International Journal of Design & Nature and Ecodynamics
Vol. 21, No. 10, October, 2026, pp. 3021-3035
Journal homepage: http://iieta.org/journals/ijdne

Near Infrared Spectroscopy and Ridge Regression for Soil Fertility Prediction in Tropical Agricultural Soils

Siti Mechram*, Agus Arip Munawar

Department of Agricultural Engineering, DSAI Research Center, University of Syiah Kuala, Banda Aceh 23111, Indonesia

Corresponding Author Email: mechram@usk.ac.id

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ijdne.211014

Received: 18 July 2026
Revised: 18 September 2026
Accepted: 23 September 2026
Available online: 31 October 2026

Keywords:
near infrared spectroscopy, soil fertility, agriculture, prediction, chemometrics

ABSTRACT

Near infrared spectroscopy (NIRS) offers a rapid, non-destructive alternative to wet-chemical soil testing, but small calibration sets and the high dimensionality and collinearity of spectral predictors remain persistent obstacles to reliable models. This study evaluated Ridge regression with six spectral preprocessing strategies for six soil fertility properties: total nitrogen (N), available phosphorus (P), exchangeable potassium (K), pH, exchangeable magnesium (Mg), and exchangeable calcium (Ca) from near-infrared spectra (1000–2500 nm, 1557 wavelengths) of 40 soil samples from Aceh Besar, Indonesia. Spectra were preprocessed with Savitzky-Golay smoothing, first and second derivative transforms, standard normal variate (SNV), multiplicative scatter correction (MSC), and polynomial detrending. Preprocessing and regularization strength were optimized strictly inside cross-validation folds to prevent leakage; performance was assessed with repeated five-fold and leave-one-out cross-validation, while a full-data calibration fit visualized training set agreement only. Cross-validated performance was weak to modest across all six properties. The best-performing preprocessing strategies were MSC for N, P, and K and polynomial detrending for pH, Mg, and Ca. The corresponding cross-validated ratio of performance to deviation (RPD) values ranged from 0.95 to 1.08, and RMSE was 0.136% (N), 11.6 ppm (P), 0.475 cmol(+)/kg (K), 2.34 pH units (pH), 4.90 cmol(+)/kg (Mg), and 6.61 cmol(+)/kg (Ca). Calibration fits were markedly more optimistic, notably for K (R^2 = 0.83), confirming cross-validated statistics, not calibration fit, as the basis for all claims here. For pH, Mg, and Ca, the selected regularization parameter reached the upper limit of the search grid, suggesting limited predictive information recoverable by the evaluated models in this dataset. Every model fell in the RPD < 1.4 unreliable category, so this 40-sample calibration is not yet practically useful for soil fertility screening. The study nonetheless provides a leakage-free, reproducible cross-validation workflow and identifies N and K as the properties with the least unfavourable spectral signal. A supplementary PLSR benchmark, run under an identical nested cross-validation scheme on the same 40-sample dataset, showed that Ridge regression matched or outperformed PLSR for five of the six soil properties, with PLSR only marginally ahead for K, and RPD values for both methods remaining below the 1.4 unreliability threshold for every property. These results should be read strictly as a regional baseline derived from a very small, 40-sample calibration set, not as a validated predictive solution, and they are presented within the broader aim of linking rapid, low-cost spectral soil assessment to sustainable agricultural ecosystem management in Aceh Province.

1. INTRODUCTION

Soil fertility assessment underpins nearly every decision in modern agronomic management, from fertilizer prescription to land suitability classification and the diagnosis of degradation risk. Reliable assessment, however, still depends overwhelmingly on conventional wet-chemical analyses: Kjeldahl digestion for total nitrogen, Bray or Olsen extraction for phosphorus, ammonium-acetate extraction for exchangeable cations, and electrode-based pH determination. These procedures are accurate and well standardized, but they are also slow, reagent-intensive, and expensive to scale across the large numbers of samples that precision agriculture and digital soil mapping increasingly demand [1]. In smallholder-dominated regions such as Aceh Province, Indonesia, the cost and turnaround time of conventional laboratory analysis are a genuine barrier to routine soil testing, and there is a clear practical incentive to develop faster, cheaper, and non-destructive alternatives. Near infrared spectroscopy (NIRS) has already demonstrated this potential for other smallholder commodities in the same region, including technology adoption for cocoa quality screening among Aceh farmers [2] and rapid, non-destructive prediction of quality attributes in mango and cocoa beans [3-5], indicating that regional

infrastructure and farmer acceptance for spectral screening already exist, and that extending this approach to soil fertility testing is a natural next step for the province.

NIRS has been investigated as such an alternative for more than three decades [6]. The technique exploits the overtone and combination vibrations of C-H, O-H, and N-H bonds that occur in the 1000–2500 nm region, which are sensitive, directly or indirectly, to organic matter, moisture, and, through covariation with these constituents, several inorganic fertility indicators [6]. Because a single scan yields information relevant to multiple soil properties simultaneously, NIRS has been explored as a screening tool for soil organic carbon, texture, cation exchange capacity, and macronutrient status, with reported performance ranging from excellent to essentially uninformative depending on the property, the geographic and pedogenic scope of the calibration set, and the chemometric approach used [7, 8]. Properties with a direct spectral chromophore, such as organic-matter-related total carbon and nitrogen, are typically predicted more successfully than properties such as exchangeable cations or pH, whose relationship to the spectrum is indirect and mediated by mineralogy, clay content, and organic matter [6, 9]. Beyond its role as a chemometric screening tool, rapid spectral characterization of soil fertility sits within a broader eco-informatics agenda: soil nutrient status and cation exchange capacity are direct indicators of ecosystem nutrient cycling and soil quality, so faster and cheaper access to these indicators can support digital soil mapping, precision agriculture, and sustainable natural resource management at a scale that conventional wet chemistry cannot match [10-13]. Framed this way, the present study contributes not only a chemometric method comparison but also a step toward the kind of low-cost, spatially scalable soil monitoring that underpins sustainable agricultural ecosystem management in smallholder tropical settings.

A central methodological challenge in NIRS chemometrics is that raw spectra are corrupted by physical phenomena unrelated to chemical composition, light scattering from particle size and packing density variation, baseline drift, and instrumental noise that must be removed or suppressed before a regression model can isolate the chemically informative signal [14]. A large family of preprocessing techniques has been developed for this purpose, broadly divided into scatter-correction methods, such as standard normal variate (SNV) and multiplicative scatter correction (MSC) [15, 16], and spectral derivative methods based on Savitzky-Golay polynomial filtering [17], which remove additive and, in the case of second derivatives, linear multiplicative baseline effects while sharpening overlapping absorption features. No single preprocessing strategy is universally optimal; performance depends on the specific noise structure of the instrument and sample set, which motivates systematic, data-driven comparison rather than the a priori selection of a single default pipeline [14].

The second methodological challenge specific to NIRS calibration is the combination of high spectral dimensionality (in this study, 1557 wavelength variables) with severe multicollinearity among adjacent wavelengths and, frequently, a comparatively small number of reference samples. Ordinary least squares regression is ill-posed under these conditions, which is why partial least squares regression (PLSR) has historically dominated the soil-NIRS literature [7, 18]. Ridge regression [19] offers an alternative regularization strategy: rather than projecting the spectral matrix onto a small number of latent factors, as PLSR does, Ridge regression retains the full wavelength space but shrinks the regression coefficients toward zero in proportion to a penalty parameter, λ (or α), thereby controlling variance at the cost of a small, well-characterized bias. Ridge regression is, in theory, well suited to strongly collinear predictor sets because its L2 penalty distributes coefficient weight smoothly across correlated wavelengths rather than selecting an arbitrary subset, and because the regularization strength can be tuned efficiently via generalized (leave-one-out) cross-validation without repeated model refitting. This theoretical property motivates the choice of Ridge regression for this study. Despite this theoretical suitability, comparative applications of Ridge regression to tropical soil-NIRS datasets, evaluated jointly against multiple candidate preprocessing pipelines, remain comparatively uncommon relative to the PLSR literature [20, 21]. This study therefore benchmarks Ridge regression directly against PLSR, the dominant incumbent method, under a single, identical leakage-free nested cross-validation scheme applied to the same 40-sample dataset and the same preprocessing pipelines for both methods, so that neither method is given an information or tuning advantage (Section 2.6, Section 3.9); this fair, within-study comparison, rather than a claim of outright methodological novelty, is the study's primary methodological contribution.

A third, methodological requirement is the avoidance of data leakage during model validation. Any preprocessing step that estimates a parameter from the data, a mean reference spectrum for MSC, a polynomial baseline for detrending, or the regularization strength itself, must be estimated using only the training portion of each cross-validation split; estimating it from the full dataset before splitting artificially inflates apparent accuracy and is a documented source of overoptimistic reporting in the chemometrics literature [8, 21]. This is especially consequential for small sample sets, where the ratio of estimated parameters to independent observations is already unfavourable, and where the gap between a model's fit to its own training data (calibration performance) and its performance on unseen data (cross-validated performance) can be very large, a gap this study measures directly and reports throughout, rather than presenting calibration statistics alone. With only 40 soil samples and no independent test set logistically available, cross-validation is the only defensible basis for claims about generalization, and its internal integrity, specifically, the exclusion of validation fold information from every preprocessing and hyperparameter selection step is treated as a non-negotiable methodological requirement.

A further, and in this case unavoidable, limitation concerns the sampling structure of the dataset. The samples analysed here originate from a broader Aceh Besar soil sampling and NIRS calibration effort that produced a first version of this spectral chemical dataset [22]. While field collection in that programme spanned a defined depth range, the specific depth label of each of the 40 samples used in the present analysis is not encoded in the working spreadsheet, and no reconstruction of depth assignment from the available fields could be justified without risking exactly the kind of invented metadata that responsible spectroscopic reporting must avoid [8]. Depth-resolved analysis, a common and valuable line of inquiry in soil NIRS studies, since organic matter, moisture, and several nutrients vary systematically with depth, and this variation can interact with spectral prediction accuracy [23], is therefore explicitly not attempted here. This is stated as a scope limitation rather than concealed or approximated, consistent

with the broader chemometric norm that absent data should never be substituted with assumed values.

Against this background, the present study had three specific objectives: (i) to conduct a rigorous data audit and exploratory analysis of the Aceh Besar soil NIRS dataset, characterizing sample size, spectral quality, and the statistical distribution of six fertility properties; (ii) to compare six chemometrically motivated spectral preprocessing strategies, Savitzky–Golay smoothing, first- and second derivative transforms, SNV, MSC, and polynomial detrending as inputs to Ridge regression, using a nested, leakage free cross-validation scheme appropriate to a 40-sample dataset; and (iii) to report model performance honestly and comprehensively, including, explicitly, the difference between how well a model fits the data it was calibrated on and how well it generalizes under cross-validation and to interpret Ridge regression coefficients across wavelength only to the extent that the underlying cross-validated models carry genuine predictive signal. The study does not claim, and explicitly cautions against inferring, that NIRS Ridge calibration is broadly applicable to Aceh soils beyond the specific sample set analysed; rather, it aims to provide a transparent, reproducible methodological benchmark including its performance ceiling under these particular conditions for future soil spectroscopic work in the region.

2. MATERIALS AND METHODS

2.1 Study area

The soil samples analysed in this study were collected from the Aceh Besar district, Aceh Province, Indonesia, as part of a broader programme developing near-infrared spectroscopic calibration models for agricultural soil fertility properties in the region [22]. Aceh Besar surrounds the provincial capital, Banda Aceh, at the northern tip of Sumatra, and supports a mosaic of smallholder agricultural land uses under a humid tropical climate. The study-area description is therefore limited to the regional information available in the original data descriptor [22]. This absence of soil type, land use, sampling distribution, and climatic metadata is treated as a material constraint on the ecological representativeness of the dataset (Section 3.10).

2.2 Soil sampling

The working dataset comprised 40 discrete bulk soil samples, each associated with a single near-infrared spectrum and a corresponding set of six wet-chemistry fertility measurements. The originating sampling and calibration programme for this spectral collection covered soil depths from 0 to 30 cm [22]; however, the specific depth interval (e.g., 5 cm increment) corresponding to each of the 40 individual samples used in the present analysis is not recorded in the available spreadsheet. Consistent with the requirement not to invent sampling metadata that the data do not support, this study does not assign, infer, or approximate depth labels for individual samples, and no depth-stratified analysis is presented. This is treated throughout as an explicit and material limitation on the interpretability of the results (Section 3.10), rather than as a gap to be filled by assumption.

2.3 Laboratory analysis

Six soil fertility properties were available for all 40 samples: total nitrogen content (N, %), available phosphorus (P, ppm), exchangeable potassium (K, cmol(+)/kg), soil pH measured in KCl suspension, exchangeable magnesium (Mg, cmol(+)/kg), and exchangeable calcium (Ca, cmol(+)/kg). These properties and units match those reported for the originating Aceh Besar soil calibration dataset [22]. Detailed laboratory protocols were not included in the supplied data file; therefore, Table 1 presents the reported property definitions and units rather than specific analytical methods.

Table 1. Dataset and soil fertility property characteristics

Property	Unit	Method Basis	Number of Samples
N	%	Total N	40
P	ppm	Available P	40
K	cmol(+)/kg	Exchangeable K	40
pH	KCl (unitless)	Soil pH, KCl extraction	40
Mg	cmol(+)/kg	Exchangeable Mg	40
Ca	cmol(+)/kg	Exchangeable Ca	40

2.4 Near infrared spectroscopy analysis

Near-infrared spectra were provided for the wavelength range 999.9–2500.2 nm at 1557 nominally-spaced wavelength variables mean spectral interval ≈ 0.96 nm, range 0.39–2.41 nm, indicating a non-uniform, sensor-native wavelength grid rather than a fixed instrumental step. Six wavelength channels at the extreme low- and high-wavelength edges (three at each end) showed zero or near-zero signal in more than half of the samples and negligible variance across the sample set; these channels were judged to be detector edge artefacts rather than informative signal and were removed prior to further analysis, leaving 1551 working wavelength variables spanning 1001.1–2493.0 nm. No further wavelength selection or dimensionality reduction was applied prior to modelling, consistent with the objective of allowing Ridge regression to manage the full, high-dimensional, collinear predictor space directly.

2.5 Spectral preprocessing

Six preprocessing methods were evaluated, chosen to represent the two principal families of spectral pretreatment, scatter correction and spectral derivatives [14], with co added complementary smoothing and baseline-removal treatments commonly reported for soil spectra. Savitzky-Golay smoothing (window = 15 points, second-order polynomial, zero-order derivative) is a moving polynomial least-squares filter that reduces high-frequency instrumental noise while preserving the overall shape and area of absorption features [17]. The Savitzky-Golay first derivative (window = 17 points, second-order polynomial) removes additive baseline offsets and sharpens the boundaries of overlapping absorption bands, while the Savitzky-Golay second derivative (window = 21 points, third-order polynomial) removes both additive and linear multiplicative baseline effects and further resolves overlapping bands, at the cost of amplifying high-frequency noise, which is why a wider window and higher polynomial order were used relative to the first derivative treatment. SNV is a per-spectrum, row-wise mean centring and unit variance scaling operation that corrects multiplicative light scatter and

path length differences between samples without reference to any other spectrum in the dataset [15]. MSC instead regresses each spectrum linearly against a reference spectrum. Here, the mean spectrum of the training fold only and corrects it for the fitted multiplicative and additive offsets [16]. Finally, polynomial detrending fits and subtracts a second-order polynomial baseline from each spectrum individually, removing smooth curvilinear baseline drift unrelated to genuine absorption features.

Full parameter specifications and the chemometric rationale for each method are summarized in Table 2. Critically, any preprocessing step that estimates a parameter from the sample set in this study—only MSC, whose reference spectrum is the mean of the training fold was fitted exclusively on the training partition of each cross-validation split and then applied, without refitting, to the corresponding held-out validation samples. The remaining five preprocessing methods operate independently on each spectrum (SNV, detrending) or via a fixed convolution filter (the three Savitzky-Golay variants) and therefore carry no leakage risk in principle, but were nonetheless applied within the same fold-wise pipeline for procedural consistency.

Table 2. Near infrared spectroscopy (NIRS) spectral preprocessing methods and parameters

Code	Preprocessing Method	Parameters	Rationale
SG-Smoothing	Savitzky–Golay smoothing	window = 15, polynomial order = 2, derivative = 0	Reduces high-frequency instrumental noise while preserving band shape
SG-1st-Derivative	Savitzky–Golay 1st derivative	window = 17, polynomial order = 2, derivative = 1	Removes additive baseline offsets and enhances overlapping band resolution
SG-2nd-Derivative	Savitzky–Golay 2nd derivative	window = 21, polynomial order = 3, derivative = 2	Removes additive and linear baseline effects; further resolves overlapping absorption features
SNV	Standard Normal Variate	per-spectrum mean-centring and unit-variance scaling	Corrects multiplicative scatter and path-length effects
MSC	Multiplicative Scatter Correction	reference = mean spectrum of the training fold	Corrects multiplicative and additive scatter relative to an idealised reference spectrum
Detrend	Polynomial detrending	2nd-order polynomial baseline subtraction per spectrum	Removes smooth curvilinear baseline drift unrelated to absorption features

2.6 Ridge regression and partial least squares regression benchmark

Ridge regression estimates the coefficient vector β that minimizes the penalized residual sum of squares,

$$(\hat{\beta}_0, \beta) = \underset{\beta_0,\beta}{\arg\min} \{\|y - \beta_0 \cdot 1 - X\beta\|^2 + \alpha\|\beta\|^2\}$$

where, X is the (preprocessed and standardized) spectral matrix, y is the property vector, and $\alpha \geq 0$ is the regularization strength [19]. As $\alpha \to 0$, the estimator approaches the ordinary least-squares solution; ordinary least-squares coefficient estimates are not unique in this underdetermined setting, where the number of predictors (1551 wavelengths) greatly exceeds the number of training samples (32–39, depending on the fold); as $\alpha \to \infty$, all coefficients shrink toward zero, and the prediction converges to the sample mean of y. Because the L2 penalty shrinks correlated predictors together rather than selecting an arbitrary subset among them, Ridge regression is well matched to the strong local collinearity of adjacent NIR wavelengths. Prior to model fitting within each training fold, the preprocessed spectral matrix was standardized (zero mean, unit variance) column-wise using statistics computed on that training fold alone; the corresponding validation-fold spectra were transformed using the training-fold mean and standard deviation (SD), again to prevent leakage.

As a benchmark against Ridge regression, PLSR [18] was fitted under the identical preprocessing and outer cross-validation scheme described above, so that the two methods differ only in how they handle the collinear predictor set: Ridge via coefficient shrinkage across all 1551 wavelengths, PLSR via projection onto a small number of latent variables that maximize covariance with y. The number of latent variables was not fixed in advance; within each outer training fold, it was selected by an inner, nested 5-fold cross-validation over candidates from 1 to min(15, n_train − 2), choosing the value minimizing inner cross-validated Root-mean-square error (RMSE), so that no information from the outer validation fold influenced component selection, mirroring the leakage-free design used for Ridge's regularization strength. As with Ridge, the preprocessed spectral matrix was standardized column-wise using training-fold statistics only before PLSR fitting, with the corresponding transformation applied to the validation fold.

2.7 Cross-validation and calibration fit

With only 40 available samples, and no depth, plot, or replicate group identifier that would justify a grouped cross-validation design, a conventional single train/test split was judged inappropriate: a held-out test set of a size sufficient to be informative would leave too few samples for calibration, while a small held-out set would produce an unstable, single-draw performance estimate. Cross-validation was therefore used as the sole basis for all generalization claims made in this manuscript. An independent, geographically separate test set was not logistically possible for this 40-sample regional collection, and no plot coordinates, field identifiers, or sampling distances between samples were available to construct a spatially blocked cross-validation design; random fold assignment was therefore used, and if the 40 samples include spatially proximate or clustered locations, the reported cross-validated statistics could be optimistically biased relative to a spatially independent evaluation, a caveat now stated explicitly as part of the validation framework rather than left implicit. Two complementary cross-validation schemes were implemented:

- Repeated five-fold cross-validation (5 splits × 20 independent repeats with different random partitions, 100 fold-level evaluations in total) served as the primary evaluation scheme, chosen because it retains a reasonable per-fold training-set size (32 samples) while the repetition stabilizes the variance of the resulting performance estimate, an important consideration given the modest sample size.

- Leave-one-out cross-validation (LOOCV) was used as a robustness check on the best preprocessing model selected for each property, providing a second, maximally data-efficient estimate against which the repeated five-fold results could be compared, and as the basis for the residual diagnostics reported in Section 3.7.

For every combination of the six candidate preprocessing methods and the six soil properties, the regularization parameter α was selected using RidgeCV over a base-10 logarithmic grid of 60 candidate values spanning 0.01 to 10^7; within each outer training fold, α selection used scikit-learn's efficient analytic leave-one-out generalized cross-validation on that training fold alone, so that hyperparameter tuning never had access to the corresponding outer validation samples. This nested structure, with preprocessing parameters and α both estimated strictly within the training partition of each outer fold, was applied identically across all 36 property x preprocessing combinations.

In addition to the two cross-validation schemes, a full data calibration fit was produced for the best preprocessing model of each property: the model was refit, with its own internally selected α, on all 40 samples, and then used to predict those same 40 samples. This calibration fit is not a generalization estimate by construction; it reuses the training data as its own evaluation set and is expected to be optimistic, in some cases substantially so, and it is used in this manuscript for exactly one purpose: to visualize, via a single observed versus predicted scatter plot per property, how closely each model's fitted values track the 1:1 line under its own training data. Every quantitative performance claim about generalization (ratio of performance to deviation (RPD), cross-validated R^2, cross-validated RMSE, and the associated comparison between preprocessing methods) is based on the repeated five-fold and LOOCV schemes described above, never on the calibration fit, and the two are kept explicitly distinct throughout Section 3.

2.8 Model evaluation

For every property x preprocessing combination, cross-validated predictions were pooled across the held-out folds of each of the 20 repeated five-fold cross-validation repeats, and the following metrics were computed per repeat and then summarized (mean ± SD) across repeats:

- Coefficient of determination (R^2) = 1 − SS_residual / SS_total, computed on the pooled out-of-fold predictions for that repeat;
- Root-mean-square error (RMSE) = sqrt(mean(($\hat{y}$ − y)²));
- Mean absolute error (MAE) = mean(|$\hat{y}$ − y|);
- Bias (mean error) = mean($\hat{y}$ − y);
- RPD = SD(y) / RMSE, following Chang et al. [7];
- Ratio of performance to interquartile distance (RPIQ) = IQR(y) / RMSE, following Bellon-Maurel et al. [8], which is less sensitive than RPD to the non-normal, right-skewed distributions observed for several properties in this dataset;
- Relative RMSE (RRMSE, %) = 100 × RMSE / mean(y).

The same set of metrics (R^2, RMSE, MAE, bias, RPD) was additionally computed for the full-data calibration fit described in Section 2.7, strictly for comparison against the cross-validated figures and for the scatter plot visualization; these calibration statistics are reported alongside, and clearly labelled apart from, the cross-validated statistics wherever both appear.

Model comparison and selection of the best preprocessing method for each property used cross-validated RMSE as the primary criterion (lower is better), with cross-validated RPD as a secondary tie-breaker, rather than R^2 alone, because R^2 can be inflated by the wide dynamic range of a skewed property without reflecting genuine absolute predictive accuracy. Following the widely used, if debated, benchmark of Chang et al. [7], RPD < 1.4 is interpreted as poor/unreliable prediction, 1.4 ≤ RPD < 2.0 as fair, and RPD ≥ 2.0 as good to excellent; the same qualitative bands are applied to RPIQ following Bellon-Maurel et al. [8].

2.9 Statistical analysis

Descriptive statistics (mean, median, SD, minimum, maximum, range, coefficient of variation (CV), skewness, and excess kurtosis) were computed for all six properties. Univariate distributional normality was assessed with the Shapiro-Wilk test. Principal component analysis (PCA) of the mean-centred, trimmed spectral matrix was used to characterize overall spectral variability and to screen for multivariate spectral outliers via Hotelling's T^2 statistic (using the number of components required to explain 99% of spectral variance, with the 95% F-distribution critical value) and via Q-residuals (95th-percentile empirical threshold), following standard chemometric outlier-screening practice. Given the absence of depth or other grouping metadata, no depth-stratified statistical comparison could be conducted; this is reported as a scope limitation rather than omitted silently. All analyses were implemented in Python 3 using NumPy, SciPy, and scikit-learn [24] for regression and cross-validation, with a fixed random seed of 42 for all repeated cross-validation partitions to ensure reproducibility.

3. RESULTS AND DISCUSSION

3.1 Dataset characteristics and data audit

The working dataset comprised 40 soil samples with matched spectral (1557 raw wavelengths, 1551 after edge-artefact trimming) and reference-chemistry records; sample identifiers in the spectral and chemistry sheets matched exactly, with no missing values, no duplicated spectra, and no duplicated chemistry records detected anywhere in the dataset. The trimmed spectral range spanned 1001.1–2493.0 nm. No depth or other sample-grouping metadata accompanied the individual records (Section 2.2), and this is reflected throughout as an explicit gap rather than filled by assumption; its consequences for interpretation are discussed in Section 3.10.

3.2 Exploratory analysis of soil fertility properties

Descriptive statistics for the six fertility properties are reported in Table 3. All six properties spanned wide ranges relative to their means, with coefficients of variation from 38.2% (pH) to 96.9% (N), and pH itself ranging from strongly acidic (2.57) to strongly alkaline (11.21) across the 40 samples, a spread that is unusually wide for a single regional sampling campaign and suggests the pooled sample set may span markedly heterogeneous soil conditions. Total N (skewness = 1.25), P (0.97), K (1.44), pH (0.73), Mg (0.43), and Ca (0.69) were all right-skewed to varying degrees; Shapiro–Wilk tests rejected normality ($p < 0.05$) for all six

properties, confirming that none of the six response variables is well approximated by a Gaussian distribution at the sample sizes available. This motivated the inclusion of RPIQ, which is less sensitive to distributional skew than RPD [8], alongside RPD in all subsequent performance reporting. As discussed further in Section 3.10, this wide, heterogeneous spread, particularly in pH, is itself a plausible contributor to the weak calibration accuracy reported below, since a single linear model is poorly suited to a pooled sample that may span more than one underlying soil population.

Table 3. Descriptive statistics of soil fertility properties (n = 40)

Property	N	Mean	Median	SD	Min	Max	Range	CV%	Skewness	Excess Kurtosis
N content (%)	40	0.1482	0.0868	0.1437	0.0188	0.521	0.5022	96.9428	1.251	0.4677
P (ppm)	40	14.4907	9.765	11.457	1.68	40.92	39.24	79.0646	0.9712	-0.1613
K (cmol(+)/kg)	40	0.8805	0.69	0.5144	0.26	2.58	2.32	58.4216	1.4384	2.1417
pH (KCl)	40	5.7837	5.31	2.2121	2.57	11.21	8.64	38.2475	0.7323	-0.1226
Mg (cmol(+)/kg)	40	6.431	6.54	4.7538	0.31	18.07	17.76	73.9206	0.4336	-0.5185
Ca (cmol(+)/kg)	40	7.4532	5.545	6.4084	0.39	20.18	19.79	85.9809	0.6851	-0.9523

Note: Standard deviation (SD), coefficient of variation (CV).

Univariate outlier screening (Tukey's 1.5 × IQR rule) flagged two samples for K (IDs 36 and 40) and one for pH (ID 38); no property-level outliers were flagged for N, P, Mg, or Ca. Multivariate spectral outlier screening via PCA Hotelling's T^2 (using seven principal components, jointly explaining 99% of spectral variance) identified no sample exceeding the 95% critical threshold, indicating no spectrum was multivariately extreme relative to the overall spectral population. A Q-residual screen (95th-percentile empirical threshold) flagged two samples (IDs 12 and 31) as showing comparatively poor reconstruction by the dominant PCA subspace, consistent with modestly atypical, but not necessarily erroneous, spectral shape. Following standard chemometric practice, and in the absence of an independent basis (e.g., a documented sampling or instrumental error) for exclusion, none of the flagged samples were removed from the analysis; all 40 samples were retained throughout modelling.

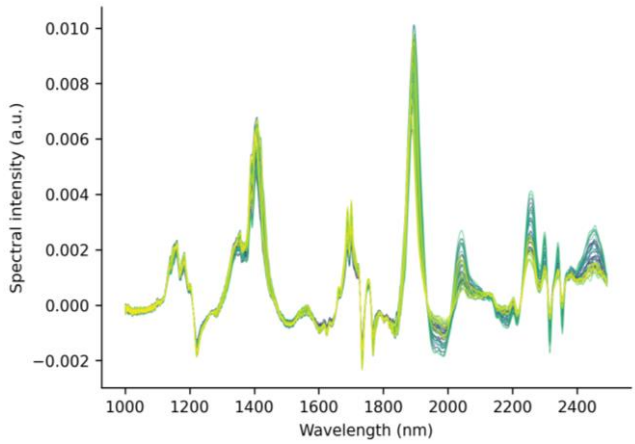

Figure 1. Raw (trimmed) near-infrared spectra of all 40 soil samples, 1001–2493 nm

3.3 Spectral characteristics

The trimmed raw spectra (Figure 1) showed the absorption pattern typical of mineral soils in this wavelength range, with the most prominent features centred near 1400 nm and 1900 nm, regions classically associated with the first overtone and combination bands of O-H stretching in structural and adsorbed water and hydroxyl bearing clay minerals together with smaller features near 1150–1250, 1700–1750, and 2200–2350 nm associated with C–H and N-H combination and overtone vibrations in organic matter. Spectral variability across the 40 samples was concentrated in a small number of dominant modes: principal component 1 explained 66.9% of total spectral variance and principal component 2 explained a further 25.6% (92.5% cumulative for the first two components), with seven components required to reach 99% cumulative variance (Figure 2). This concentration of variance in the first one or two components is consistent with the water/organic matter dominated absorption pattern visible in the raw spectra, but as discussed in Section 3.5, does not guarantee that the dominant spectral modes of variation are aligned with the specific fertility properties targeted here, and the weak cross-validated accuracy reported below suggests that, for most properties, they were not.

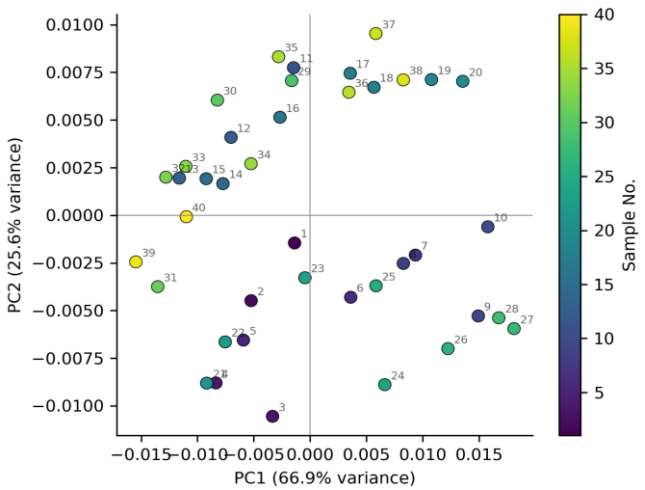

Figure 2. Principal component score plot (PC1 vs PC2) of the trimmed spectral matrix, labelled by sample number

3.4 Effect of spectral preprocessing

The six preprocessing transforms produced visually distinct spectral representations (Figure 3): Savitzky-Golay smoothing and detrending closely preserved the shape of the raw spectra while suppressing high-frequency noise or slow curvature, respectively; the first- and second derivative transforms progressively resolved the two dominant absorption features into sharper, sign-changing peak structures while amplifying noise, most visibly in the second-derivative spectra; and SNV and MSC rescaled the spectra onto comparable amplitude and baseline scales across samples, with MSC additionally removing a training-fold-referenced multiplicative offset.

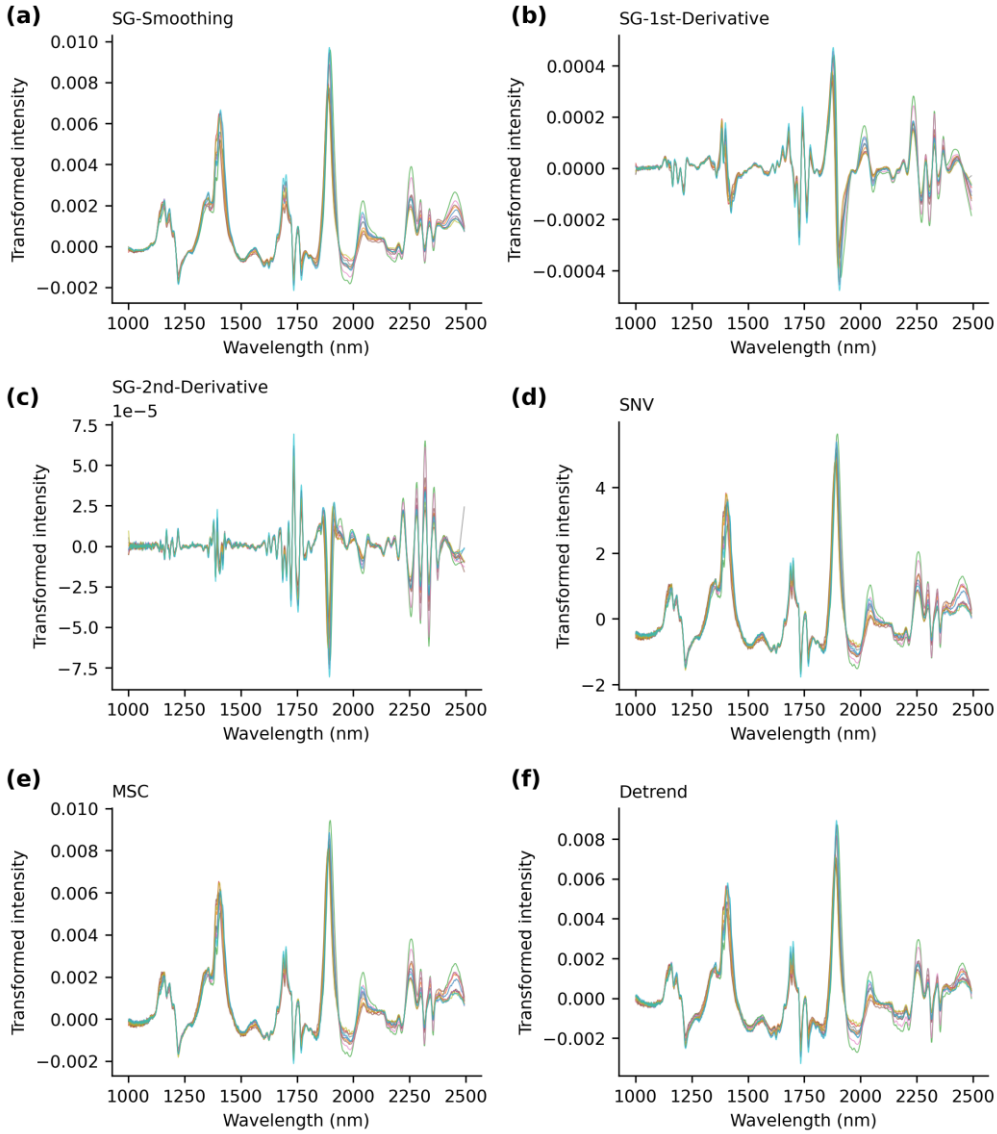

Figure 3. Representative soil spectra after six preprocessing treatments: (a) Savitzky–Golay smoothing; (b) first derivative; (c) second derivative; (d) standard normal variate (SNV); (e) multiplicative scatter correction (MSC); and (f) second-order polynomial detrending

3.5 Ridge regression performance across preprocessing methods

Full cross-validated performance for all 36 property x preprocessing combinations is reported in Table 4. Across all six properties, differences in RMSE among the six preprocessing methods were modest in absolute terms, typically spanning less than 10-15% of the best-performing model's RMSE, and the ranking of preprocessing methods differed by property. MSC yielded the lowest mean RMSE for N, P, and K. SNV ranked second for N and P, whereas SG smoothing ranked second for K; both derivative transforms also outperformed Savitzky–Golay smoothing itself for N and P, and only for K did the two derivative transforms underperform every scatter-correction and smoothing method (Table 4). For pH and Mg, second-order polynomial detrending produced the lowest mean RMSE, closely followed by the first derivative transform; for Ca, detrending was again best, followed most closely by the second derivative transform. Unlike N, P, and K, pH, Mg, and Ca achieved their lowest mean RMSE with detrending, although the rankings of the remaining preprocessing methods varied among these properties.

This split is broadly consistent with the general chemometric expectation that different pretreatments target different sources of unwanted spectral variation [14]: properties with at least a weak organic-matter-mediated spectral signal (N, in particular, given nitrogen's close biological association with soil organic matter) benefited modestly from MSC, which removes particle size and packing-related scaling differences that would otherwise compete with the comparatively subtle organic matter absorption features, whereas for pH, Mg, and Ca properties without a direct spectral chromophore and whose relationship to reflectance, where it exists at all, is mediated indirectly through mineralogy and cation exchange capacity [6, 12]. Formal significance testing between preprocessing methods was not conducted; however, the repeated five-fold RMSE SDs reported in Table 4 and plotted in Figure 4 provide an indication of the spread around each mean estimate, and in most cases this spread is comparable to, or larger than, the difference between the best and worst preprocessing method for a given property, so the preprocessing rankings above should be interpreted as indicative rather than statistically definitive.

Table 4. Cross-validated performance (repeated 5-fold CV, 20 repeats) of all 36 property × preprocessing Ridge regression models

Property	Preprocessing	R^2 (Mean)	R^2 (SD)	RMSE (Mean)	RMSE (SD)	MAE (Mean)	RPD (Mean)	RPIQ (Mean)	Bias (Mean)	Rank
Ca (cmol(+)/kg)	Detrend	-0.093	0.0428	6.6144	0.1287	5.7816	0.9692	1.6118	-0.0054	1.0
Ca (cmol(+)/kg)	SG-2nd-Derivative	-0.096	0.0445	6.6233	0.1338	5.7918	0.9679	1.6097	-0.0044	2.0
Ca (cmol(+)/kg)	SNV	-0.1299	0.1292	6.7168	0.3625	5.865	0.9564	1.5906	0.0659	3.0
Ca (cmol(+)/kg)	SG-Smoothing	-0.1589	0.1789	6.7953	0.4914	5.9079	0.9472	1.5752	0.0572	4.0
Ca (cmol(+)/kg)	SG-1st-Derivative	-0.1598	0.165	6.7999	0.4594	5.9156	0.9461	1.5734	0.1188	5.0
Ca (cmol(+)/kg)	MSC	-0.2165	0.1801	6.9615	0.5078	6.0178	0.925	1.5384	0.2073	6.0
K (cmol(+)/kg)	MSC	0.1241	0.0709	0.475	0.0191	0.3599	1.0846	1.2967	-0.0077	1.0
K (cmol(+)/kg)	SG-Smoothing	0.1089	0.0983	0.4788	0.0263	0.3601	1.0774	1.2881	-0.0027	2.0
K (cmol(+)/kg)	SNV	0.1054	0.0741	0.48	0.0197	0.3649	1.0733	1.2832	-0.0094	3.0
K (cmol(+)/kg)	Detrend	0.0953	0.0943	0.4825	0.0252	0.3692	1.0689	1.2779	-0.0016	4.0
K (cmol(+)/kg)	SG-1st-Derivative	-0.0667	0.0801	0.5242	0.0197	0.4102	0.9825	1.1747	-0.0013	5.0
K (cmol(+)/kg)	SG-2nd-Derivative	-0.0838	0.0841	0.5284	0.0204	0.4173	0.9749	1.1655	0.0007	6.0
Mg (cmol(+)/kg)	Detrend	-0.0887	0.0763	4.8951	0.1661	4.0571	0.9721	1.5945	0.0149	1.0
Mg (cmol(+)/kg)	SG-1st-Derivative	-0.0921	0.0751	4.9028	0.1638	4.0681	0.9706	1.592	0.0235	2.0
Mg (cmol(+)/kg)	MSC	-0.1086	0.0992	4.9379	0.2152	4.1031	0.9644	1.5818	0.0272	3.0
Mg (cmol(+)/kg)	SG-2nd-Derivative	-0.1105	0.0909	4.9428	0.1967	4.1179	0.9631	1.5798	0.003	4.0
Mg (cmol(+)/kg)	SNV	-0.1164	0.0856	4.9563	0.1878	4.1125	0.9604	1.5753	0.0256	5.0
Mg (cmol(+)/kg)	SG-Smoothing	-0.1303	0.1239	4.9839	0.2636	4.1339	0.9562	1.5684	0.0176	6.0
N content (%)	MSC	0.0853	0.0691	0.1356	0.005	0.1075	1.0609	1.4654	-0.0009	1.0
N content (%)	SNV	0.0792	0.0701	0.136	0.0051	0.1073	1.0574	1.4606	-0.0005	2.0
N content (%)	SG-1st-Derivative	0.0722	0.0706	0.1365	0.005	0.109	1.0533	1.455	-0.0015	3.0
N content (%)	SG-2nd-Derivative	0.0624	0.091	0.1372	0.0065	0.1092	1.0492	1.4493	-0.0018	4.0
N content (%)	Detrend	0.0558	0.0703	0.1377	0.005	0.1094	1.0442	1.4423	-0.0009	5.0
N content (%)	SG-Smoothing	0.0408	0.1115	0.1387	0.0077	0.1104	1.0383	1.4342	-0.0013	6.0
P (ppm)	MSC	-0.0467	0.126	11.555	0.6743	9.5735	0.9945	1.3939	-0.0955	1.0
P (ppm)	SNV	-0.0598	0.1501	11.6212	0.7858	9.6464	0.9898	1.3872	-0.0637	2.0
P (ppm)	SG-1st-Derivative	-0.0652	0.0943	11.6657	0.5015	9.7052	0.9837	1.3788	-0.0387	3.0
P (ppm)	SG-2nd-Derivative	-0.0945	0.098	11.8245	0.5189	9.8895	0.9706	1.3604	0.0482	4.0
P (ppm)	SG-Smoothing	-0.1223	0.1071	11.972	0.5666	9.912	0.959	1.3441	0.0037	5.0
P (ppm)	Detrend	-0.1339	0.1393	12.0257	0.7231	10.0384	0.9559	1.3397	0.0299	6.0
pH (KCl)	Detrend	-0.1522	0.1123	2.3421	0.1126	1.8699	0.9465	1.0783	0.0191	1.0
pH (KCl)	SG-1st-Derivative	-0.157	0.098	2.3475	0.0986	1.8785	0.9439	1.0752	0.006	2.0
pH (KCl)	SG-2nd-Derivative	-0.1611	0.1023	2.3515	0.1021	1.8865	0.9423	1.0735	0.0415	3.0
pH (KCl)	SG-Smoothing	-0.1832	0.1432	2.372	0.141	1.8951	0.9356	1.0658	0.0296	4.0
pH (KCl)	SNV	-0.206	0.112	2.3963	0.1118	1.9225	0.9251	1.0538	0.0243	5.0
pH (KCl)	MSC	-0.277	0.163	2.464	0.1506	1.9648	0.9007	1.0261	0.0202	6.0

Note: RMSE = root-mean-square error; MAE = mean absolute error; RPD = ratio of performance to deviation; RPIQ = ratio of performance to interquartile distance; SNV = standard normal variate; MSC = multiplicative scatter correction.

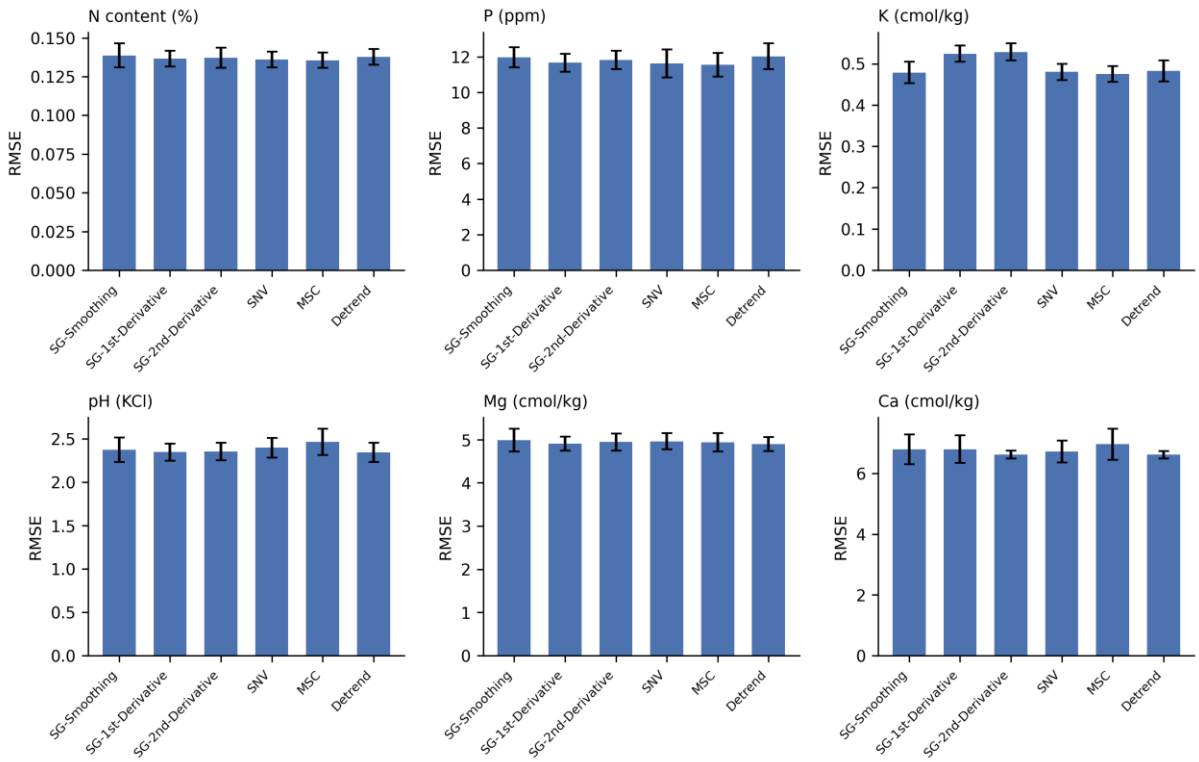

Figure 4. Cross-validated root-mean-square error (RMSE) (mean ± SD across 20 repeats of 5-fold CV) for each of the six preprocessing methods

Table 5. Best-performing model per soil property

Metric	Ca (cmol(+)/kg)	K (cmol(+)/kg)	Mg (cmol(+)/kg)	N Content (%)	P (ppm)	pH (KCl)
Best preprocessing	Detrend	MSC	Detrend	MSC	MSC	Detrend
CV R^2	-0.093	0.1241	-0.0887	0.0853	-0.0467	-0.1522
CV RMSE	6.6144	0.475	4.8951	0.1356	11.555	2.3421
CV MAE	5.7816	0.3599	4.0571	0.1075	9.5735	1.8699
CV RPD	0.9692	1.0846	0.9721	1.0609	0.9945	0.9465
CV RPIQ	1.6118	1.2967	1.5945	1.4654	1.3939	1.0783
CV Bias	-0.0054	-0.0077	0.0149	-0.0009	-0.0955	0.0191
LOOCV R^2	-0.0522	0.1737	-0.0539	0.0955	0.0025	-0.2096
LOOCV RMSE	6.4907	0.4617	4.819	0.1349	11.2987	2.4023
LOOCV RPD	0.9873	1.1141	0.9865	1.0649	1.014	0.9208
Calib R^2	0.0001	0.8273	0.0002	0.33	0.173	0.0004
Calib RMSE	6.3273	0.2111	4.6935	0.1161	10.2878	2.1839
Calib MAE	5.5455	0.1561	3.8861	0.0914	8.5926	1.7567
Calib Bias	0.0	0.0	0.0	0.0	0.0	0.0
Calib RPD	1.0128	2.4373	1.0128	1.2373	1.1136	1.0129
Final alpha	10000000.0	265.2948	10000000.0	4406.2364	12638.482	10000000.0

Note: RMSE = root-mean-square error; MAE = mean absolute error; RPD = ratio of performance to deviation; RPIQ = ratio of performance to interquartile distance; LOOCV = leave-one-out cross-validation.

Absolute predictive accuracy, however, was weak across every property and every preprocessing method. Table 5 summarizes the best preprocessing model selected for each property. Cross-validated RMSE for the best models was 0.136% for N, 11.6 ppm for P, 0.475 cmol(+)/kg for K, 2.34 pH units for pH, 4.90 cmol(+)/kg for Mg, and 6.61 cmol(+)/kg for Ca; corresponding mean cross-validated R^2 values ranged from −0.1522 for pH to 0.1241 for K, and RPD ranged from 1.08 (K) to 0.95 (pH). Following the RPD categories of Chang et al. [7], every model in this study, including the best-performing preprocessing for each property, falls in the RPD < 1.4 "poor/unreliable" band; none reached the RPD >= 1.4 threshold conventionally regarded as the floor for practically informative prediction. RPIQ values (1.08–1.61) were somewhat more favourable in absolute terms, reflecting the interquartile rather than full range spread of the skewed properties, but remained below the RPIQ >= 1.4 threshold suggested as a minimum for "fair" performance for three of six properties (pH, K, P), and only modestly above it for the remainder.

LOOCV performance for the same best-preprocessing models (Table 5) closely tracked the repeated five-fold results; R^2 differed by no more than 0.06 for any property, and RPD differed by no more than 0.05, confirming that the repeated five-fold cross-validation estimates were not an artefact of the particular fold structure used and that the weak performance reported here is a stable feature of this dataset and modelling approach rather than a validation-scheme artefact.

Read in the context of the wider soil-NIRS literature, this outcome differs from the substantially higher accuracies

reported in several published studies; for example, Chang et al. [7] reported $R^2 > 0.80$ for several soil properties using a spectral library of 802 samples spanning four Major Land Resource Areas, and more recent regional calibrations combining preprocessing optimization with partial least squares or machine learning regressors have reported RPD values well above 2 for properties such as soil organic carbon and total nitrogen [20, 25]. Comparable mid-infrared-based fertiliser recommendation calibrations built on larger regional soil libraries have likewise achieved substantially stronger predictive performance than reported here [13]. The comparison is instructive less as a shortfall of the modelling approach used here than as an illustration of how strongly soil-NIRS calibration accuracy depends on sample size and calibration-set homogeneity [6, 8]. At n = 40, with 1551 spectral predictors, the ratio of samples to predictors is extremely unfavourable even for a strongly regularized estimator such as Ridge regression, and the wide, non-normal ranges observed for several properties, including a pH span from strongly acidic to strongly alkaline, suggest the pooled sample set may itself be pedologically heterogeneous in ways not captured by any variable in the working dataset, which would further dilute any consistent spectral chemical relationship that Ridge regression could exploit.

3.6 Regularization behaviour as a diagnostic of signal strength

The Ridge regularization parameter selected on the full data calibration fit (Table 5) differed sharply between the two groups of properties identified above. For N, P, and K, alpha settled at intermediate values (4406, 12638, and 265, respectively) that meaningfully shrink but do not eliminate the fitted coefficients. For pH, Mg, and Ca, however, the selected alpha saturated at the upper bound of the 60-point search grid (10^7) in every case under cross-validation and again on the independent full data refit, indicating that, within the range explored, cross-validated error continued to improve as regularization strengthened all the way to near-total coefficient shrinkage. In practice, this means the best available Ridge model for pH, Mg, and Ca is one that predicts very close to the sample mean of each property for every soil, regardless of its spectrum, a conservative, low-variance outcome that Ridge regression correctly identifies as minimizing cross-validated error when no reliable linear spectral signal is present, rather than a failure of the optimization procedure itself.

This behaviour is methodologically informative rather than merely disappointing: it shows Ridge regression's regularization path functioning as an honest diagnostic of signal strength. For pH, Mg, and Ca, cross-validated error monotonically favoured stronger shrinkage all the way to the edge of the search grid, which is the expected behaviour of an L2-penalized estimator when the true linear relationship between predictors and response is weak or absent relative to sampling noise [19]: rather than overfitting to noise, as an unregularized or weakly regularized estimator would, Ridge regression correctly detected that the safest cross-validated prediction was close to the unconditional mean. This is a substantively different, and more informative, failure mode than a model that reports a spuriously high in-sample R^2 while generalizing poorly, and it is only visible because hyperparameter selection was nested strictly inside the cross-validation loop (Sections 2.6–2.7); had alpha been tuned on the full dataset before cross-validation, this shrinkage-to-the-mean behaviour could easily have been masked by leakage-inflated apparent accuracy.

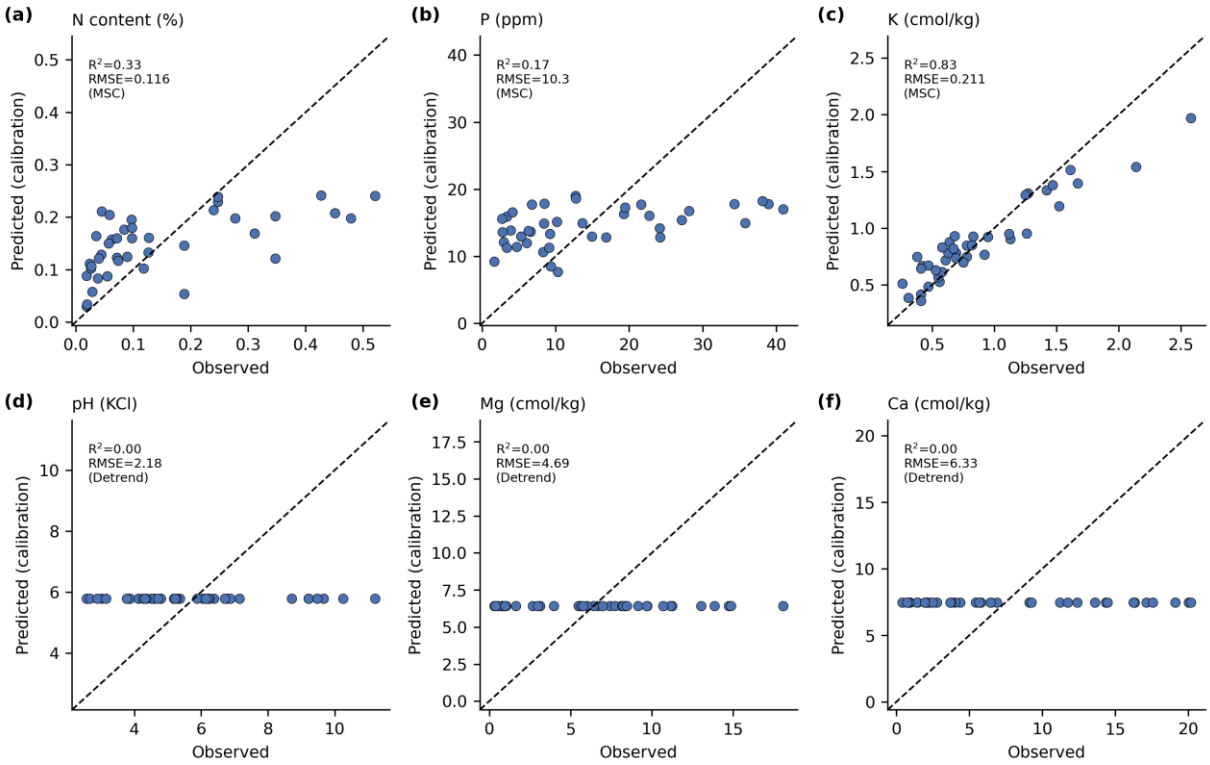

Figure 5. Observed versus predicted values from the full data calibration fit for each of the six soil fertility properties (Section 2.7). Panels: (a) N content; (b) P; (c) K; (d) pH; (e) Mg; (f) Ca

3.7 Calibration fit versus cross-validated generalization: Model diagnostics

Because the gap between a model's fit to its own training data and its performance on unseen data is central to interpreting these results correctly, Figure 5 shows observed-versus-predicted values from the full data calibration fit only (Section 2.7) for the best model of each property, and this gap is quantified explicitly here rather than left implicit. Calibration R^2 was 0.83 for K, 0.33 for N, 0.17 for P, and 0.00 for pH, Mg, and Ca; the corresponding cross-validated R^2 values (Table 5) were 0.12, 0.09, -0.05, -0.15, -0.09, and -0.09, respectively. The calibration-versus-cross-validation gap is therefore large for K in particular, a calibration R^2 of 0.83 alongside a cross-validated R^2 of 0.12, and this discrepancy is not a contradiction, but the expected signature of a high-dimensional, regularized model fit to a small sample: with 1551 correlated predictors and 40 training samples, even a well regularized estimator can trace a substantial share of in-sample variation that does not reproduce when the same samples are held out in turn. Reporting only the calibration fit, as some NIRS studies have done, would have suggested a materially more useful K model than cross-validation supports; the two are therefore presented side by side, with the cross-validated statistics treated as the only basis for any claim about predictive usefulness. For pH, Mg, and Ca, calibration R^2 of 0.00 confirms that the near-total coefficient shrinkage described in Section 3.6 holds even when the model is allowed to fit the very data it will be judged against, reinforcing that these three properties carry no detectable linear spectral signal in this sample, under either evaluation scheme.

Cross-validated residual diagnostics (LOOCV predictions; Figure 6) show no strong evidence of curvature or of a systematic funnel (heteroscedastic) pattern for any property when residuals are plotted against predicted values, and mean cross-validated bias was small in absolute terms for all six properties (|bias| <= 0.12 in the property's own units; Table 5), indicating that, to the limited extent the cross-validated models carry signal, they are not systematically over- or under-predicting on average. Cross-validation fold-to-fold variability (RMSE SD across the 20 repeats; Table 4) was proportionally larger for MSC and SNV in the pH, Mg, and Ca models (CV of RMSE up to ~8%) than for the best-performing preprocessing on those same properties, suggesting that detrending produced not only the lowest average cross-validated error but also the most stable cross-validated error for this subset of properties.

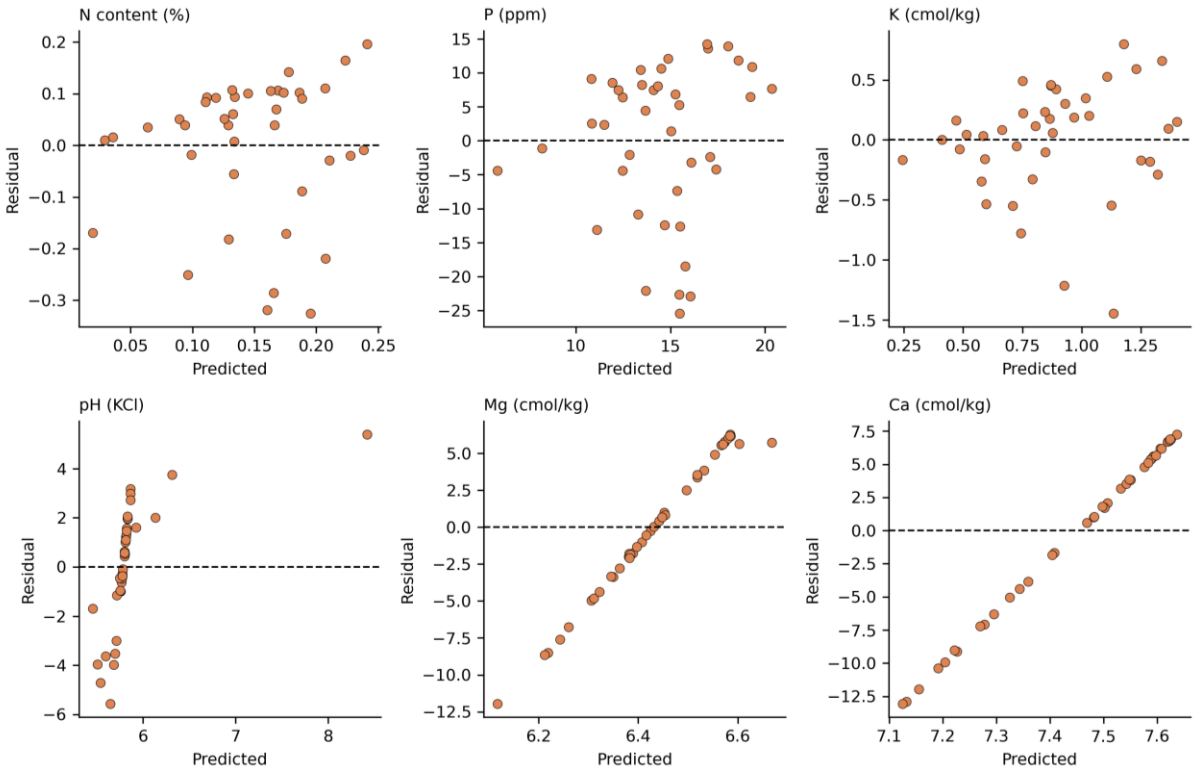

Figure 6. Leave-one-out cross-validation (LOOCV) residual (predicted minus observed) versus predicted value for the best model of each soil property

3.8 Ridge coefficient profiles and spectral regions

For the N, P, and K models, whose regularization parameters retained meaningful coefficient structure (Section 3.6), the largest magnitude Ridge coefficients concentrated in the shorter-wavelength region of the working range, broadly between 1076 and 1116 nm and, secondarily, near 1270–1290 nm and 1720–1815 nm (Table 6, Figure 7) regions associated in the general NIRS literature with combination and second overtone C-H and N-H vibrations relevant to organic matter and, indirectly, nutrient holding capacity [6]. This wavelength-to-coefficient correspondence is the explicit basis for interpreting these three models chemically; conversely, for pH, Mg, and Ca (below), the coefficient profile is deliberately not read wavelength by wavelength, since a near-null, heavily shrunk coefficient vector carries no reliable chemical information to interpret.

For pH, Mg, and Ca, whose full data models were shrunk almost to the sample mean, the corresponding coefficient profiles (Figure 7) are correspondingly close to flat across the

whole spectrum and are not interpreted here as evidence of specific informative spectral regions; presenting a wavelength-by-wavelength interpretation of a near-null coefficient vector would overstate the chemical information content actually captured by these particular models, and we deliberately avoid doing so.

Ridge regression's tendency to average across correlated wavelengths rather than select a sparse subset contrasts with PLSR's latent variable compression, historically the dominant approach in soil-NIRS chemometrics [7, 18], and with more recent machine learning approaches, random forests, gradient boosting, and neural architectures that have shown promise for larger soil spectral libraries but themselves demand larger sample sizes to avoid overfitting [20].

Table 6. Top wavelength regions ranked by |Ridge coefficient| magnitude for the best model of each property

Property	Top Wavelength Regions (|coef| Ranked)
N content (%)	1076–1076 nm; 1084–1085 nm; 1726–1726 nm; 1801–1813 nm
P (ppm)	1088–1092 nm; 1273–1273 nm; 1727–1727 nm; 1789–1789 nm
K (cmol(+)/kg)	1116–1116 nm; 1273–1273 nm; 1794–1804 nm
pH (KCl)	1085–1085 nm; 1111–1125 nm
Mg (cmol(+)/kg)	1044–1045 nm; 1080–1087 nm; 1102–1108 nm
Ca (cmol(+)/kg)	1066–1074 nm; 1095–1095 nm; 1117–1118 nm; 2233–2233 nm

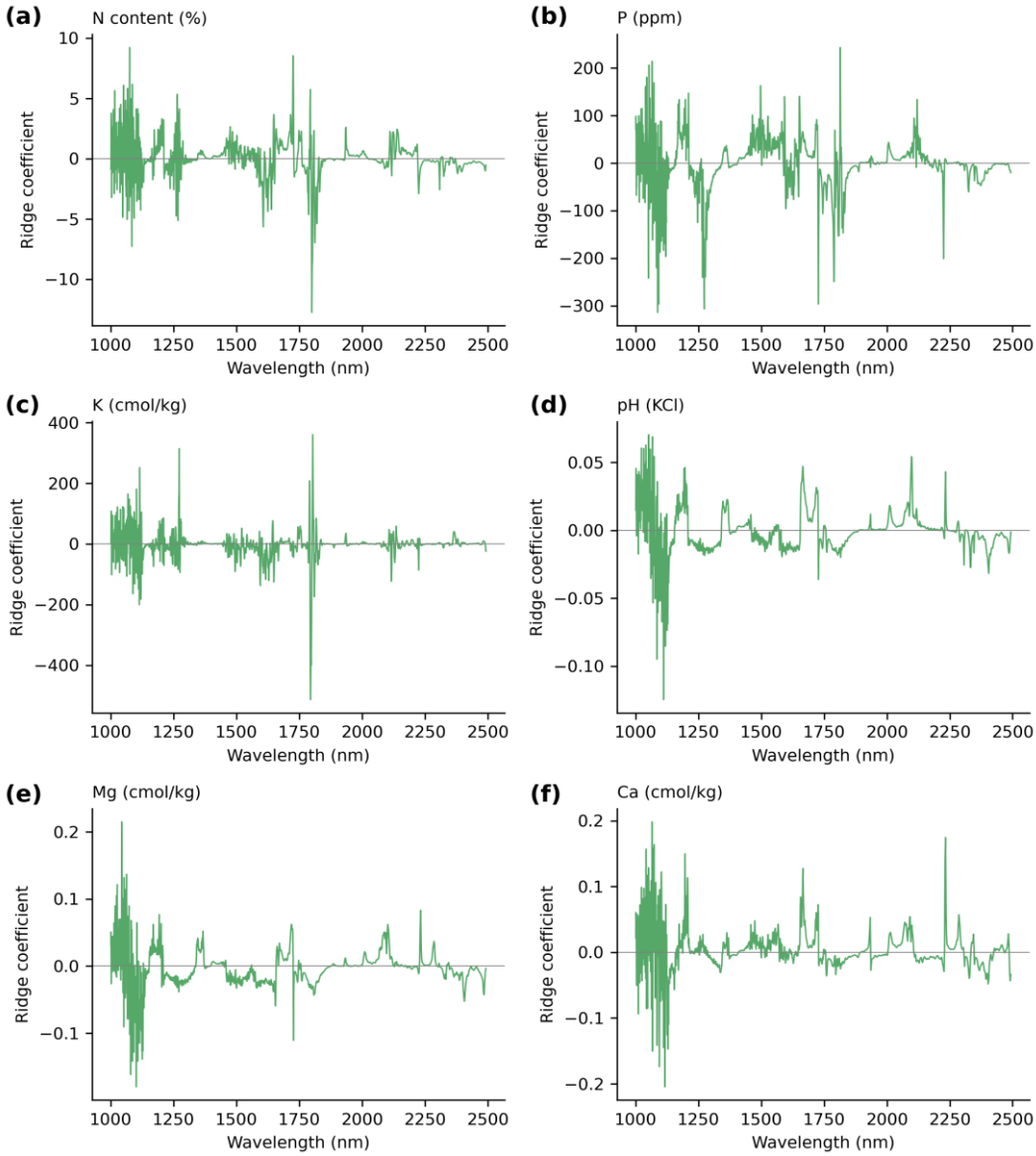

Figure 7. Ridge regression coefficient profiles across wavelength for the best full data model of each soil property. Panels: (a) N content; (b) P; (c) K; (d) pH; (e) Mg; (f) Ca

Recent methodological reviews have specifically argued that embedding soil science domain knowledge into machine learning training through informative priors, constrained loss functions, or explicit incorporation of known soil forming relationships can partially compensate for small or heterogeneous calibration sets [20, 21]; the present study did not attempt such knowledge-informed extensions, and this is a plausible avenue by which the weak cross-validated performance documented here for pH, Mg, and Ca in particular might be improved without simply enlarging the sample size, although enlarging the sample size, and reducing pedological heterogeneity within the calibration set, would

likely remain the single most effective intervention [6, 8].

3.9 Partial least squares regression benchmark comparison with Ridge regression

To directly address the absence of a within-study benchmark, a PLSR model was fitted and validated under exactly the same leakage-free, repeated 5-fold cross-validation scheme used for Ridge (Section 2.6), with PLSR latent-variable counts selected by nested inner 5-fold cross-validation within each outer training fold rather than fixed in advance, so that neither method had any information advantage over the other (Table 7).

Table 7. Ridge regression versus partial least squares regression (PLSR) under an identical leakage-free, repeated 5-fold cross-validation scheme, reporting the best preprocessing method for each model and its cross-validated root-mean-square error (RMSE), ratios of performance to deviation (RPD), and R^2

Property	Ridge: Prep.	RMSE	RPD	R^2	PLSR: Prep.	RMSE	RPD	R^2
N (%)	MSC	0.136	1.06	0.085	SG-1st-Derivative	0.136	1.06	0.074
P (ppm)	MSC	11.555	0.99	-0.047	MSC	11.640	0.99	-0.062
K (cmol(+)/kg)	MSC	0.475	1.08	0.124	Detrend	0.468	1.10	0.149
pH (KCl)	Detrend	2.342	0.95	-0.152	Detrend	2.499	0.89	-0.318
Mg (cmol(+)/kg)	Detrend	4.895	0.97	-0.089	Detrend	5.310	0.90	-0.286
Ca (cmol(+)/kg)	Detrend	6.614	0.97	-0.093	SNV	7.503	0.86	-0.412

Ridge regression matched or outperformed PLSR, measured by cross-validated RMSE, for five of the six soil properties: N, P, pH, Mg, and Ca. PLSR was only marginally ahead of Ridge for K (RMSE 0.468 versus 0.475 cmol(+)/kg; R^2 0.149 versus 0.124), the single property for which a latent-variable compression evidently captured a small additional increment of exploitable signal relative to Ridge's coefficient shrinkage. For every property and both methods, however, RPD remained below the 1.4 threshold conventionally regarded as the floor for practically useful prediction, so this benchmark should be read as establishing that Ridge's performance on this 40-sample dataset is not an artefact of an unfavourable or inequitable comparison against PLSR, rather than as evidence that either method is yet practically reliable for soil fertility screening at this sample size.

Several limitations bound the interpretation of these results and are stated explicitly rather than glossed over. First, and most consequentially, sample size. With only 40 samples, even a well-regularized model has limited power to detect genuine but modest spectral chemical relationships, and cross-validated performance estimates, while internally consistent between the repeated five-fold and LOOCV schemes used here, inevitably carry wide uncertainty at this scale; the results should be read as a rigorous estimate of what this particular sample achieves, not as a definitive statement about the achievable ceiling for NIRS prediction of these properties in Aceh soils generally.

Second, absence of depth metadata. As detailed in Section 2.2, the depth interval associated with each of the 40 samples could not be recovered from the working dataset, and the study therefore could not address depth-related variation in fertility properties, spectral characteristics, or prediction accuracy, an omission that departs from a fuller analytical brief that requested depth-stratified analysis, but one that was necessary to avoid fabricating sample-level metadata the data do not contain. Depth is known, in other soil-NIRS studies with resolved depth information, to influence both baseline fertility levels and spectral prediction accuracy [23], and its absence here should be regarded as a meaningful gap rather than a neutral omission.

Third, validation only by cross-validation, and the calibration/validation distinction. In the absence of a genuinely independent, geographically or temporally distinct test set, the cross-validation estimates reported here, however carefully implemented to avoid leakage, cannot substitute for external validation, and the reported cross-validated performance figures should not be read as evidence of transferability to soils outside the 40-sample calibration set [8]. Equally, the calibration scatter plot (Figure 5) should not be read as a validation result at all: as Section 3.7 demonstrates quantitatively for potassium, calibration fit can substantially overstate genuine predictive accuracy, and every performance claim in this manuscript beyond Figure 5's descriptive role rests on the cross-validated statistics in Tables 4 and 5.

Fourth, potential pedological heterogeneity. The unusually wide pH range observed (2.57–11.21) suggests the 40 samples may represent a more heterogeneous mixture of soil conditions than a single, spatially compact calibration set would typically contain; if so, this heterogeneity would be expected to weaken any single global linear calibration, and future work stratifying or explicitly modelling this heterogeneity may be more productive than a single pooled model. Beyond this pH range, no further site-specific geographic, land-use, or environmental metadata accompanied the working dataset; since digital soil mapping and soil health monitoring frameworks typically rely on exactly this kind of site-level environmental information to contextualize a spectral calibration [10, 11], its collection is recommended for any future extension of this work and would also help test directly for the kind of pooled-population heterogeneity suggested above.

Fifth, linear-model scope. Ridge regression is, by construction, a linear method; if the true spectral property relationships for some of these fertility indicators are substantially non-linear, a linear regularized model, however well validated, would be structurally unable to capture them, and this cannot be distinguished, from the present results alone, from a genuine absence of exploitable spectral signal.

Sixth, ecological framing of practical value. Even at the modest accuracy reported here, rapid and non-destructive spectral screening has a role to play in sustainable agricultural ecosystem management: a low-cost, high-throughput first-pass indicator, once its accuracy ceiling is improved with a larger and more representative sample set, could support more frequent soil fertility monitoring than conventional wet chemistry allows, informing fertilizer input decisions, flagging candidate sites for soil conservation intervention, and contributing observations to regional digital soil mapping and ecosystem monitoring efforts [10-13, 26]. The present results should therefore be read as a methodological step toward that goal rather than as evidence that the goal has already been

reached [27].

Finally, consistent with the scientific integrity requirements guiding this study, no result reported here has been adjusted, selectively reported, or extrapolated beyond what the cross-validated analysis directly supports; where predictive accuracy was poor, it is reported as poor, calibration statistics are never substituted for cross-validated statistics, and no external validation, depth effect, or additional accuracy is claimed beyond what this section documents.

4. CONCLUSIONS

This study conducted a comprehensive, leakage-free evaluation of Ridge regression combined with six spectral preprocessing strategies for predicting six soil fertility properties from near-infrared spectra of 40 soil samples from Aceh Besar, Indonesia. MSC produced the lowest cross-validated error for N, P, and K, while second-order polynomial detrending performed best for pH, magnesium, and calcium, but in absolute terms every resulting model fell within the poor RPD category (RPD < 1.4) established in the soil-NIRS literature, with the least unfavourable results obtained for potassium (RPD ≈ 1.08–1.11) and nitrogen (RPD ≈ 1.06). For pH, Mg, and Ca, the selected Ridge regularization parameter reached the upper bound of the parameter search space under both cross-validation and the full data calibration fit, suggesting limited predictive information recoverable from the spectra for these three properties with the models and search range evaluated here, rather than demonstrating conclusively that no such information exists; in practice, this means the optimal prediction for these three properties, given the available spectral information and sample size, is close to the sample mean rather than a spectrum-informed estimate. The calibration fit scatter plot presented here shows a substantially more optimistic picture for potassium in particular (R^2 = 0.83) than cross-validation supports (R^2 = 0.12), a discrepancy that is reported explicitly, not resolved in the model's favour, and that underscores why cross-validated, not calibration, statistics are used throughout this manuscript as the basis for any claim about predictive usefulness. A direct PLSR benchmark, fitted and validated under this identical nested cross-validation scheme, confirmed that Ridge regression matched or outperformed PLSR for five of the six soil properties, with PLSR only marginally ahead for potassium, indicating that Ridge's performance here reflects the limits of this small, heterogeneous calibration set rather than an unfavourable comparison against PLSR.

These findings do not support the practical use of the present 40-sample calibration for quantitative soil fertility screening in Aceh Besar, and they should be read as a transparent, methodologically rigorous baseline rather than as a validated field tool. They do, however, demonstrate a fully reproducible, nested cross-validation workflow with preprocessing and hyperparameter selection strictly confined to training folds, and calibration and cross-validated performance kept explicitly distinct, which is directly extensible to larger soil spectral libraries as they become available, and they identify nitrogen and potassium as the properties carrying the least unfavourable, if still weak, linear spectral signal in this particular dataset. Given the sensitivity of soil-NIRS calibration accuracy to sample size and pedological homogeneity documented throughout the wider literature, the most direct path to a practically useful calibration for this region would combine (i) a substantially larger and more systematically documented sample set, explicitly recording depth and, where feasible, geographic coordinates and soil-type classification for every scanned sample; (ii) evaluation of non-linear or domain knowledge informed modelling approaches alongside Ridge regression; and (iii) genuine external validation on soils not used in any stage of model calibration. Until such a dataset is assembled, the results reported here stand as a candid, fully reproducible benchmark of what a rigorously validated, leakage-free Ridge regression calibration achieves and does not yet achieve for Aceh Besar soil fertility prediction from near-infrared spectra. Consistent with the scope of the evidence gathered, these results should be described as a preliminary reference point for future NIRS-Ridge calibration studies in tropical soils, not as a regional predictive solution or a deployable soil monitoring system; a spatially aware validation design, alongside the larger, environmentally documented sample set described above, is recommended before any operational claim is made. Within that limited but honest scope, the workflow demonstrated here is intended as one incremental building block toward the kind of rapid, low-cost soil monitoring that can support sustainable agricultural ecosystem management in smallholder tropical regions such as Aceh Province.

ACKNOWLEDGMENTS

We sincerely acknowledge the Directorate for Research and Community Services, LPPM, Universitas Syiah Kuala, for funding this work through the *Penelitian Lektor* PL scheme 2026.

REFERENCES

[1] McBratney, A.B., Mendonça Santos, M.L., Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1-2): 3-52. https://doi.org/10.1016/S0016-7061(03)00223-4

[2] Agussabti, Rahmaddiansyah, Satriyo, P., Munawar, A.A. (2020). Data analysis on near infrared spectroscopy as a part of technology adoption for cocoa farmer in Aceh Province, Indonesia. Data in Brief, 29: 105251. https://doi.org/10.1016/j.dib.2020.105251

[3] Hayati, R., Munawar, A.A., Fachruddin, F. (2020). Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango. Data in Brief, 30: 105571. https://doi.org/10.1016/j.dib.2020.105571

[4] Hayati, R., Zulfahrizal, Munawar, A.A. (2021). Robust prediction performance of inner quality attributes in intact cocoa beans using near infrared spectroscopy and multivariate analysis. Heliyon, 7(2): e06286. https://doi.org/10.1016/j.heliyon.2021.e06286

[5] Munawar, A.A., Kusumiyati, Wahyuni, D. (2019). Near infrared spectroscopic data for rapid and simultaneous prediction of quality attributes in intact mango fruits. Data in Brief, 27: 104789. https://doi.org/10.1016/j.dib.2019.104789

[6] Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. Advances in Agronomy, 107: 163-215. https://doi.org/10.1016/S0065-2113(10)07005-7

[7] Chang, C.W., 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. https://doi.org/10.2136/sssaj2001.652480x

[8] 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. TrAC Trends in Analytical Chemistry, 29(9): 1073-1081. https://doi.org/10.1016/j.trac.2010.05.006

[9] Du, C.W., Zhou, J.M. (2009). Evaluation of soil fertility using infrared spectroscopy: A review. Environmental Chemistry Letters, 7: 97-113. https://doi.org/10.1007/s10311-008-0166-x

[10] Barra, I., Haefele, S.M., Sakrabani, R., Kebede, F. (2021). Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. TrAC Trends in Analytical Chemistry, 135: 116166. https://doi.org/10.1016/j.trac.2020.116166

[11] Bostani, A., Mohebbi Tafreshi, A., Bijeh Keshavarzi, M.H. (2025). Assessment of soil fertility and nutrient distribution for enhanced soil health and field management through an innovative approach. Agrosystems, Geosciences & Environment, 8(2): e70088. https://doi.org/10.1002/agg2.70088

[12] Ćirić, V., Prekop, N., Šeremešić, S., et al. (2023). The implication of cation exchange capacity (CEC) assessment for soil quality management and improvement. Agriculture and Forestry, 69(4): 113-133. https://doi.org/10.17707/AgricultForest.69.4.08

[13] Jeon, S.H., Jang, H.J., Ng, W., et al. (2024). Predicting soil properties for fertiliser recommendation in South Korea using MIR spectroscopy. Geoderma Regional, 39: e00901. https://doi.org/10.1016/j.geodrs.2024.e00901

[14] Rinnan, Å., van den Berg, F., Engelsen, S.B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry, 28(10): 1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

[15] Barnes, R.J., Dhanoa, M.S., Lister, S.J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy, 43(5): 772-777. https://doi.org/10.1366/0003702894202201

[16] Geladi, P., MacDougall, D., Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. Applied Spectroscopy, 39(3): 491-500. https://doi.org/10.1366/0003702854248656

[17] Savitzky, A., Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8): 1627-1639. https://doi.org/10.1021/ac60214a047

[18] Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2): 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1

[19] Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1): 55-67. https://doi.org/10.1080/00401706.1970.10488634

[20] Ma, Y.X., Minasny, B., Demattê, J.A.M., McBratney, A.B. (2023). Incorporating soil knowledge into machine-learning prediction of soil properties from soil spectra. European Journal of Soil Science, 74(6): e13438. https://doi.org/10.1111/ejss.13438

[21] Minasny, B., Bandai, T., Ghezzehei, T.A., et al. (2024). Soil science-informed machine learning. Geoderma, 452: 117094. https://doi.org/10.1016/j.geoderma.2024.117094

[22] Munawar, A.A., Yunus, Y., Devianti, Satriyo, P. (2020). Calibration models database of near infrared spectroscopy to predict agricultural soil fertility properties. Data in Brief, 30: 105469. https://doi.org/10.1016/j.dib.2020.105469

[23] Shi, Z., Yin, J.X., Li, B.G., et al. (2023). Comparison of depth-specific prediction of soil properties: MIR vs. Vis-NIR spectroscopy. Sensors, 23(13): 5967. https://doi.org/10.3390/s23135967

[24] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12: 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html?utm_source=chatgpt.com.

[25] Metzger, K., Liebisch, F., Herrera, J.M., Guillaume, T., Walder, F., Bragazza, L. (2024). The use of visible and near-infrared spectroscopy for in-situ characterization of agricultural soil fertility: A proposition of best practice by comparing scanning positions and spectrometers. Soil Use and Management, 40(1): e12952. https://doi.org/10.1111/sum.12952

[26] Mechram, S., Ayu, I.W., Farni, Y. (2024). Pretreatment method standard normal variate (SNV) and baseline shift correction (BSC) on the NIRS-based soil spectrum for rapid prediction of soil nitrogen content. IOP Conference Series: Earth and Environmental Science, 1290: 012026. https://doi.org/10.1088/1755-1315/1290/1/012026

[27] Mechram, S., Rahadi, B., Kusuma, Z. (2021). Evaluation of soil fertility: A case study in Aceh Besar district. IOP Conference Series: Earth and Environmental Science, 922: 012069. https://doi.org/10.1088/1755-1315/922/1/012069