



Article

Prediction of Soil Total Nitrogen Through Vis–NIR Spectroscopy and Machine Learning: From Model Comparison to Explainability

Shengchang Huai ^{1,2} , Qingyue Zhang ¹, Yuwen Jin ¹, Shenzhong Tian ² , Yueming Chen ², Xilin Guan ² , Tao Sun ² , Shenqiang Lv ², Zichao Zhao ², Weijia Yu ³ , Ran Li ⁴, Gilles Colinet ⁵ , Changai Lu ^{1,*} and Xinhao Gao ^{2,*}

¹ Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China

² State Key Laboratory of Nutrient Use and Management, Key Laboratory of Wastes Matrix Utilization, Ministry of Agriculture and Rural Affairs, Institute of Agricultural Resources and Environment, Shandong Academy of Agricultural Sciences, Jinan 250100, China

³ Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan 250100, China

⁴ Beijing Engineering Research Centre for Ecological Restoration and Carbon Fixation of Saline–Alkaline and Desert Land, Department of Energy and Power Engineering, Tsinghua University, Beijing 100084, China

⁵ TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liege, Passage des Deportes 2, 5030 Gembloux, Belgium

* Correspondence: luchangai@caas.cn (C.L.); sdnkygxxh@163.com (X.G.); Tel.: +01-(08)-2108703 (C.L.); Tel.: +15-(00)-5310201 (X.G.)

Abstract

Rapid and cost-effective estimation of soil total nitrogen (TN) is essential for soil fertility assessment and nutrient management. However, the performance of laboratory visible–near-infrared (Vis–NIR) models is shaped not only by preprocessing and modeling strategy but also by sample preparation and the soil’s compositional background. In this study, TN prediction was evaluated using 376 topsoil samples from two contrasting datasets: Mollisols from the black-soil region of Northeast China and Ultisols from Qiyang County, Hunan Province, southern China. Spectra acquired over 350–2500 nm for three particle-size fractions were preprocessed using Savitzky–Golay smoothing combined with standard normal variate (SNV), first-derivative, or second-derivative transformations, and modeled using partial least squares regression (PLSR), support vector regression (SVR), and extreme gradient boosting (XGBoost). Model development used a 5×5 nested cross-validation followed by evaluation on a sample-grouped held-out test set. Among all combinations, XGBoost with first-derivative preprocessing on the 0.25 mm fraction produced the best performance, with test R^2 values of 0.91 for Mollisol and 0.78 for Ultisol. Shapley additive explanations (SHAP) and principal component analysis (PCA) consistently identified informative spectral regions at 430–480 and 1330–1450 nm for Mollisol and at 585–635, 820–900, and 2180–2240 nm for Ultisol. Prediction errors were larger in the sampled Ultisol dataset and increased with DCB-extractable Fe and mineral backgrounds. A second-stage log-domain residual correction incorporating ancillary soil properties further reduced the Ultisol RMSE from 0.30 to 0.27 g kg^{−1}. These findings support the 0.25 mm, first-derivative, XGBoost workflow as a robust laboratory Vis–NIR approach for TN prediction and indicate that composition-aware residual correction can improve prediction in oxide- and mineral-rich soils.

Keywords: soil total nitrogen; Vis–NIR; particle size; SHAP interpretability; XGBoost



Academic Editor: Claudio Cocozza

Received: 27 March 2026

Revised: 16 May 2026

Accepted: 18 May 2026

Published: 20 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Soil total nitrogen (TN) is a key indicator of soil fertility and provides a practical basis for nutrient management and ecosystem assessment. Conventional reference methods (e.g., dry combustion or Kjeldahl digestion) are accurate but labor-intensive and relatively costly, which limits dense, repeated monitoring when large numbers of samples must be analyzed. Visible–near-infrared (Vis–NIR) diffuse reflectance spectroscopy has therefore been widely adopted as a rapid, nondestructive approach to estimate TN alongside other soil properties, supported by extensive soil spectroscopy research and methodological syntheses [1,2]. The development of regional and global soil spectral libraries has further strengthened the empirical basis for broad-scale inference by showing that Vis–NIR spectra capture systematic variation in organic and mineral composition across diverse landscapes and climates [3]. However, TN is typically inferred indirectly through covariation with organic matter and organo-mineral associations rather than from distinct TN-specific absorption features. Vis–NIR models for TN are therefore sensitive to external measurement conditions and compositional factors, particularly moisture, mineralogical background, and particle size-related scattering, all of which can alter baseline shape and spectral contrast and can obscure weak TN-related information [4].

Over the past decade, Vis–NIR modeling has shifted from predominantly linear partial least squares regression (PLSR) calibrations toward more flexible workflows that combine scatter correction, derivative preprocessing, variable selection, and nonlinear regression methods such as SVR and tree-based ensembles [5–7]. With the availability of large and diverse spectral libraries, nonlinear learners have often shown clear advantages for several soil properties, including TN, and have improved the feasibility of large-scale modeling [8]. In parallel, model interpretation has received increasing attention; Shapley additive explanations (SHAP) provide a useful framework for summarizing wavelength contributions and supporting interpretation of model behavior [9].

Despite these advances, inter-study conclusions about best practice remain difficult to reconcile. Laboratory measurement and reporting protocols are not always harmonized across instruments and laboratories, which complicates reproducibility and comparability [10]. Sample preparation, which includes sieving and particle-size handling, can also change scattering and noise characteristics and, thereby, alter the apparent benefit of specific preprocessing choices or algorithms [11]. Moreover, model accuracy is also strongly dependent on the validation design: when hyperparameter tuning and performance assessment are not strictly separated, cross-validation can yield optimistic error estimates. Nested cross-validation with strict separation of tuning and performance assessment is therefore needed to reduce optimistic bias, while a held-out test set provides an external check on out-of-sample performance [12,13]. Together, these methodological issues make it difficult to identify robust choices among sieve fraction, preprocessing strategy, and model type within a single leakage-controlled evaluation framework.

From a spectroscopic and pedological perspective, the challenge is compounded because TN is typically inferred indirectly through covariation with organic matter and organo-mineral associations rather than through distinct TN-specific absorption features. Moisture produces strong absorption features around 1400 and 1900 nm and can dominate spectral variance and model behavior if not controlled [14]. Ferric oxides generate strong visible-range color effects and baseline shifts, while dithionite–citrate–bicarbonate (DCB)-extractable Fe and Al (Fe_d and Al_d) and quartz-rich, oxide-coated matrices may enhance non-selective scattering and further obscure subtle TN-related signals [15,16]. Yet evidence remains limited on how measured ancillary soil properties and mineral background structure systematically influence residual variation across contrasting soil groups,

and on whether composition-informed correction can reduce prediction error under a leakage-controlled evaluation [17].

Accordingly, this study investigates how soil compositional background influences laboratory Vis–NIR prediction of TN using two regionally contrasting Chinese soil datasets: Mollisols from the black-soil region of Northeast China and Ultisols from Qiyang County, Hunan Province, southern China. Within a leakage-controlled framework based on 5×5 nested cross-validation and a sample-grouped held-out test set, we compare sieve fraction, spectral preprocessing, and model type, with the model and preprocessing comparisons serving as a benchmarking basis for evaluating particle-size and soil-background effects. Dominant wavelength regions and systematic residual variation are then examined using SHAP and principal component analysis (PCA), with reference to organic matter, Fe_d, Al_d, and mineral background. Particular attention is given to whether composition-aware residual correction can reduce prediction error in the sampled oxide-rich Ultisol dataset. Therefore, the main contribution is placed on the comparison of sieve fractions, the influence of contrasting soil backgrounds, and the proposed composition-aware residual correction strategy.

2. Materials and Methods

2.1. Study Regions, Sampling, and Reference Measurements

This study investigated two regionally sampled and compositionally contrasting soil groups, classified as Mollisols and Ultisols. The Mollisol samples were collected from cropland sites in the black-soil region of Northeast China, including Beian, Keshan, Hailun, Lishu, Gongzhuling, and Yushu, whereas the Ultisol samples were collected from cropland sites in Qiyang County, Hunan Province, southern China (Figure 1). These two soil groups were selected to provide strong pedogenic and compositional contrasts relevant to laboratory visible–near-infrared (Vis–NIR) prediction of soil nitrogen, particularly contrasts in organic matter status, oxide abundance, and mineral background. Soil orders were assigned following the USDA Soil Taxonomy and supported by field soil profile descriptions [18]. The Mollisol classification was supported by a dark, organic-rich surface horizon and high fertility typical of black soils in Northeast China, whereas the Ultisol classification was supported by the acidic, strongly weathered red-soil background with clayey texture and enrichment of Fe and Al oxides in Qiyang.

The two regions differ markedly in hydrothermal conditions and parent materials. The Northeast China black-soil sites are located in a cool temperate monsoon region, with low to rolling cropland landscapes, a mean annual temperature generally around 4–5 °C in representative Mollisol areas, and annual precipitation of about 450–600 mm. The Qiyang sites are located in a subtropical monsoon region of southern China, with an elevation of approximately 120 m at the red-soil experimental area, a mean annual temperature of about 18 °C, annual precipitation of about 1250 mm, and about 300 frost-free days. The current land use at the sampling sites was cropland, and samples were collected after the maize (*Zea mays* L.) harvest.

In 2021, after the maize harvest, a total of 376 cropland topsoil samples (0–20 cm) were collected, including 188 Mollisols and 188 Ultisols. Mollisol sampling targeted six counties or cities, namely, Beian ($n = 32$), Keshan ($n = 31$), Hailun ($n = 30$), Lishu ($n = 31$), Gongzhuling ($n = 33$), and Yushu ($n = 31$), spanning a north-to-south transition from thicker to thinner black-soil horizons that is typically accompanied by decreasing organic matter. The Mollisol parent materials are mainly Quaternary loess-like, fluvial, and lacustrine sediments, whereas the Qiyang Ultisols are mainly developed from Quaternary red clay. Ultisol samples were collected within the administrative boundary of Qiyang County in subtropical China. At each sampling site, five soil cores were collected within a 10 m × 10 m

plot as field subsamples and composited to form one bulk sample. The composite bulk sample, rather than the individual cores, was used as the experimental unit in subsequent analyses. Samples were air-dried, gently disaggregated, homogenized, and sieved to obtain three sieve fractions, namely, <2.00, <0.25, and <0.15 mm. Unless otherwise stated, reference analyses were conducted on the <0.25 mm subsample. The three sieve preparations were generated from the same homogenized bulk sample, and a single reference value from the corresponding <0.25 mm subsample was linked to the alternative spectral preparations in order to isolate sample-preparation effects on spectral prediction. Unless otherwise stated, all subsequent analyses were conducted separately for Mollisols and Ultisols.

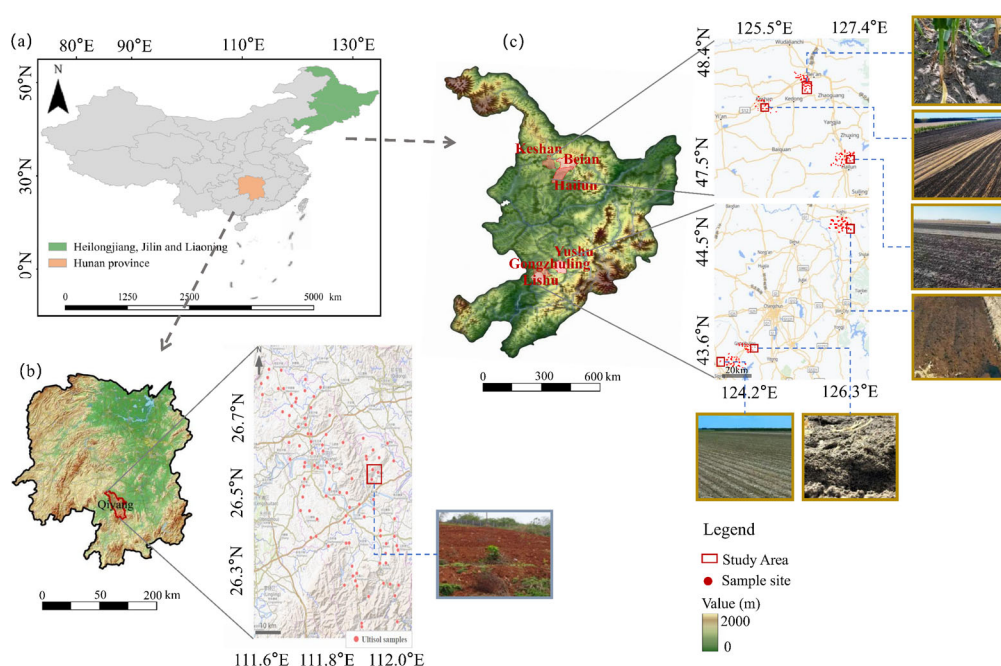


Figure 1. Study regions and sampling locations for topsoil samples collected at 0 to 20 cm depth across two contrasting pedogenic settings in China. (a) Location of the Northeast China black soil belt, including Heilongjiang, Jilin, and Liaoning, and the location of Hunan Province. (b) Ultisol sampling sites within the administrative boundary of Qiyang County in Hunan Province. A total of 188 sites were sampled. (c) Mollisol sampling sites across six counties and cities in Northeast China, including Beian, Keshan, Hailun, Lishu, Gongzhuling, and Yushu. A total of 188 sites were sampled. Elevation is shown as the background.

Reference measurements included soil total nitrogen (TN), organic carbon (OC), estimated organic matter (OM), dithionite–citrate–bicarbonate (DCB)–extractable Fe and Al (Fe_d and Al_d), and residual mineral composition. TN, operationally defined here by the Kjeldahl procedure, was determined using a Kjeldahl nitrogen analyzer (Kjeltec 8400, FOSS, Hillerød, Denmark) following the Kjeldahl method. OC was measured by dry combustion and, where required for subsequent analyses, converted to OM using the Van Bemmelen factor ($OM = 1.724 \times OC$) [19]. Residual mineral composition was quantified by powder X-ray diffraction (XRD) after sequential removal of OM, carbonates, and free Fe and Al oxides using H₂O₂, dilute acid, and the DCB procedure, respectively [20,21]. XRD analyses were conducted with an Empyrean diffractometer (Malvern Panalytical, Almelo, The Netherlands); diffraction patterns were collected over 5–70° 2θ with a step size of 0.02° and a counting time of 1 s per step, and quantitative phase analysis was performed by Rietveld refinement in HighScore [21]. Phase totals were checked for closure on a residue basis and were typically 100 ± 3 wt%. Calcite and dolomite were removed during the dilute-acid carbonate-removal step and were excluded from the reported residual mineral composi-

tion, which was calculated on a non-carbonate residue basis to avoid carbonate-related dilution of silicate and clay mineral proportions in the subsequent mineral-background analysis [22,23]. Each analytical batch included blanks, certified reference materials, and at least 10% duplicate measurements.

2.2. Spectral Acquisition and Preprocessing

Visible–near-infrared diffuse reflectance spectra were acquired for all three sieve fractions using a full-range spectroradiometer (RS-8800, Spectral Evolution Inc., Haverhill, MA, USA; nominal range 350–2500 nm) equipped with a nadir fore-optic. Illumination was provided by a single quartz–tungsten–halogen lamp positioned at a 15° incidence angle relative to the sample normal. The probe-to-sample distance was 5 cm, yielding an approximate footprint of 10–12 mm. The acquisition geometry and the overall analytical workflow are schematically shown in Figure 2. Measurements followed recommended laboratory soil spectroscopy practice to improve consistency and comparability [10].

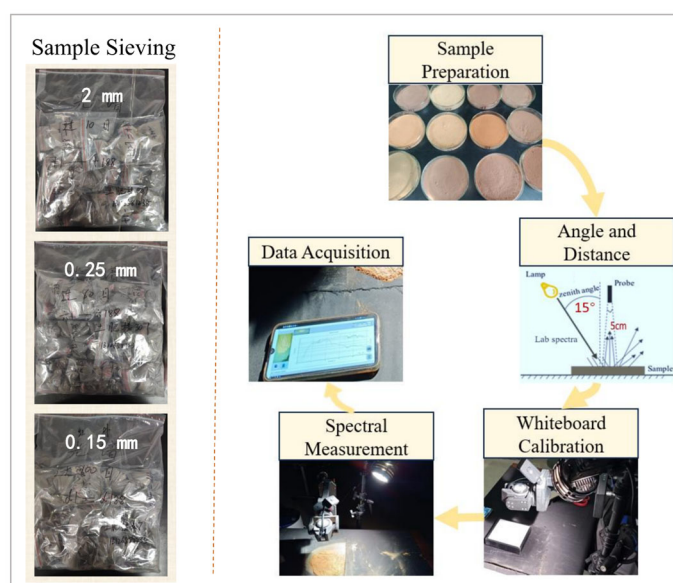


Figure 2. Workflow for laboratory Vis–NIR spectral acquisition and calibration of soil samples.

Prior to spectral measurement, soil samples were removed from the storage bags, spread into a uniform layer, and gently flattened to provide a smooth and reproducible exposed soil surface. The same sample holder, surface preparation procedure, and measurement geometry were used throughout. The instrument was warmed up for at least 30 min before measurement. Dark current was recorded at the beginning of each batch, and reflectance was calibrated against a 99% Spectralon® white reference panel. The white reference was measured at the start of each batch and every 10–15 min thereafter, and remeasured whenever the measurement geometry changed. Stray light was minimized using a matte enclosure, and laboratory temperature was maintained at 20–25 °C. For each sample and sieve fraction, spectra were recorded three times after repositioning the sample, and the final spectrum used for modeling was the arithmetic mean of the three replicate scans. Reflectance spectra were retained over the full nominal 350–2500 nm range at the instrument-exported 1 nm interval, and no additional spectral resampling was applied. All wavelength variables within this range were used for model construction after preprocessing.

Three preprocessing pipelines were compared: Savitzky–Golay smoothing followed by standard normal variate correction (SG-SNV), Savitzky–Golay first derivative (SG-1st), and Savitzky–Golay second derivative (SG-2nd) [5,24,25]. Savitzky–Golay filters used

symmetric window lengths of 11, 21, or 31 points and polynomial orders of 2 or 3 [24]. Under SG-SNV, smoothing was applied first, and SNV was then performed spectrum-wise to reduce multiplicative scatter effects [25]. Under SG-1st and SG-2nd, smoothing and derivative calculation were performed within the same parameter grid. Window length and polynomial order were selected only within the inner cross-validation loop of the training data. All preprocessing steps were fitted on the training split only and then applied to the corresponding validation and test splits to prevent information leakage.

2.3. Model Development and Hyperparameter Tuning

Model development was carried out separately for each soil group $\times$ sieve fraction $\times$ preprocessing combination. Three regression methods were benchmarked: partial least squares regression (PLSR), support vector regression with a radial basis function kernel (SVR), and extreme gradient boosting (XGBoost) [26–28]. For each soil group, samples were first split at the sample level into a development set ($n = 150$) and a held-out test set ($n = 38$) using a fixed random seed (seed = 38) and stratification by sampling site where applicable. The same sample-level split indices were reused across sieve fractions, preprocessing strategies, and models to enable paired comparison while preventing cross-fraction leakage.

Within each development set, model tuning was performed using 5×5 nested cross-validation [12,13], with a fivefold outer loop for performance estimation and a fivefold inner loop for hyperparameter selection. Because three sieve preparations were available for each sample, folds were grouped by sample ID so that spectra from the same sample could not appear simultaneously in training and validation partitions. All preprocessing steps and model-specific transformations were fitted only on the training partition of each fold.

For PLSR, predictors and response were mean-centered, and the number of latent variables was selected from 1 to 30 using the NIPALS implementation (Table S1) [26]. For SVR, predictors were standardized within each training fold, and hyperparameters were selected by grid search over the penalty parameter C , the kernel parameter γ , and ε in the ε -insensitive loss [27]. For XGBoost, hyperparameters tuned in the inner loop included the number of trees, learning rate, maximum tree depth, minimum child weight, row subsampling, column subsampling, and L1/L2 regularization [28]. Early stopping was applied within the inner loop, and the setting minimizing inner-loop RMSE was selected. When performance differences were negligible, simpler configurations with fewer latent variables, smaller C values, or shallower trees and stronger regularization were preferred to improve generalization. Data processing, statistical analysis, model development, SHAP interpretation, and visualization were conducted with Python 3.11 using NumPy 1.26.1, pandas 2.1.4, SciPy 1.11.3, scikit-learn 1.4.0, xgboost 2.0.3, shap 0.44.1, and Matplotlib 3.8.2.

2.4. Evaluation Framework and Performance Metrics

For each soil group, the held-out test set was defined at the sample level and was not accessed until final model reporting. Nested cross-validation on the corresponding development set produced one out-of-fold (OOF) prediction per sample. Within each outer fold, preprocessing and hyperparameter tuning were fitted on the training subset only, and the resulting fitted model was applied to the outer hold-out subset. Model performance was evaluated using the coefficient of determination (R^2), root-mean-squared error (RMSE), and residual prediction deviation (RPD). RMSE and R^2 were calculated as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (1)$$

where y_i and $\hat{y}_i$ are measured and predicted TN values, $\bar{y}$ is the sample mean, and n is the number of specimens. RPD was computed as $SD(TN)/RMSE$, with SD calculated within the same soil type and dataset using $ddof = 1$, i.e., the sample standard deviation with a denominator of $n - 1$, to ensure comparability across models evaluated on the same population. Nested cross-validation performance was summarized as the mean $\pm$ standard deviation across the five outer folds. Final model comparison in the main text was based primarily on the held-out test sets. Because the same sample-level split indices were reused across competing models within each soil group, the resulting test predictions were directly comparable across sieve fractions, preprocessing strategies, and algorithms. To quantify the uncertainty associated with the relatively small held-out test sets, 95% bootstrap confidence intervals were calculated for held-out R^2 , $RMSE$, and RPD by resampling paired observed and predicted TN values at the sample level, with replacement for 2000 iterations. The 2.5th and 97.5th percentiles of the bootstrap distributions were used as confidence limits. Monte Carlo-based outlier identification using squared OOF residuals was used for quality control and mechanistic diagnosis rather than for data exclusion (Figure S1).

2.5. Spectral Structure and Interpretability Analyses

To characterize broad spectral structure, principal component analysis (PCA) was performed on SG-smoothed reflectance spectra [29]. Wavelength variables were mean-centered before decomposition. PCA was used to summarize the major sources of spectral variance, visualize separation between soil groups and sieve fractions, and identify wavelength regions strongly contributing to the overall spectral structure.

Pearson correlation spectra between TN and spectral reflectance were calculated separately for Mollisols and Ultisols for each sieve fraction under the three preprocessing strategies (SG-SNV, SG-1st, and SG-2nd). Pearson coefficients were used descriptively to represent linear TN-reflectance covariance patterns, not for formal significance testing. Normality of TN and selected soil properties was assessed using the Shapiro–Wilk test. These analyses were used descriptively to compare wavelength regions associated with TN and to support the interpretation of the model-based importance patterns.

Model interpretability was further assessed using Shapley additive explanations (SHAP) for the best-performing first-stage configuration, namely, XGBoost with SG-1st preprocessing on the 0.25 mm fraction [9]. SHAP analysis was performed separately for Mollisols and Ultisols using TreeExplainer [30]. To assess the stability of wavelength importance, 100 bootstrap refits were generated within each soil group by resampling training samples with replacement. For each refit, wavelengths were ranked according to mean absolute SHAP values, and the 50 highest-ranked wavelengths were aggregated into 20-nm bins to quantify selection frequency across refits. Interpretation focused on contiguous wavelength windows rather than isolated wavelengths because adjacent spectral bands were strongly collinear.

2.6. Composition-Related Residual Analysis and Second-Stage Residual Correction

Composition-related error structure was examined for the best-performing first-stage configuration using OOF predictions from the development data. For diagnostic analyses, prediction error was expressed as absolute relative prediction error (ARE):

$$ARE_i = \frac{|y_i - \hat{y}_i^{(1)}|}{y_i} \quad (2)$$

where y_i is the observed TN value and $\hat{y}_i^{(1)}$ is the first-stage Vis–NIR prediction. Absolute relative prediction error was then examined against OM and Fe_d to assess whether prediction error covaried with soil composition.

For multivariable correction, mineral phase proportions were treated as compositional data and transformed using the centered log-ratio (CLR) transformation prior to dimensionality reduction [31]. To avoid undefined log-ratios, zero values, if present, were replaced by a small positive constant equal to half of the smallest non-zero proportion in the mineral dataset [32]. PCA was then applied to the CLR-transformed mineral composition, and the leading mineral principal components together explaining more than 80% of the total mineralogical variance were retained for downstream modeling.

A second-stage residual correction was fitted using OM, Fe_d, Al_d, and the retained mineral principal components as predictors. Before ridge regression, these ancillary predictors were standardized to zero mean and unit variance using parameters estimated from the development set, and the same scaling parameters were then applied to the held-out test set. This step was intended to model residual structure not captured by the first-stage spectral model, rather than to decompose TN prediction into independent spectral and soil property components. The first-stage residual term was defined in log space as

$$r_i = \log \left(\frac{y_i}{\hat{y}_i^{(1)}} \right) \quad (3)$$

This residual term was modeled by ridge regression [33]. To avoid information leakage, the correction model was trained only on development set OOF residuals and ancillary variables and was never fitted on the held-out test set. For the final external evaluation, the first-stage spectral model was refitted on the full development set to generate baseline predictions for the test set, and the ridge correction estimated from the development data was then applied to the test set ancillary variables. Corrected predictions were obtained as

$$\hat{y}_i^{(2)} = \hat{y}_i^{(1)} \times \exp(\hat{r}_i) \quad (4)$$

Residual trend analyses were based on OOF predictions from the development data, whereas baseline and corrected model performance were compared on the held-out test sets.

3. Results

3.1. Differences in TN and Vis–NIR Spectra Between the Mollisol and Ultisol

3.1.1. TN Distribution and Mean Vis–NIR Spectral Response

TN concentrations differed between the two datasets (Figure 3a), with the Mollisol dataset showing higher values and a broader distribution than the Ultisol dataset. Mean Vis–NIR reflectance spectra also differed, with the Mollisol dataset showing lower reflectance than the Ultisol dataset, particularly in the visible region (Figure 3b). In both datasets, reflectance decreased with decreasing particle size, and the largest shift was generally observed between the 2.00 and 0.25 mm fractions. PCA of SG-smoothed spectra separated the two datasets primarily along PC1, which accounted for 91.8% of the spectral variance, whereas PC2 accounted for 6.6%; together, the first two PCs explained 98.3% of the total variance (Figure 3c). The Ultisol dataset showed greater dispersion in the PC space. The loading patterns were concentrated mainly in the 400–800 nm region and near 1400 and 2200 nm (Figure S2).

3.1.2. TN-Spectral Correlation Patterns and Associations with Soil Components

Pearson correlation spectra between TN and reflectance were calculated for each dataset and particle-size fraction using SG-SNV, SG-1st, and SG-2nd preprocessing (Figure S3; Figure 3d). Compared with the Ultisol dataset, the Mollisol dataset showed stronger and smoother TN-spectral correlation patterns, with the clearest correlations in

the visible region (400–800 nm) and near 2200 nm. In the Ultisol dataset, the correlations were weaker and more variable across preprocessing methods. TN also showed different relationships with measured soil components in the two datasets (Table S2). In both datasets, TN was positively correlated with OM, with a stronger relationship in the Mollisol dataset ($r = 0.91$) than in the Ultisol dataset ($r = 0.70$). In the Mollisol dataset, TN was also positively associated with Al_d ($r = 0.36$) and several clay minerals, including vermiculite and kaolinite. In the Ultisol dataset, TN showed weak negative correlations with Fe_d ($r = -0.11$) and Al_d ($r = -0.13$). The TN-quartz relationship was negative in the Mollisol dataset ($r = -0.21$) and weakly positive in the Ultisol dataset ($r = 0.11$).

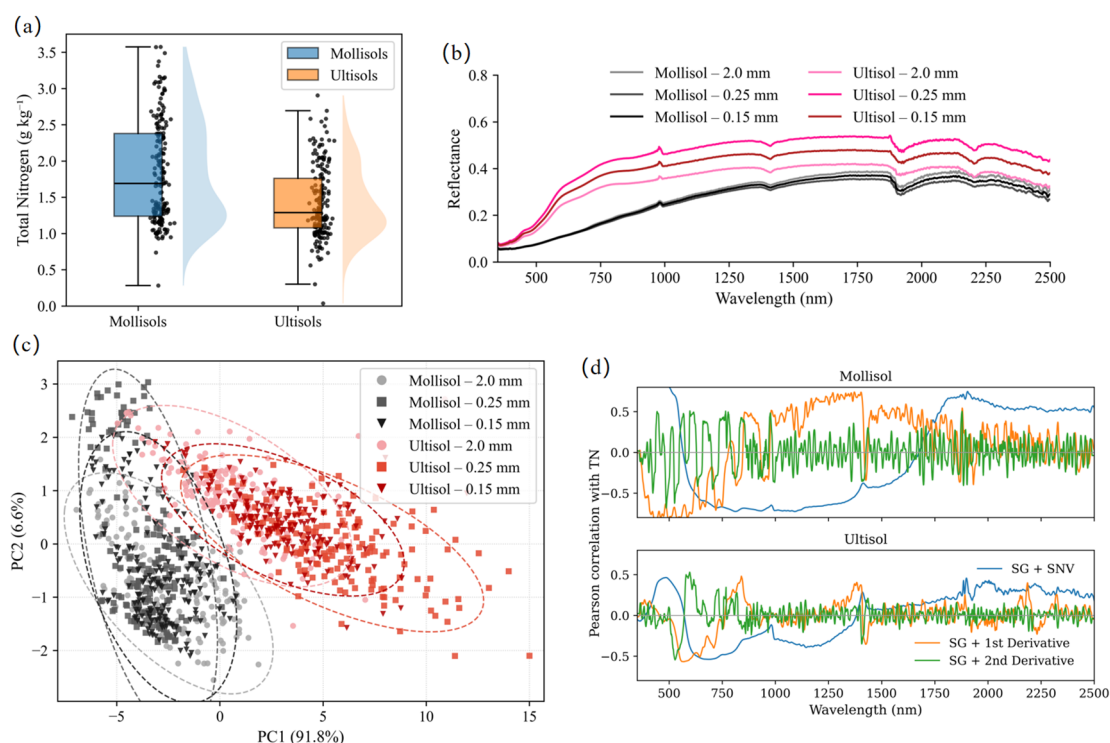


Figure 3. Differences in total nitrogen distribution and Vis–NIR spectral responses between Mollisols and Ultisols across particle-size fractions. (a) Distribution of total nitrogen (TN, g kg⁻¹) by soil type. (b) Mean Vis–NIR reflectance spectra for Mollisols and Ultisols at three particle-size fractions (2.00, 0.25, and 0.15 mm). (c) PCA score plot of SG-smoothed spectra showing separation of samples by soil type and particle size. (d) Pearson correlation between TN and spectral reflectance under three preprocessing methods (SG + SNV, SG + 1st derivative, and SG + 2nd derivative), shown separately for Mollisols and Ultisols.

3.2. Benchmark Performance Across Soil Group, Particle-Size Fraction, Preprocessing Strategy, and Algorithm

Across the held-out test sets, XGBoost achieved the highest accuracy, followed by SVR and PLSR (Figure 4). In the preprocessing comparison (Figure 4a–c), SG-1st provided the closest agreement between observed and predicted TN, whereas SG-2nd was less consistent. In the particle-size comparison under SG-1st (Figure 4d–f), the 0.25 mm fraction gave the most consistent performance across the two datasets, while the 0.15 mm fraction did not consistently improve performance over the 0.25 mm fraction. Under SG-1st + 0.25 mm (Figure 4g,h), prediction was more accurate in the Mollisol dataset than in the Ultisol dataset. For XGBoost + SG-1st + 0.25 mm, held-out test performance was $R^2 = 0.91$ and $RMSE = 0.21$ g kg⁻¹ in the Mollisol dataset, and $R^2 = 0.78$ and $RMSE = 0.30$ g kg⁻¹ in the Ultisol dataset (Table 1), with paired-bootstrap 95% confidence intervals reported in

Table S3. In the Ultisol dataset, XGBoost + SG-1st + 0.15 mm reduced RMSE to 0.27 g kg⁻¹, while R² was 0.76 (Table 1).

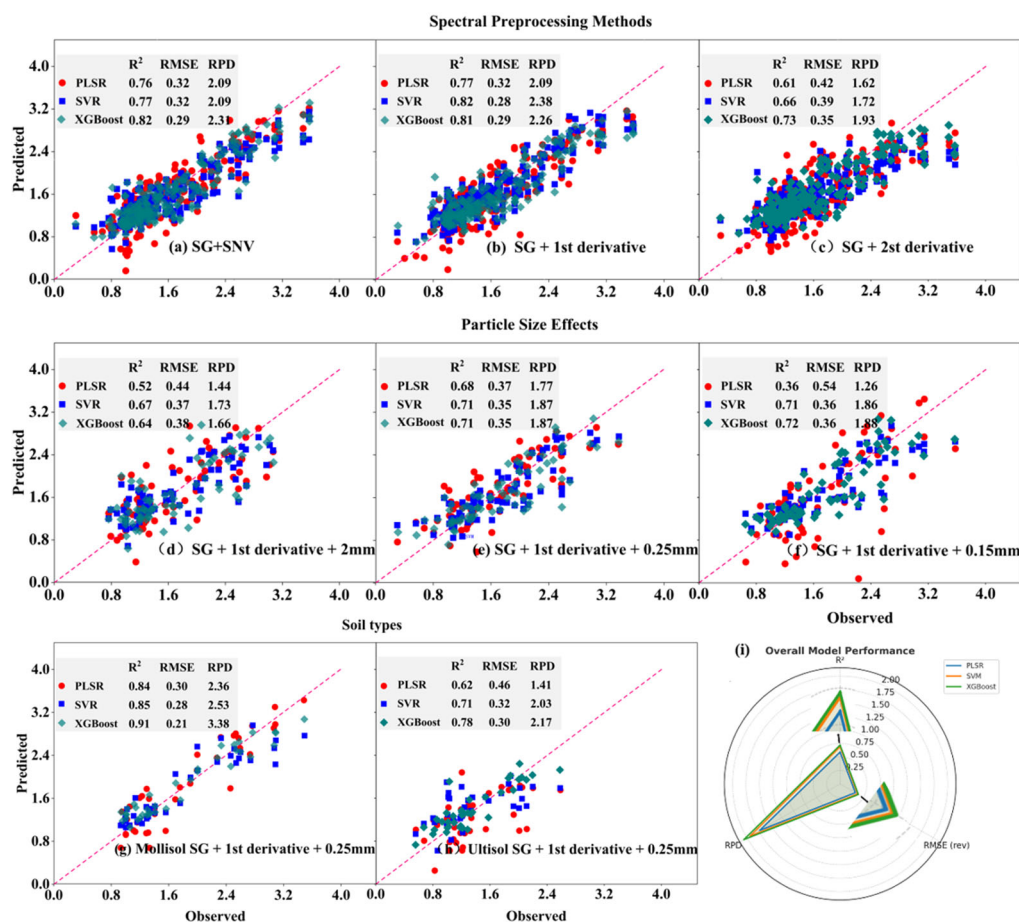


Figure 4. Held-out test performance of Vis-NIR models for predicting soil total nitrogen (TN) across preprocessing strategy, particle-size fraction, dataset, and algorithm. Observed versus held-out-test predicted TN is shown for PLSR, SVR, and XGBoost. Panels (a–c) compare preprocessing pipelines SG-SNV, SG-1st, and SG-2nd with N = 225. Panels (d–f) compare particle-size fractions 2.00, 0.25, and 0.15 mm under SG-1st with N = 75. Panels (g,h) compare the Mollisol and Ultisol datasets under SG-1st at 0.25 mm with N = 38. The dashed line indicates the 1:1 relationship, and panel annotations report R², RMSE in g kg⁻¹, and RPD. Panel (i) summarizes overall performance using a radar plot, with RMSE displayed on a reversed scale as labeled.

Table 1. Top-performing TN prediction models for the Mollisol and Ultisol datasets under 5 × 5 nested cross-validation and held-out test evaluation.

Soil Type	Preprocessing	Particle Size (mm)	Model	Nested CV (5 × 5)			Held-Out Test		
				R ²	RMSE	RPD	R ²	RMSE	RPD
Mollisol	SG-1st	0.25	XGBoost	0.93	0.20	3.79	0.91	0.21	3.38
	SG-SNV	0.25	XGBoost	0.92	0.22	3.45	0.90	0.23	3.08
	SG-1st	0.25	SVR	0.88	0.28	2.71	0.85	0.30	2.36
	SG-SNV	0.15	XGBoost	0.87	0.26	2.92	0.85	0.26	2.73
Ultisol	SG-1st	0.15	SVR	0.86	0.25	3.03	0.84	0.27	2.63
	SG-1st	0.25	XGBoost	0.80	0.29	2.24	0.78	0.30	2.17
	SG-1st	0.15	XGBoost	0.79	0.25	2.60	0.76	0.27	2.40
	SG-1st	0.25	SVR	0.72	0.28	2.23	0.71	0.32	2.03

Table 1. Cont.

Soil Type	Preprocessing	Particle Size (mm)	Model	Nested CV (5×5)			Held-Out Test		
				R ²	RMSE	RPD	R ²	RMSE	RPD
	SG-SNV	0.15	XGBoost	0.70	0.38	1.71	0.67	0.39	1.66
	SG-SNV	2.00	SVR	0.69	0.40	1.63	0.65	0.44	1.47

Entries report R², RMSE (g kg⁻¹), and RPD. Rows are ordered within each dataset by nested cross-validation performance.

3.3. SHAP-Derived Key Wavelengths and Soil Types Differences

SHAP analysis was conducted for the best-performing first-stage configuration, XGBoost with SG-1st on the 0.25 mm fraction (Figure 5). In the Mollisol dataset, high-frequency wavelengths were concentrated mainly in the 430–480 nm and 1330–1450 nm regions. In the Ultisol dataset, the main clusters occurred at 585–635 nm, 820–900 nm, and 2180–2240 nm. These wavelengths were grouped within broader regions in the visible domain, around 1400 nm, and in the 2180–2240 nm range (Figure 5c). The main SHAP clusters broadly overlapped with the PCA loading domains shown in Figure S2.

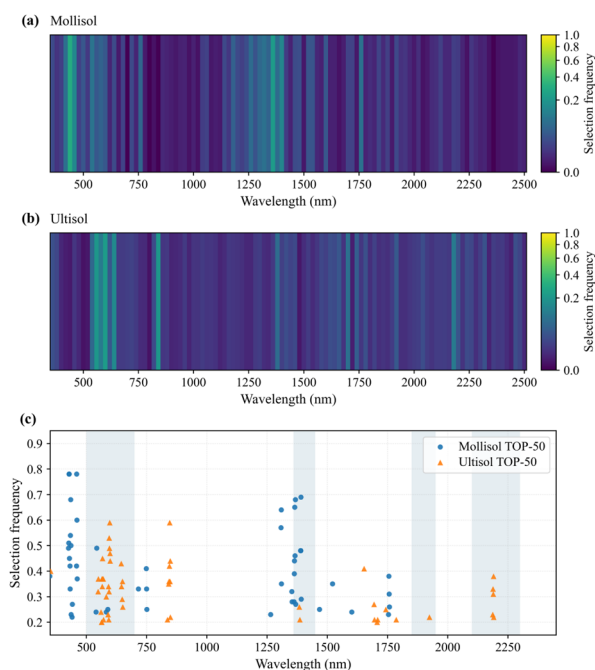


Figure 5. Stability of SHAP-selected wavelengths under the best-performing first-stage configuration using XGBoost, SG-1st, and the 0.25 mm fraction. Panels (a,b) show consensus stability heatmaps for the Mollisol and Ultisol datasets, respectively, based on 100 bootstrap refits and summarized in 20 nm bins, with color indicating selection frequency. Panel (c) compares the top 50 wavelengths selected for the two datasets. Shaded regions indicate broad wavelength windows used to summarize major clusters of selected wavelengths and are intended for visualization only, rather than as statistically derived boundaries.

3.4. Composition-Related Error Structure and Residual Correction

Using out-of-fold predictions from the best first-stage model, XGBoost + SG-1st + 0.25 mm, the absolute relative prediction error (ARE) was plotted against OM and Fe_d (Figure 6a,b). In the Mollisol dataset, ARE decreased over the intermediate OM range and then changed little at the upper end of the gradient, whereas the relationship with Fe_d was relatively weak. In the Ultisol dataset, ARE showed a clearer association with Fe_d

and remained more variable along the OM gradient. A second-stage composition-aware residual correction using OM, Fe_d, Al_d, and mineral principal components produced modest improvements in both datasets (Figure 6c,d). In the Mollisol dataset, RMSE decreased from 0.21 to 0.20 g kg⁻¹, and R² increased from 0.91 to 0.92. In the Ultisol dataset, RMSE decreased from 0.30 to 0.27 g kg⁻¹, and R² increased from 0.78 to 0.82.

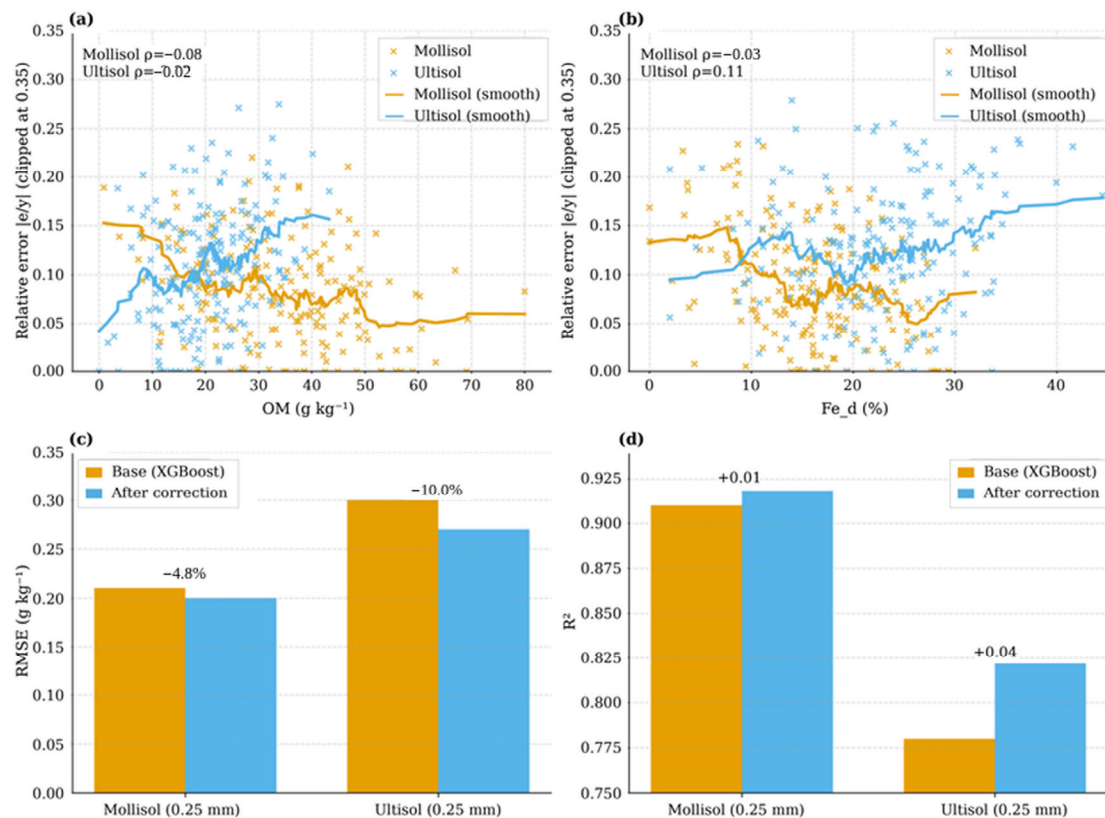


Figure 6. Composition-related error structure and composition-aware residual correction for TN prediction in the Mollisol and Ultisol datasets under the best-performing first-stage configuration (XGBoost, SG-1st, and the 0.25 mm fraction). Panels (a,b) show relationships between absolute relative prediction error (ARE) and organic matter (OM) or dithionite–citrate–bicarbonate–extractable Fe (Fe_d), respectively, based on out-of-fold predictions, with smoothed trend lines overlaid. Panels (c,d) compare baseline and corrected performance in terms of RMSE and R² for the Mollisol and Ultisol datasets, respectively.

4. Discussion

4.1. Determinants of Predictive Accuracy Across Particle-Size Fraction, Preprocessing Strategy, and Algorithm

Laboratory Vis–NIR modeling is often expected to benefit from finer sieving and derivative preprocessing because both can reduce particle-size scattering and enhance spectral contrast [7,11]. In the present study, however, the effect of particle size was conditional rather than monotonic. The combination of the 0.25 mm fraction, SG-1st preprocessing, and XGBoost gave the most consistent overall performance across the Mollisol and Ultisol datasets. Additional grinding to the 0.15 mm fraction did not yield a uniform gain, indicating that reducing particle size beyond 0.25 mm was not sufficient to overcome background effects associated with oxide- and mineral-rich matrices. From a practical perspective, the 0.25 mm fraction appears to provide a workable balance between improved sample homogeneity and avoidance of unnecessary extra grinding. A similar pattern was observed for preprocessing: SG-1st generally improved robustness, whereas SG-SNV was less effective and SG-2nd was less stable. SNV mainly corrects multiplicative

scatter and spectrum-level scaling but may suppress broad brightness differences associated with organic matter and mineral background. The first derivative, in contrast, reduces baseline offsets while emphasizing local slope changes and weak absorption features, which likely explains its better performance in this study [5].

With particle size fraction and preprocessing fixed, the ranking of XGBoost > SVR > PLSR suggests that nonlinear models captured broader spectral interactions more effectively than linear latent-variable regression. Similar advantages of ensemble and other nonlinear learners have been reported for large or compositionally diverse soil spectral datasets [8,34]. For example, Clingensmith and Grunwald [8] reported strong Vis–NIR prediction of TN across a large continental U.S. soil dataset, while Recha et al. [34] found that ensemble modeling improved prediction of several soil health indicators, including TN. Because preprocessing and hyperparameter tuning were separated within the inner loop of nested cross-validation and final performance was then checked on a held-out test set [12,13], the contrast between the two datasets is unlikely to be only a tuning artifact. Instead, the weaker held-out performance in the sampled Ultisol dataset suggests that soil compositional background constrained generalization beyond what algorithm choice alone could overcome. Given that each held-out test set contained 38 samples, these external test metrics should be interpreted together with the nested cross-validation results and the paired-bootstrap confidence intervals reported in Table S3.

4.2. Spectral Regions Supporting TN Prediction and the Limits of SHAP Interpretation

SHAP analysis for the best-performing first-stage configuration, namely, XGBoost with SG-1st on the 0.25 mm fraction, highlighted several broad predictive domains, including the visible region (430–635 nm), the NIR around 820–900 and 1330–1450 nm, and the SWIR around 2180–2240 nm. These windows are consistent with spectral domains commonly associated with color and brightness variation linked to organic matter and iron oxides, water-related absorptions near 1400 nm, and clay or organo-mineral combination bands in the SWIR [4,14–17]. This interpretation does not imply that the residual correction used independent sources of TN information; rather, it tested whether measured composition variables could explain remaining bias after the spectral model had already used their spectral proxies. This agrees with previous soil spectroscopy studies showing that TN prediction is usually indirect and depends strongly on covariation with organic matter, clay minerals, moisture status, and oxide-related spectral features rather than on distinct N-specific absorption bands. This interpretation also agrees with the PCA loadings and the correlation spectra, which showed that the dominant spectral structure differed between the two datasets rather than converging on a single TN-specific feature.

The incomplete overlap between the SHAP-selected intervals is therefore interpreted as a consequence of different indirect covariation pathways in the two soil backgrounds. In the Mollisol dataset, the stronger TN-OM association and darker organic-rich surface material likely favored the 430–480 nm and 1330–1450 nm regions, whereas in the Ultisol dataset, Fe/Al oxide abundance and a brighter mineral background shifted importance toward Fe-related visible/NIR features at 585–635 and 820–900 nm and clay/mineral combination bands around 2180–2240 nm. Because neighboring spectral bands are highly collinear, SHAP importance may also be distributed differently among adjacent proxy wavelengths in different soil domains. At the same time, the SHAP results should be interpreted as model-conditional importance patterns rather than direct evidence of TN absorption features. SHAP quantifies the contribution of predictors within a fitted model [9,30], but TN itself has weak direct Vis–NIR absorptions. Accordingly, the highlighted windows are more plausibly interpreted as domains through which TN covaries with organic matter, oxide abundance, hydration status, and mineral composition [4,17]. This caution is especially important for

water-sensitive regions around 1400 and 1900 nm because their apparent importance can vary with moisture status and measurement conditions [14]. For this reason, interpretation at the level of contiguous wavelength windows is more defensible than interpretation of isolated single wavelengths.

4.3. Soil-Type Dependence of TN Prediction in Mollisols and Ultisols

The higher predictive accuracy in the Mollisol dataset is unlikely to reflect model choice alone and instead illustrates the influence of contrasting soil backgrounds on TN prediction. Part of the difference is statistical because R^2 tends to increase more easily when the target variable spans a broader range. However, the contrast was also evident in RMSE and in the residual patterns, which indicates a substantive difference in signal-to-background conditions between the two datasets. The stronger TN-OM association in the Mollisol dataset suggests that TN prediction depended more directly on spectral variation linked to organic matter and associated organo-mineral assemblages, which is consistent with the indirect nature of Vis–NIR prediction for TN [4,17]. In the Ultisol dataset, higher Fe_d and Al_d contents together with brighter quartz-rich backgrounds likely increased albedo and introduced broader background variation, thereby weakening the indirect spectral signal of TN [15,16]. The clearer error structure along the Fe_d gradient supports this interpretation. Similar soil-background effects have been reported beyond Mollisols and Ultisols, including NIRS studies of C and N in Alfisols, Oxisols, and Ultisols [31]. The poorly predicted samples can therefore be interpreted as diagnostic cases of stronger composition-related spectral interference, particularly in the Ultisol dataset, where errors were more closely associated with Fe_d and mineral-background variation. Even so, this contrast should be framed cautiously. Because the two datasets were collected from different regional settings, the present comparison is best interpreted as a between-dataset contrast represented by Mollisols and Ultisols, rather than as a strictly soil-order effect.

4.4. Composition-Aware Correction: Gains, Limits, and Practical Implications

The proposed composition-aware residual correction represents a complementary contribution of this study and was most useful in the Ultisol dataset, where the gain in held-out performance was larger. This pattern suggests that part of the residual structure in the Ultisol dataset was composition-related and remained unresolved after the first-stage spectral model. In the Mollisol dataset, by contrast, the improvement was small, which is consistent with the idea that a greater proportion of the predictable structure was already captured by the spectral model. Because the correction step incorporated OM, Fe_d, Al_d, and mineral principal components, it should be interpreted as a hybrid ancillary-assisted model rather than as a purely spectral recalibration and should not be viewed as a replacement for the rapid first-stage Vis–NIR workflow.

The correction strategy is statistically appropriate because mineral phase data are compositional, the centered log-ratio transformation addresses the closure problem, and zero values must be handled before log-ratio analysis [32]. Its practical value, however, depends on the availability and cost of ancillary measurements, particularly Fe_d, Al_d, and mineralogical data that require additional laboratory analyses. The present results were obtained under controlled laboratory conditions; under field or imaging applications, moisture and surface roughness would likely become stronger sources of variation [4,14]. In addition, when models are applied to new pedological settings, local adaptation strategies such as spiking or global–local calibration are often more reliable than direct deployment of an unadapted model [35,36]. The residual-correction approach used here is consistent with this broader view, although it targets composition-related residual bias using ancillary soil variables rather than expanding or locally adapting the spectral calibration set. For

routine laboratory prediction within a comparable domain, the combination of 0.25 mm sieving, SG-1st preprocessing, and XGBoost appears to be a practical default workflow, whereas composition-aware correction is better reserved for existing ancillary datasets, calibration-library development, or oxide- and mineral-rich soils where additional analyses are justified.

5. Conclusions

Within this benchmarking framework, XGBoost achieved the best overall performance, followed by SVR and PLSR. The combination of the 0.25 mm fraction and Savitzky–Golay first-derivative preprocessing provided the most stable results, whereas further grinding to 0.15 mm did not yield consistent gains. SHAP identified dataset-specific Vis–NIR windows, with major contributions at 430–480 and 1330–1450 nm in the Mollisol dataset and at 585–635, 820–900, and 2180–2240 nm in the Ultisol dataset, consistent with the PCA loading structure. These patterns indicate that TN was inferred mainly through covariation with organic matter and mineral background rather than through a distinct TN absorption feature. Held-out performance was lower in the sampled Ultisol, where composition-related errors were more evident under oxide- and quartz-rich backgrounds, and where composition-aware residual correction produced a larger improvement than in the Mollisol. Taken together, the results support the 0.25 mm, SG-1st, XGBoost combination as a practical first-stage laboratory workflow for TN prediction. Composition-aware residual correction may further reduce residual bias in oxide- and mineral-rich datasets, but its use is conditional on the availability or justified acquisition of ancillary composition data.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/soilsystems10050059/s1>, Table S1: Hyperparameter grids and selection rules for PLSR, SVR, and XGBoost; Table S2: Pearson correlation coefficients between total nitrogen (TN) and selected soil components in Mollisols and Ultisols; Table S3: Paired-bootstrap 95% confidence intervals for held-out test performance metrics of the best-performing first-stage models; Figure S1: Monte Carlo outlier detection using the mean and standard deviation of squared prediction error. Figure S2: Loadings of the first two principal components (PC1 and PC2) derived from SG-smoothed soil reflectance spectra. Figure S3: Pearson correlation spectra between total nitrogen (TN) and Vis–NIR reflectance (350–2500 nm) for Mollisols and Ultisols across three particle-size fractions. Correlation curves are shown for (a–c) Mollisols and (d–f) Ultisols at 2.0, 0.25, and 0.15 mm under three preprocessing strategies: SG + SNV, SG + 1st derivative, and SG + 2nd derivative.

Author Contributions: Conceptualization, S.H., C.L., and X.G. (Xinhao Gao); methodology, S.H., Q.Z., and Y.J.; software, S.H.; validation, S.H., Q.Z., Y.J., and X.G. (Xinhao Gao); formal analysis, S.H.; investigation, S.H., Q.Z., Y.J., S.T., Y.C., X.G. (Xilin Guan), T.S., S.L., Z.Z., W.Y., and R.L.; resources, C.L., X.G. (Xinhao Gao), and G.C.; data curation, S.H., Q.Z., and Y.J.; writing—original draft preparation, S.H.; writing—review and editing, G.C., C.L. and X.G. (Xinhao Gao); visualization, S.H.; supervision, C.L., X.G. (Xinhao Gao), and G.C.; project administration, C.L. and X.G. (Xinhao Gao); funding acquisition, C.L. and X.G. (Xinhao Gao). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Mechanism of Soil Spectral Response in Complex Environments and Field Validation (CAAS-CSAL-202402), the National Key Research and Development Program of China (2022YFD1500100), and the Key R&D Program of Shandong Province, China (2024SFGC0405).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data that support the findings of this study are available upon reasonable request to the corresponding author.

Acknowledgments: The authors sincerely thank the Qiyang Red Soil Experimental Station, Hunan, for providing the Ultisol samples used in this study and Hongliang Wu of the Center for Agricultural Resources Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, for providing the Mollisol samples.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and Near Infrared Spectroscopy in Soil Science. *Adv. Agron.* **2010**, *107*, 163–215. [\[CrossRef\]](#)
2. Nocita, M.; Stevens, A.; van Wesemael, B.; Aitkenhead, M.; Bachmann, M.; Barthès, B.; Dor, E.B.; Brown, D.J.; Clairotte, M.; Csorba, A.; et al. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *Adv. Agron.* **2015**, *132*, 139–159. [\[CrossRef\]](#)
3. Erebara, A.; Kopali, A.; Shallari, S.; Di Iorio, E.; Manzo, A.; Longobardi, E.; Napoletano, P.; De Marco, A.; Colombo, C. Pedological features affecting carbon accumulation in submerged-soils sequence assessed with Vis-NIR spectroscopy in a coastal wetland in Albania. *Catena* **2025**, *257*, 109187. [\[CrossRef\]](#)
4. Soriano-Disla, J.M.; Janik, L.J.; Viscarra Rossel, R.A.; Macdonald, L.M.; McLaughlin, M.J. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. *Appl. Spectrosc. Rev.* **2014**, *49*, 139–186. [\[CrossRef\]](#)
5. Rinnan, Å.; van den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222. [\[CrossRef\]](#)
6. Yun, Y.H.; Li, H.D.; Deng, B.C.; Cao, D.S. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends Anal. Chem.* **2019**, *113*, 102–115. [\[CrossRef\]](#)
7. Viscarra Rossel, R.A.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [\[CrossRef\]](#)
8. Clingensmith, C.M.; Grunwald, S. Predicting Soil Properties and Interpreting Vis-NIR Models from across Continental United States. *Sensors* **2022**, *22*, 3187. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [\[CrossRef\]](#)
10. Ben-Dor, E.; Ong, C.; Lau, I.C. Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma* **2015**, *245–246*, 112–124. [\[CrossRef\]](#)
11. Coutinho, M.A.N.; Alari, F.D.O.; Ferreira, M.M.C.; do Amaral, L.R. Influence of soil sample preparation on the quantification of NPK content via spectroscopy. *Geoderma* **2019**, *338*, 401–409. [\[CrossRef\]](#)
12. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. [\[CrossRef\]](#)
13. Krstajic, D.; Buturovic, L.J.; Leahy, D.E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminformatics* **2014**, *6*, 10. [\[CrossRef\]](#)
14. Minasny, B.; McBratney, A.B.; Bellon-Maurel, V.; Roger, J.-M.; Gobrecht, A.; Ferrand, L.; Joalland, S. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma* **2011**, *167–168*, 118–124. [\[CrossRef\]](#)
15. Sherman, D.M.; Waite, T.D. Electronic spectra of Fe³⁺ oxides and oxide hydroxides in the near IR to near UV. *Am. Mineral.* **1985**, *70*, 1262–1269.
16. Scheinost, A.C.; Chavernas, A.; Barrón, V.; Torrent, J. Use and limitations of second-derivative diffuse reflectance spectroscopy in the visible to near-infrared range to identify and quantify Fe oxide minerals in soils. *Clays Clay Miner.* **1998**, *46*, 528–536. [\[CrossRef\]](#)
17. Viscarra Rossel, R.A.; Walvoort, D.J.J.; McBratney, A.B.; Janik, L.J.; Skjemstad, J.O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75. [\[CrossRef\]](#)
18. Soil Survey Staff. *Keys to Soil Taxonomy*, 13th ed.; U.S. Department of Agriculture, Natural Resources Conservation Service: Washington, DC, USA, 2022.
19. Pribyl, D.W. A critical review of the conventional SOC to SOM conversion factor. *Geoderma* **2010**, *156*, 75–83. [\[CrossRef\]](#)
20. Moore, D.M.; Reynolds, R.C., Jr. *X-Ray Diffraction and the Identification and Analysis of Clay Minerals*, 2nd ed.; Oxford University Press: Oxford, UK, 1997.
21. Rietveld, H.M. A profile refinement method for nuclear and magnetic structures. *J. Appl. Crystallogr.* **1969**, *2*, 65–71. [\[CrossRef\]](#)

22. Mehra, O.P.; Jackson, M.L. Iron oxide removal from soils and clays by a dithionite-citrate system buffered with sodium bicarbonate. In *Clays and Clay Minerals: Proceedings of the Seventh National Conference on Clays and Clay Minerals*; Pergamon Press: Oxford, UK, 1960; pp. 317–327.
23. United States Environmental Protection Agency. *SW-846 Test Method 6010D: Inductively Coupled Plasma-Optical Emission Spectrometry (ICP-OES). Revision 5*; U.S. Environmental Protection Agency: Washington, DC, USA, 2018.
24. Savitzky, A.; Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [[CrossRef](#)]
25. Barnes, R.J.; Dhanoa, M.S.; Lister, S.J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43*, 772–777. [[CrossRef](#)]
26. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [[CrossRef](#)]
27. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [[CrossRef](#)]
28. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [[CrossRef](#)]
29. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150202. [[CrossRef](#)]
30. Lundberg, S.M.; Erion, G.G.; Lee, S.I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* **2018**, arXiv:1802.03888. [[CrossRef](#)]
31. Brunet, D.; Barthès, B.G.; Chotte, J.L.; Feller, C. Determination of carbon and nitrogen contents in Alfisols, Oxisols and Ultisols from Africa and Brazil using NIRS analysis: Effects of sample grinding and set heterogeneity. *Geoderma* **2007**, *139*, 106–117. [[CrossRef](#)]
32. Martín-Fernández, J.A.; Barceló-Vidal, C.; Pawłowsky-Glahn, V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* **2003**, *35*, 253–278. [[CrossRef](#)]
33. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [[CrossRef](#)]
34. Recha, J.W.; Olale, K.O.; Sila, A.; Ambaw, G.; Radeny, M.; Solomon, D. Ensemble Modeling on Near-Infrared Spectra as Rapid Tool for Assessment of Soil Health Indicators for Sustainable Food Production Systems. *Soil. Syst.* **2021**, *5*, 69. [[CrossRef](#)]
35. Seidel, M.; Hutengs, C.; Ludwig, B.; Thiele-Bruhn, S.; Vohland, M. Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations. *Geoderma* **2019**, *354*, 113856. [[CrossRef](#)]
36. St. Luce, M.; Ziadi, N.; Viscarra Rossel, R.A. GLOBAL-LOCAL: A new approach for local predictions of soil organic carbon content using large soil spectral libraries. *Geoderma* **2022**, *425*, 116048. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Reproduced with permission of copyright owner. Further reproduction
prohibited without permission.