377.84391.54>
- Mevik, B.-H., & Wehrens, R. (2015). Introduction to the pls Package. *Help Section of the “Pls” Package of RStudio Software*, Section 7, 18, 1–23.
- Minasny, B., & Mcbratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32, 1378–1388. <https://doi.org/10.1016/j.cageo.2005.12.009>
- Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, 264, 301–311. <https://doi.org/10.1016/j.geoderma.2015.07.017>
- Nocita, M., Kooistra, L., Bachmann, M., Müller, A., Powell, M., & Weel, S. (2011). Predictions of soil surface and topsoil organic carbon 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. <https://linkinghub.elsevier.com/retrieve/pii/S0016706111002850>
- Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. B., Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., Demattè, J. A. M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., ... Wetterlind, J. (2015). Soil spectroscopy: An alternative to wet chemistry for

- soil monitoring. *Advances in Agronomy*, 132, 139–159. <https://doi.org/10.1016/BS.AGRON.2015.02.002>
- Nocita, M., Stevens, A., van Wesemael, B., Brown, D. J., Shepherd, K. D., Towett, E., Vargas, R., & Montanarella, L. (2015). Soil spectroscopy: An opportunity to be seized. *Global Change Biology*, 21(1), 10–11. <https://doi.org/10.1111/gcb.12632>
- Peng, J., Ji, W., Ma, Z., Li, S., Chen, S., Zhou, L., & Shi, Z. (2016). Predicting total dissolved salts and soluble ion concentrations in agricultural soils using portable visible near-infrared and mid-infrared spectrometers. *Biosystems Engineering*, 152, 94–103. <https://doi.org/10.1016/j.biosystemseng.2016.04.015>
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/>
- Safanelli, J. L., Hengl, T., Parente, L., Minarik, R., Bloom, D. E., Todd-Brown, K., Gholizadeh, A., Mendes, W. d. S., & Sanderman, J. (2023). Open soil spectral library (OSSL): Building reproducible soil calibration models through open development and community engagement. *BioRxiv*, 2023. 572011. <https://doi.org/10.1101/2023.12.16.572011>
- Schmaltz, R. (2017). *What is precision agriculture?* (pp. 1–13). University of Sydney. <https://precision-agriculture.sydney.edu.au/home/what-is-precision-agriculture/>
- 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. <https://doi.org/10.1080/02571862.2023.2180098>
- 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. <https://doi.org/10.2136/sssaj2002.9880>
- Sila, A. M., Shepherd, K. D., & Pokhariyal, G. P. (2016). Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. *Chemometrics and Intelligent Laboratory Systems*, 153, 92–105. <https://doi.org/10.1016/j.chemolab.2016.02.013>
- Smith, B. C. (1999). *Infrared spectral interpretation: A systematic approach*. CRC press. <http://lib.ugent.be/catalog/rug01:000469608>
- Soil and Plant Analysis Council, Inc. (2000). *Soil analysis handbook of reference methods* (4th ed.). CRC Press LLC.
- Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., MacDonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews*, 49(2), 139–186. <https://doi.org/10.1080/05704928.2013.811081>
- Statistics South Africa. (2022). The world at 8 billion. <https://www.statssa.gov.za/?p=15918>
- Stevens, A., & Ramirez-Lopez, L. (2022). An introduction to the prospectr package.
- FERTASA (The Fertilizer Society of South Africa). (2010). In G. van der Linde, P. Goede, P. Haumann, R. Rabie, H. Swart, & J. J. Bornman (Eds.), *FSSA fertilizer handbook sixth*. FERTASA.
- Thomasson, J. A., Sui, R., Cox, M. S., & Al-Rajehy, A. (2001). Soil reflectance sensing for determining soil properties in precision agriculture. *Transactions of the American Society of Agricultural Engineers*, 44(6), 1445–1453. <https://doi.org/10.13031/2013.7002>
- Tilman, D., Balzer, C., Hill, J., & Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), 20260–20264. <https://doi.org/10.1073/pnas.1116437108>
- 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. <https://doi.org/10.1080/00103620600817382>
- Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 131, 59–75. <https://doi.org/10.1016/j.geoderma.2005.03.007>
- Vona, V., Sarjant, S., Tomczyk, B., Vona, M., Kalocsai, R., Kulmány, I. M., Jakab, G., Ver, A., Milics, G., & 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. <https://doi.org/10.1007/s11119-022-09942-y>
- Wadoux, A. M. J. C., Heuvelink, G. B. M., Lark, R. M., Lagacherie, P., Bouma, J., Mulder, V. L., et al. (2021). Ten challenges for the future of pedometrics. *Geoderma*, 401, 115–155. <https://linkinghub.elsevier.com/retrieve/pii/S0016706121002354>
- Wadoux, A. M. J.-C., Malone, B., Minasny, B., Fajardo, M., & McBratney, A. B. (2021). Soil Spectral Inference with R. In *Springer International Publishing AG, Cham*. (Issue Progress in Soil Science). <https://doi.org/10.1007/978-3-030-64896-1>
- Warrick, A. W. (2002). *Soil physics companion*. CRC Press.
- 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. <https://doi.org/10.1016/j.sandf.2015.09.018>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation. <https://CRAN.R-project.org/package=dplyr>
- Wijewardane, N. K., Ge, Y., Wills, S., & Libohova Z. (2018). Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. *Soil Science Society of America Journal*, 82(3), 722–731. <https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2017.10.0361>
- Zhao, D., Wang, J., Jiang, X., Zhen, J., Miao, J., Wang, J., et al. (2022). Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. *Catena*, 1, 211.

How to cite this article: Kock, A.-L., Ramphisa-Nghondzweni, P. 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), e70014. <https://doi.org/10.1111/ejss.70014>