Tang *et al. Chem. Biol. Technol. Agric.* (2024) 11:33
https://doi.org/10.1186/s40538-024-00552-6

Chemical and Biological Technologies in Agriculture

RESEARCH

Open Access

Improving the performance of a spectral model to estimate total nitrogen content with small soil samples sizes

Weihao Tang[1†], Wenfeng Hu[1,2†], Chuang Li[1], Jinjing Wu[1], Hong Liu[1], Chao Wang[2], Xiaochuan Luo[1] and Rongnian Tang[1*]

Abstract

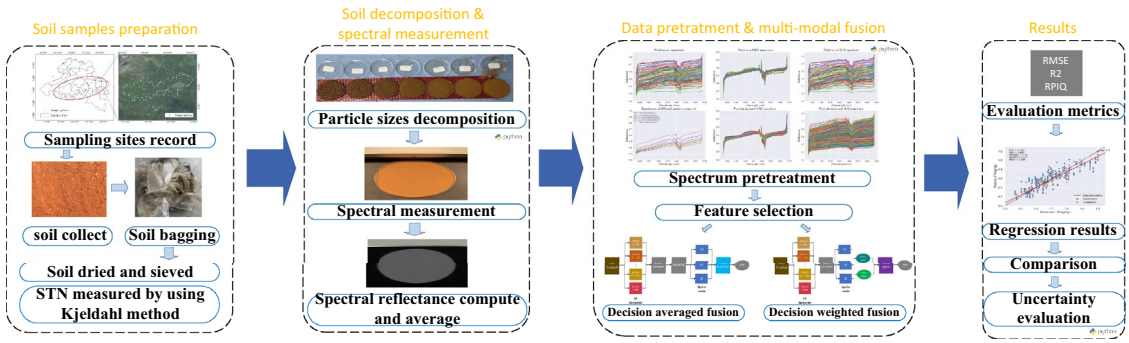
The application of near-infrared spectroscopy (NIRS) for rapid quantitative analysis of soil total nitrogen (STN) is of great significance to recycling nitrogen in the ecosystem and crops growth. However, collecting thousands of soil samples and chemical analysis are impracticable, more importantly a deviation from NIRS advantages of rapid, inexpensive and nondestructive. To more efficiently improve the estimation performance and reduce uncertainty of the model when working with small sample sizes (less than 100), solutions from soil particle size decomposition and model fusion were investigated. Elaborately, 123 Latosols samples were collected and decomposed them according to particle sizes to extend limited data at multiple scales. Based on all soil groups decomposed, a hyperspectral data recapture and model decision fusion method were implemented. The results demonstrated that the proposed method increased the scale of spectral data, extracted more STN-related spectral information, improved estimation accuracy, and reduced uncertainty. The fused model based on data from all decomposed groups yielded the best estimated results (root mean square error (*RMSE*) $= 0.075 g.kg^{-1}$, $R^2 = 0.784$, and a ratio of performance to interquartile distance (*RPIQ*) $= 3.787$) on the validation set. Through a tenfold cross-validation, the weighted fusion model with six groups of particle sizes data showed an improvement of 0.307 in R_c^2v and an improved *RPIQ* of 1.015 compared to models constructed using conventional machine learning (ML) techniques and limited pristine data ($R_c^2v = 0.442$, $RMSE = 0.119$). Therefore, when utilizing NIRS to build rapid and accurate STN predictive models, the proposed method demonstrates great potential in improving the reliability of soil spectral models under small sample sizes.

Keywords Soil total nitrogen, Near-infrared spectroscopy, Soil particle sizes decomposition, Spectral data recapture, Model fusion, Spectral modelling

[†]Weihao Tang and Wenfeng Hu contributed equally to this work.

*Correspondence:
Rongnian Tang
rn.tang@hainanu.edu.cn
Full list of author information is available at the end of the article

Graphical Abstract



Introduction

Quantitative knowledge of soil total nitrogen (STN) is of great significance to society, as it helps to ensure crop growth and reduce leakage of N from agricultural activities into the environment [1, 2]. Latosols, highly weathered and dystrophic soils, dominated by low-activity 1:1 clay minerals and *Fe* and *Al* oxyhydroxides [3]. They are primarily used for the cultivation of rubber trees (Hevea brasiliensis) in Hainan Island. Estimating STN content in Latosols is thus essential, not only for natural rubber production, but also for the health and income of the local population.

In recent years, the fast and nondestructive characteristics of NIRS technology has facilitated its use in STN quantification [4–6]. However, insufficient soil samples owing to the difficulty and expense in densely collecting and chemically analysing thousands of samples, result in large deviations in spectral estimation models and weak generalization capabilities [7–9].

To more efficiently enhance the reliability of spectral models under small sample sizes, machine learning approaches have been proposed, including ensemble regression techniques, spectral dimensionality reduction, and data pretreatment [7, 10, 11]. However, solutions from increasing sample sizes from physical levels are hardly discussed to generate more spectral observations efficiently. Providing comprehensive observations is a practical approach to mitigate the high uncertainty and poor robustness of models under small sample sizes (n<100) [12]. In view of this, we attempt to offer models with the multi-scale data by recapturing spectra of Latosols samples decoupled into different particle sizes. Based on multiple sizes data, assure that model comprehensively captures the pattern between soil spectral characteristics and STN content. Currently, there is a little discussion on dealing with small samples by means of soil particle decomposition. Hence, a strategy from this new perspective deserves further exploration, and soil samples augmenting framework should be developed to deal with the challenge of small sample size in modelling.

Previous studies have demonstrated that soil particle sizes are closely related to spectral information [13]. When soil with a particle size of less than 1 mm was further divided into multiple groups, for each group, the measured spectral data exhibited a high correlation with soil organic matter in different degrees [14, 15]. This conclusion provided sufficient motivation and evidence for the utilization of physical-scale variation on pristine soil to generate more spectral information and augment limited data. Furthermore, decision-level model fusion framework have been introduced to generate more stable estimations with satisfactory accuracy [16, 17].

In terms of this, a potential solution for increasing sample sizes is to decompose limited soil samples into multiple groups and fuse established models based on multi-scale soil data. However, before applying this approach to fill the research gap, the following questions need to be addressed: Can soil particle size decomposition yield more informative data? Can decomposition into more samples and a model fusion strategy further improve estimation performance under small sample sizes? Can this new method provide a rapid and flexible way to estimate STN under small soil sample sizes?

To answer these questions, our study aims to: (1) investigate the effect of soil particle size decomposition on model performance; (2) propose a new framework for rapid soil spectral observations expansion and multi-model fusion to improve model estimation performance

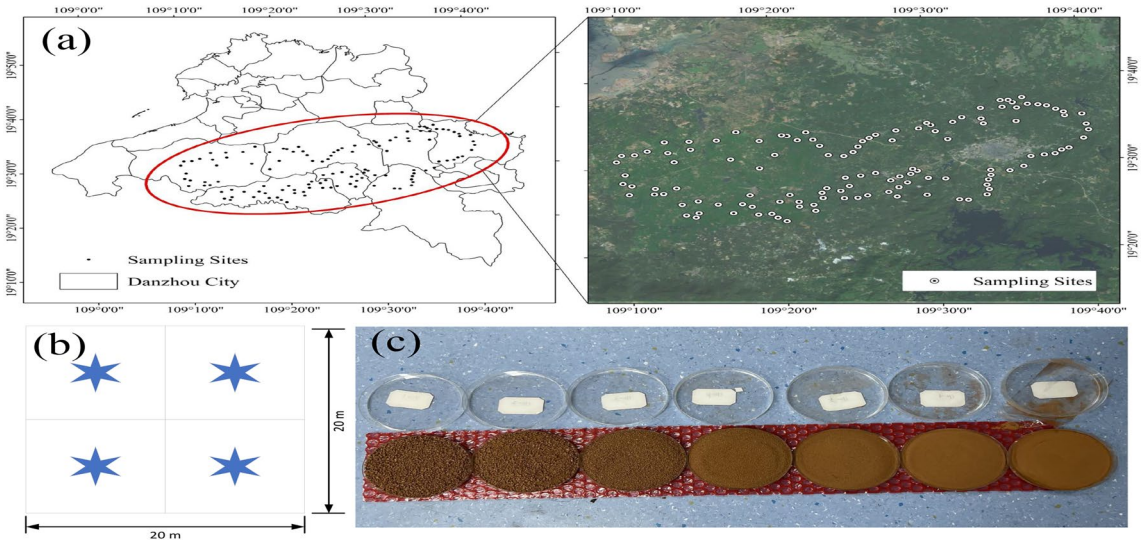


Fig. 1 Sampling sites location (**a**). The left image shows the sampling sites in Arcgis, and the figure to the right is the remote sensing image (**b**). A four-point sampling quadrat used in our study (**c**). Seven groups sieved soil samples. The particle size ranges of soil samples from left to right in the figure are < 2.0mm, 1.0—2.0 mm, 0.5—1.0 mm, 0.25—0.50 mm, 0.15—0.25 mm, 0.09—0.15 mm, and < 0.09 mm

under small sample sizes; and (3) efficiently estimate STN in Latosols using the NIRS technique.

Materials and methods

Soil samples collection

Latosols samples were collected from Danzhou, Hainan Province, in the southern part of China. The details of the location of uniform distributed sampling sites are shown in Fig. 1a, which were recorded with a handheld GPS. For each site, a drill was used to dig 20 cm below the surface layer, and 4 soil sub-samples within 400 m^2 were collected by using a four-point sampling quadrat, which is shown in Fig. 1b. They were mixed and packed into a white plastic bag as one soil sample. Subsequently, a total of 123 soil samples were fast labelled and sent to the laboratory.

The oven-drying method was employed to dry the soil samples at 100°C for 24 h [18]. After drying, they were stood for 8 h and sieved 2 mm uniformly, to minimize the systematic and random effects for subsequent spectral measurements [19]. Notably, a dry soil sample with a particle size of less than 2 mm and room temperature was considered as the pristine soil sample for further study.

Soil particle sizes decomposition

To efficiently provide a greater number of spectral observations for soil samples, we implemented an extension of soil sample sizes at the physical level. Namely, the mixed pristine soil was modified through particle size decomposition, generating additional spectral data from limited pristine soil samples. This hyperspectral recapture for each soil sample aimed to ensure that the model learned from the augmented spectral data and accurately captured the representation of soil total nitrogen (STN).

Specifically, the soil samples with sizes less than 2 mm were further sieved using five sieves with gaps of 1, 0.5, 0.25, 0.15, and 0.09 mm. As a result, the soil particles were decomposed into the following size ranges: 1.0 − 2.0, 0.5 − 1.0, 0.25 − 0.50, 0.15 − 0.25, 0.09 − 0.15, and < 0.09 mm. This decomposition process yielded a total of seven groups, including the pristine soil sample with sizes less than 2 mm. The sieved soil samples in the cylinders are shown in Fig. 1c, where the pristine mixed soil samples (< 2mm) are presented first on the left, and the rest are the six groups of sieved soil.

To demonstrate that particle size decomposition can provide more informative data, we conducted canonical correlation analysis (CCA) to assess the correlation between each group of particle sizes and STN [20]. Fig. 2 shows the correlation coefficients obtained from the six groups of decomposed soil spectra, which are higher than those of the pristine soil. This indicates that the decomposition method can extract various information with high relevance from the limited soil samples. This conclusion is consistent with the work of Wu et al. [14].

Spectral measurement and data prepossessing

Before acquiring spectra reflectance, the instrument was calibrated using a 99% reference white board [21, 22]. Subsequently, the reflectance data for each soil sample were measured using a spectrometer (GaiaField-F-N17E). An interesting observation is that as the soil particle size

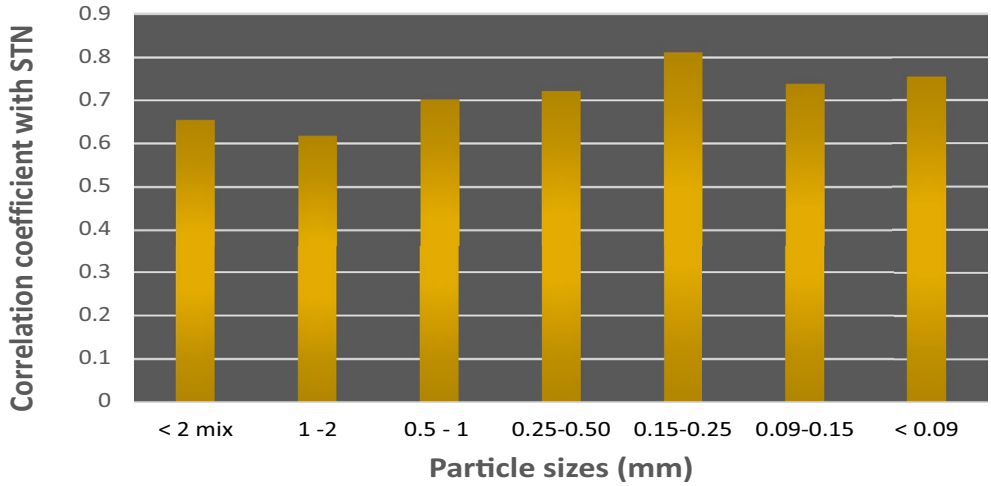


Fig. 2 Barplot of correlation coefficients between spectrum of different particle sizes and total nitrogen content

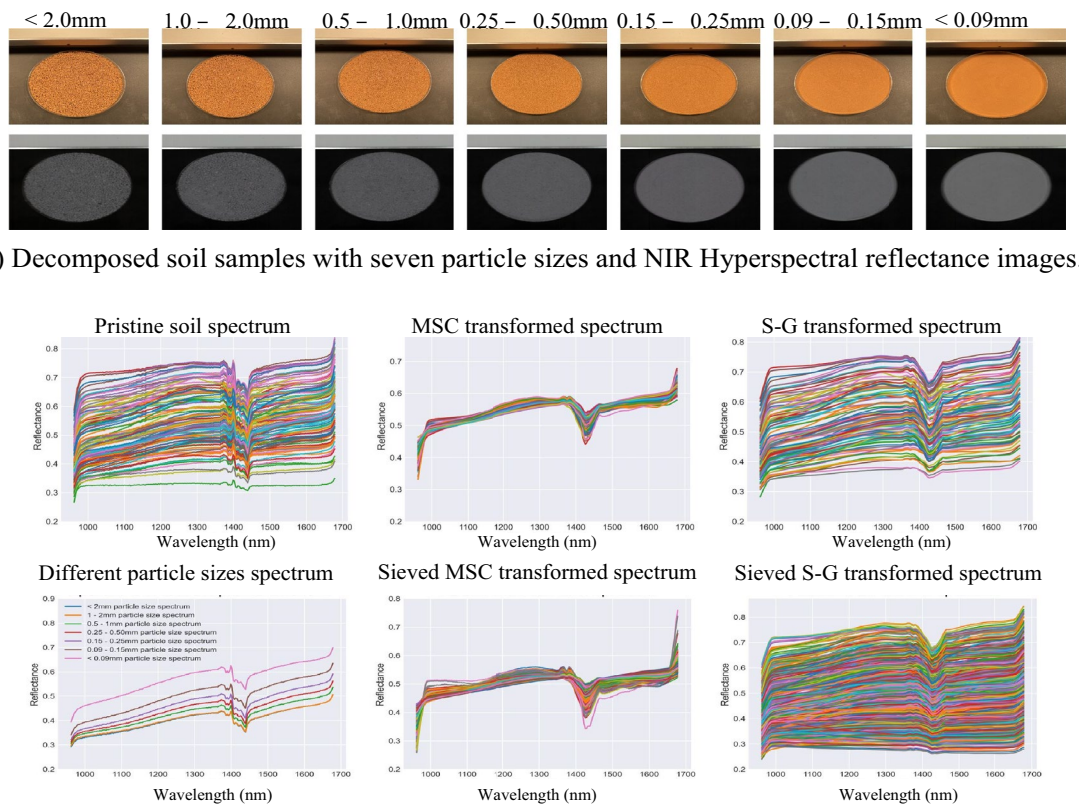


Fig. 3 NIR reflectance images of decomposed soil samples with seven particle sizes and averaged spectrums with multiple transformed methods. MSC refers to multiplicative scatter correction and S-G refers to Savitzky–Golay

Table 1 Soil sample total nitrogen content distribution of different data set

Data set	Size	Range($g.kg^{-1}$)	Mean($g.kg^{-1}$)	CV(%)
Total	111	0.13 ~ 1.45	0.582	41.6
Calibration set	78	0.13 ~ 1.45	0.589	45.1
Validation set	33	0.31 ~ 0.97	0.567	27.7
Calibration set (six particle sizes)	468	0.13 ~ 1.45	0.589	45.1
Validation set (six particle sizes)	198	0.31 ~ 0.97	0.567	27.7

decreases, the reflectance images become brighter in Fig. 3a, which aligns with the findings of Xie et al. [14]. Simultaneously, an increasing trend occurs as the soil particle size scale decreases, and the averaged spectrum has higher value, which can be viewed in Fig. 3b.

The spectrometer used for data collection had a wavelength range of 840-1700 nm, resulting in a total of 254 reflectance bands for each soil sample. However, the first 30 and last seven bands were found to contain significant high-frequency noise. Therefore, to ensure data quality, 217 bands were selected and reserved as soil spectral data within the wavelength range of 942-1678 nm. Subsequently, the acquired hyperspectral images with 217 bands were further averaged, as depicted in Fig. 3.

During the data collection process, random noise and outliers are inevitable, particularly in environments with high air humidity [23–26]. A clear example of this interference can be seen in the raw spectral data observed at wavelengths 1400 to 1450 nm (Fig. 3a). Consequently, raw spectral data cannot be directly used for modelling purposes without any preprocessing.

To address this issue, the three-sigma rule [27] was applied to identify and remove 12 abnormal samples that deviated significantly from the majority of collected samples. This step was crucial to prevent the training models from being biased by these outliers. Furthermore, a Savitzky–Golay (S-G) filter [28] was conducted to filter high-frequency noise in spectral signals at 1400nm. Multiplicative scatter correction (MSC)[29] was implemented to remove the scattering interference during spectrum measurement. The pretreated single group spectrum curve and particle sizes decomposed curves can be shown in Fig. 3b.

Moreover, the spectrum contained redundant and collinear bands that distorted the estimated model parameters, resulting in model's weak generalization ability. Thus, two classical spectral bands algorithms were conducted, the successive projection algorithm (SPA) [30] and uninformative variables elimination (UVE) [31].

Chemical analysis

The most widely employed method in chemical analysis is the Kjeldahl determination method [32, 33], which is generally used as a reference method for nitrogen concentration estimation. Thus, the semi-micro Kjeldahl distillation method was used in this study to determine the total nitrogen content in the soil, and the absolute error between the measured and standard values was less than 0.01 g/kg for the collected soil samples. This is the prerequisite for establishing an accurate STN spectral estimation model. Notably, samples of different particle size groups decomposed from the same pristine soil have a consistent STN. Physical variation does not change the STN of Latosols in different particle sizes.

Descriptive statistics

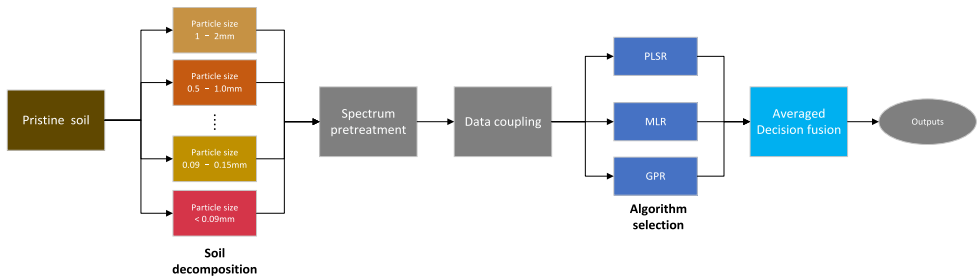
To ensure an objective evaluation of the model's performance, we employed sample set partitioning based on joint x–y distance to split the data into calibration and validation sets [34, 35]. The calibration set accounted for 70% of the samples, which corresponded to 78 samples for calibration and 33 samples for validation.

We performed soil particle size decomposition separately for the calibration and validation sets, generating six different particle size groups for each dataset. The calibration set was used to train the model, while the validation set was utilized to assess the model's performance.

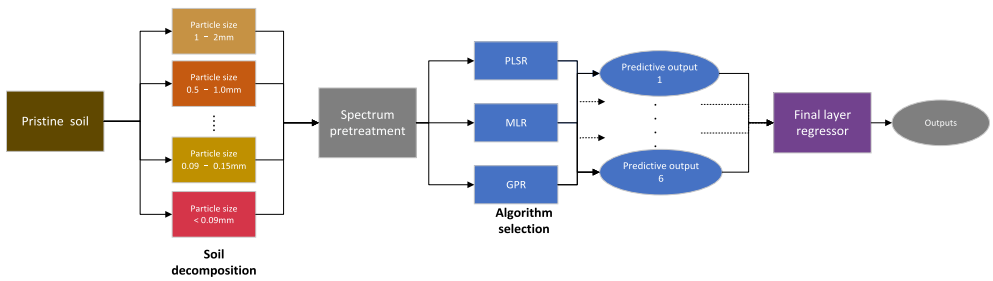
Additionally, a descriptive statistic chart of soil total nitrogen (STN), obtained through chemical analysis, is provided in Table 1.

STN content estimation modelling

Based on expanded spectral dataset, we established multiple regression sub-models separately, and performed decision fusion on multiple output results to establish a more reliable model for small sample sizes. We conducted two fusion strategies: average fusion section , and weighted fusion section . Additionally, partial least squares regression (PLSR), Gaussian process regression (GPR) and multivariate linear regression (MLR) were employed for sub-model establishment in section . Further, common model ensemble methods were introduced as a contrast to comprehensively reflect the effect of the



(a) Chart of averaged fusion framework used for STN value estimation modelling



(b) Chart of the weighted fusion framework used for STN value estimation modelling

Fig. 4 Charts of averaged fusion (**a**) and weighted fusion (**b**) framework used for STN value estimation modelling

proposed method, including Bagging (Random Forest), Stacking, and Cubist methods.

Decision-averaged fusion

To further improve the estimation accuracy and reduce uncertainty under small samples sizes, an average decision fusion on multiple sub-models is indispensable, instead of generating results, respectively [16, 36]. Therefore, we implemented the fusion framework of Bagging, which is the representative work of ensemble learning [37]. Specifically, Bagging yields numerous outputs with the subset selected from an original training set through bootstrapping [37]. The average of these outputs was used as the final result. The detailed construction is shown in Fig. 4a, and the algorithm is as follows.

Assume a fusion with T estimators, where T is equal to 6 in this study. For $i = 1, ..., T$, repeat following steps:

- *Step 1.* Bundle spectral data with the whole T groups of soil spectral data to construct a new set, D.
- *Step 2.* m_b samples are randomly picked from D.
- *Step 3.* The ith estimator is built with the m_b samples picked from D, and train the estimator by the following function: $\hat{\theta} = \underset{\theta}{argmin}\, \mathbb{E}_{(x,y)\sim m_b} \log(y - f_\theta^i(x))$.
- Since acquiring the T estimators, average the outputs of the T estimators as the final result.

Decision-weighted fusion

Considering that there exist differences in effects of STN sub-models learning, stacking method was employed to train a weighted fusion model that fuses multiple sub-models via different weights obtained from training stage [7, 17]. To obtain suitable weights, the outputs of all sub-models and true STN value are deemed as new features and labels. A last-layer regressor was further trained to determine fusion weights [38]. The framework used in our study is shown in Fig. 4b, and the detailed procedure as follows.

A fusion with (T=6) models was assumed. For $i = 1, ..., T$, start following steps:

- *Step 1.* Enter the spectral data of the ith particle size, x_i.
- *Step 2.* Build a proprietary predictive model for the ith data to generate a mapping between data and STN value.

$$\hat{y}_i = f_\theta^i(x_i) \tag{1}$$

- *Step 3.* Collect the output of entire T models.
- *Step 4.* Build a final-layer regression model, which is used to learn the optimal weights of each model's output.
- *Step 5.* Compute ω_i by solving the following equation:

$$\hat{\omega} = \underset{\omega}{argmin}\left(y - \sum_{i=1}^{K}\omega_i\hat{y}_i\right)^2. \quad (2)$$

- Finally, output the stacking result

$$\left[\hat{y}_1\ \hat{y}_2\ \cdots\ \hat{y}_T\right]\cdot\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{bmatrix} = y_p, \quad (3)$$

where ω_i is calculated from the equation in Step 5, y_p is the prediction result of the fusion model, where, $\hat{y}_i$ is the prediction outcome of the ith model, ω_i is the stacking weight for the ith model's output, and y represents the STN label.

Sub-regressor modelling method

In this study, the PLSR, MLR, and GPR algorithms were used to establish the sub-model with spectral reflectance data. MLR and PLSR have been used as reference linear modelling methods in the field of analytical chemistry [39, 40]. Further, UVE and SPA were employed to select the key bands from enhanced data, and thus, we used the UVE-PLSR and SPA-MLR methods to build the sub-models. Meanwhile, popular ensemble learning methods, including RF, Stacking, and Cubist [41], were implemented to be comparisons to our proposed method as well.

However, both MLR and PLSR are types of linear regression algorithms, and the fitting ability is limited. Considering this, we further implemented a non-linear and non-parametric learning algorithm, GPR, to build predictive a sub-model that is equivalent to kernel ridge regression. In brief, GPR assumes that a Gaussian process prior governs the set of possible latent functions (which are unobserved). Then, the likelihood of the latent function and observations shape this prior to produce posterior probabilistic estimates [42].

Evaluation metrics

The root mean square error (*RMSE*), determination coefficient (R^2) and ratio of performance to inter-quartile distance (*RPIQ*) [43] were used to evaluate the performance of the prediction model, following eqs. (4), (5), (6), (7). Generally, an ideal model exhibits higher R^2, *RPIQ* values and lower *RMSE* value:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \quad (4)$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(y_i - \overline{y_i})(\hat{y}_i - \overline{\hat{y}_i})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}_i})^2}}\right)^2, \quad (5)$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}}, \quad (6)$$

$$RPIQ = \frac{IQ}{RMSE} = \frac{Q_3 - Q_1}{RMSE}, \quad (7)$$

where *IQ* represent the inter-quartile distance.

Notably, the uncertainty is another key issue in soil mapping, especially in our scenario of applying NIRS model to predict STN under small sample sizes [36, 44]. A tenfold cross-validation method was supposed to validate the estimation uncertainty, where R^2_{cv} and $RPIQ_{cv}$ were used. To illustrate our method comprehensively, the experimental workflow presented in Fig. 5 demonstrated the experiment methodology of the study.

Results

Comparison of model fusion strategies

Table 2 summarizes the evaluation metrics of all estimated models constructed using the enhanced data. The performances of the models using the multi-model fusion were better in general. For instance, the R^2_v and $RMSE_v$ of

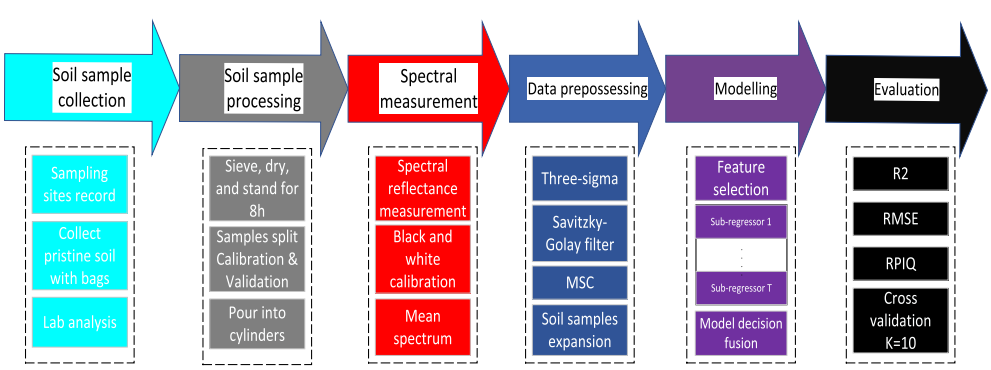


Fig. 5 The experimental workflow of our study

Table 2 Estimation results of using different fusion strategies and sub-regressors

Fusion strategy	Regressor	Calibration		Validation		Tenfold validation	
		R_c^2	$RMSE_c\,(g.kg^{-1})$	R_v^2	$RMSE_v\,(g.kg^{-1})$	R_{cv}^2	RPIQ
Single[1]	GPR[2]	0.695	0.140	0.627	0.107	0.598	2.895
Bagging	GPR	0.702	0.138	0.711	0.088	0.651	2.942
Stacking	GPR	0.711	0.131	0.723	0.085	0.653	2.951
Single	UVE-PLSR[3]	0.645	0.156	0.597	0.113	0.550	2.937
Bagging	UVE-PLSR	0.715	0.140	0.737	0.083	0.693	3.324
Stacking	UVE-PLSR	0.734	0.123	0.714	0.087	0.677	3.340
Single	SPA-MLR[4]	0.429	0.198	0.553	0.117	0.484	2.501
Bagging	SPA-MLR	0.683	0.147	0.709	0.087	0.680	3.314
Stacking	SPA-MLR	0.773	0.125	0.784	0.075	0.720	3.511

[1] Single implies without using any model fusion strategy, a single model is directly implemented on the expanded dataset

[2] GPR represents using Gaussian progress regressor to build sub-models

[3] UVE-PLSR is to use the combination of uninformative variable elimination techniques and partial least regressor to build sub-models

[4] SPA-MLR is to implement the successive projections algorithm to remove co-linear variables and establish multiple linear regression sub-models

simply using SPA-MLR and GPR algorithms were 0.553 and 0.117 $g.kg^{-1}$, 0.627 and 0.107 $g.kg^{-1}$, respectively, on the sieved soil data, as summarized in Table 2. Particularly, the results of the fusion modelling approach of SPA-Stacking was the best in our study, and the R_v^2 and $RMSE_v$ reached 0.784 and 0.075 $g.kg^{-1}$, respectively, as shown in regression scatter diagram (Fig. 6).

Additionally, the stacking construction can be more effective to elevate the STN model performance than Bagging, as shown in Fig. 4. When the SPA-Bagging-MLR framework was implemented, the R_v^2 and $RMSE_v$ of the model were 0.709 and 0.087$g.kg^{-1}$, respectively; whereas, the R_v^2 of the SPA-Stacking-MLR model improved by 0.075 than SPA-Bagging-MLR model. Simultaneously, when UVE-PLSR a were offered with stacking fusion, the R_v^2 was 0.139 higher than that without model fusion, and the $RMSE_v$ decreased by 17.5%, as summarized in Table 2.

Model performance comparison of using soil particle size decomposition method

The record of estimation results of modelling by using groups of soil data decomposed and using original soil data is summarized in Table 3. This table indicates that using the size of 666 data can generate a higher accuracy model than the pristine soil dataset. The R_v^2 and $RMSE_v$ of

Table 3 Result comparison of using soil decomposed data and particle size of 0.15 mm—0.25 mm to build the model

Data size	Data category	Regressor	Calibration		Validation		Tenfold validation	
			R_c^2	$RMSE_c\,((g.kg^{-1}))$	R_v^2	$RMSE_v\,((g.kg^{-1}))$	R_{cv}^2	RPIQ
111	Pristine mixed soil data[1]	UVE-PLSR	0.514	0.181	0.457	0.118	0.433	2.105
		GPR	0.737	0.132	0.543	0.108	0.522	2.603
		SPA-MLR	0.374	0.198	0.418	0.132	0.385	2.032
	0.15-0.25mm data[2]	UVE-PLSR	0.645	0.156	0.597	0.113	0.550	2.937
		GPR	0.695	0.140	0.627	0.107	0.598	2.895
		SPA-MLR	0.429	0.198	0.553	0.117	0.484	2.501
	0.15-0.25mm data	Bagging	0.568	0.172	0.608	0.101	0.579	2.514
		Stacking	0.625	0.161	0.548	0.103	0.510	2.227
		Cubist	0.719	0.139	0.631	0.098	0.589	2.631
666	Six groups of particle sizes data[3]	PLSR	0.643	0.167	0.637	0.097	0.600	2.650
		Stacking	0.773	0.125	0.784	0.075	0.720	3.511
		Bagging	0.715	0.140	0.737	0.083	0.693	3.324

[1] Pristine soil data refer to using a mixture soil sample data with particle size < 2 mm for modelling

[2] 0.15–0.25 mm data are for modelling based on the soil data of particle size ranging from 0.15 mm to 0.25 mm, which superior to the rest five particle sizes group

[3] Six groups of particle sizes data means using the whole particle sizes data to build the model

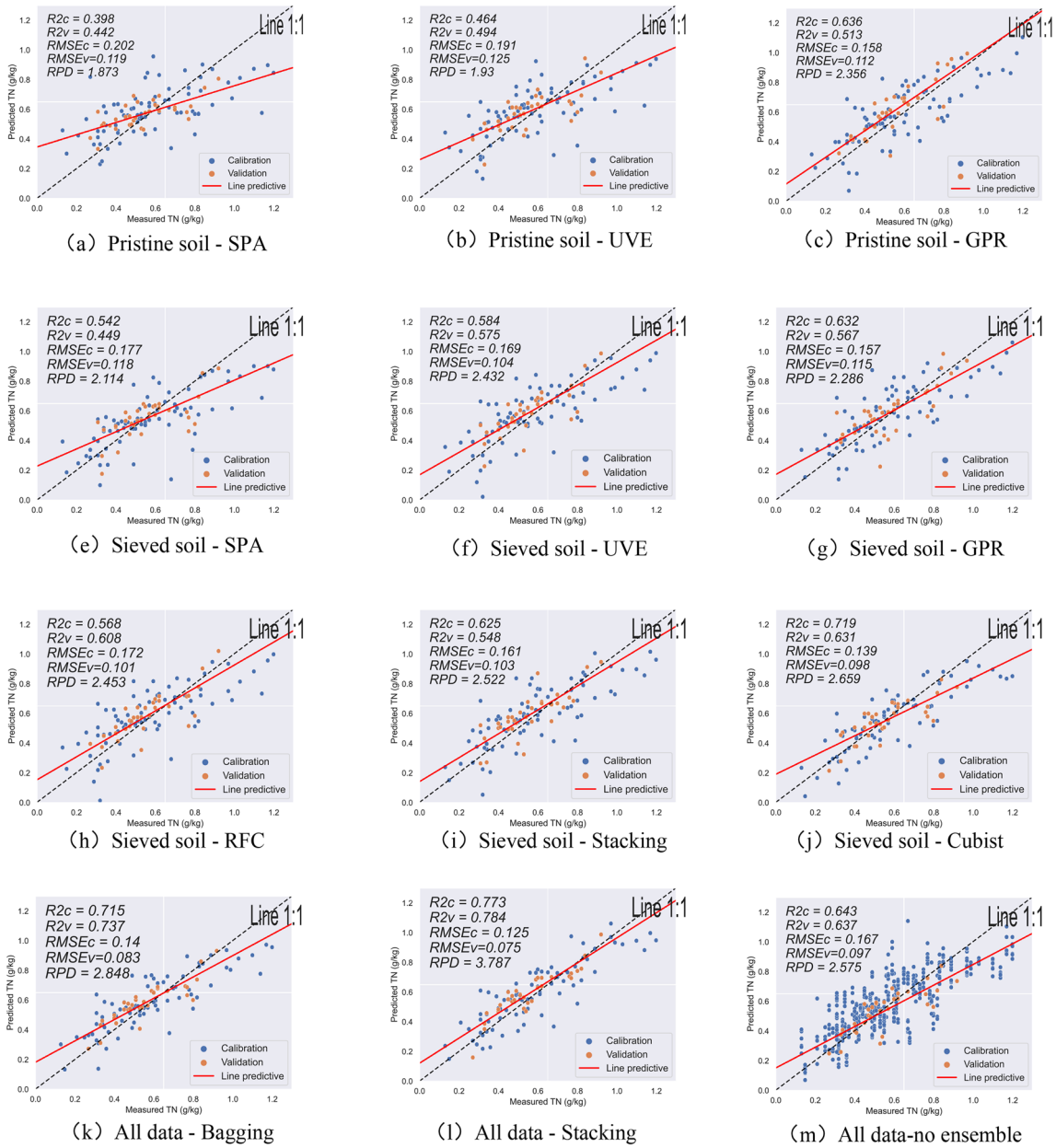


Fig. 6 Regression scatter diagram of using different model fusions. The first row of images are the scatter plots of the prediction results using pristine soil data, the second row of images are the resulting scatter plots of using sieved data with the particle size of 0.15–0.25mm, the third row are the scatter plots of regression results using model fusion method and sieved data, and the last row are the scatter plots of regression results using six group data and model fusion method

the estimation model built with pristine mixed soils data and UVE-PLSR method were 0.457 and 0.118 $g.kg^{-1}$, respectively. The model's performance was not adequate. Moreover, the estimation model was established using the filtered soil spectrum with particle size range of 0.15–0.25 mm, and the results were superior to the rest of the particle size groups. The entire results of using seven groups data modelling are shown in Table 4. When total six groups data (n=666) were implemented, the identical UVE-PLSR and GPR methods generated the better performing model, where the R_v^2 and $RMSE_v$ reached 0.637 and 0.098 $g.kg^{-1}$, 0.620 and 0.104 $g.kg^{-1}$, respectively.

In particular, the R_v^2 score of model based on ensemble method (RF, Stacking and Cubist) without soil spectral recapture were 0.608, 0.548 and 0.631, respectively, which were all lower than UVE-Stacking-PLSR with six

Table 4 Regression results of using different particle sizes soil data

Data (n=111)	Best method[1]	Calibration		Validation	
		R_c^2	$RMSE_c\,(g.kg^{-1})$	R_v^2	$RMSE_v\,(g.kg^{-1})$
Pristine mixed soil data	GPR	0.737	0.132	0.543	0.118
0.1 - 0.2mm data	UVE-PLSR	0.613	0.148	0.505	0.124
0.5 - 0.1mm data	GPR	0.739	0.132	0.570	0.107
0.25 - 0.50mm data	GPR	0.855	0.093	0.575	0.110
0.15 - 0.25mm data	Cubist	0.719	0.139	0.631	0.098
0.09 - 0.15mm data	Cubist	0.697	0.139	0.583	0.102
< 0.09mm data	GPR	0.542	0.157	0.490	0.128

[1] Best method means implemented to get the optimum results among all methods

groups spectral data. To visualize the effectiveness of soil particle sizes decomposition and spectral recapture to assist spectral model learning, the regression scatter diagrams are shown in Fig.6.

Comparison of sub-model modelling methods

In this study, we implemented the UVE-PLSR, GPR, and SPA-MLR methods to establish multiple sub-models. The R_v^2 and $RMSE_v$ results of using different regressors modelling are shown in Fig. 7. As evident, GPR was the best among the three methods in general, although the best performance of the fusion model UVE-Stacking-PLSR was constructed by the regressor UVE-PLSR. When only analysing the effect of sub-regressor selection on model accuracy, the mean R_v^2 metrics of the GPR modelling method was 0.560, while the SPA-MLR and UVE-PLSR were 0.463 and 0.527. To better compare and analyse the estimation effect of the models used in this study, the performance of all models on the validation set is shown in the form of boxplots (Fig. 7).

Discussion

Validity interpretation analysis of soil particle sizes decomposition methods

As summarized in Table 3, the estimation results obtained from modelling with seven groups of spectra are more desirable, even when employing the same model fusion strategies. We attribute this improvement to three main reasons.

First, for data equality, proposed decomposition method can eliminate the interference of physical properties caused by excessive particle size [45]. When combined with pretreatment methods, the information related to STN in the curve was highlighted, and the correlation between STN content and spectrum improved [10, 46]. The detailed results of using seven groups of data modelling are shown in Table 4.

Second, by providing multi-scale data, the approach mitigates the drawbacks associated with high-dimensional spectral data, such as spectral overlap and high levels of noise [47]. This enables the model to confidently capture STN-related information and reject noise interference.

Third, the idea of this method is aligned with the classic bootstrap integration idea [48]. This idea involves collecting spectral estimation and analysis results of the same target (STN) from different observations and integrating these estimates to enhance the accuracy and robustness of the model, particularly in small sample sizes.

Validity interpretation analysis of multi-model fusion strategy

As Table 3 suggests, the results of using model fusion outperformed other methods both on single group and six groups particle sizes data. There are three main reasons why the fusion method implemented in this study can improve the estimation model performance. First, compared with the single model, the integrated model learned from multi-scale information, which generated more reliable results based on more complete estimation. Second, the output of a single model obtained via small samples size had a high uncertainty and massive variation, thus, it is necessary to integrate outputs of multiple models at decision-level to reduce uncertainty [16]. Third, a multi-modal fusion model has the larger and deeper parameter scales than a single model to fit data. The introduction of decision fusion model can further reduce the bias in the regression results and improve the training approximation when the amount of data is suitable [7].

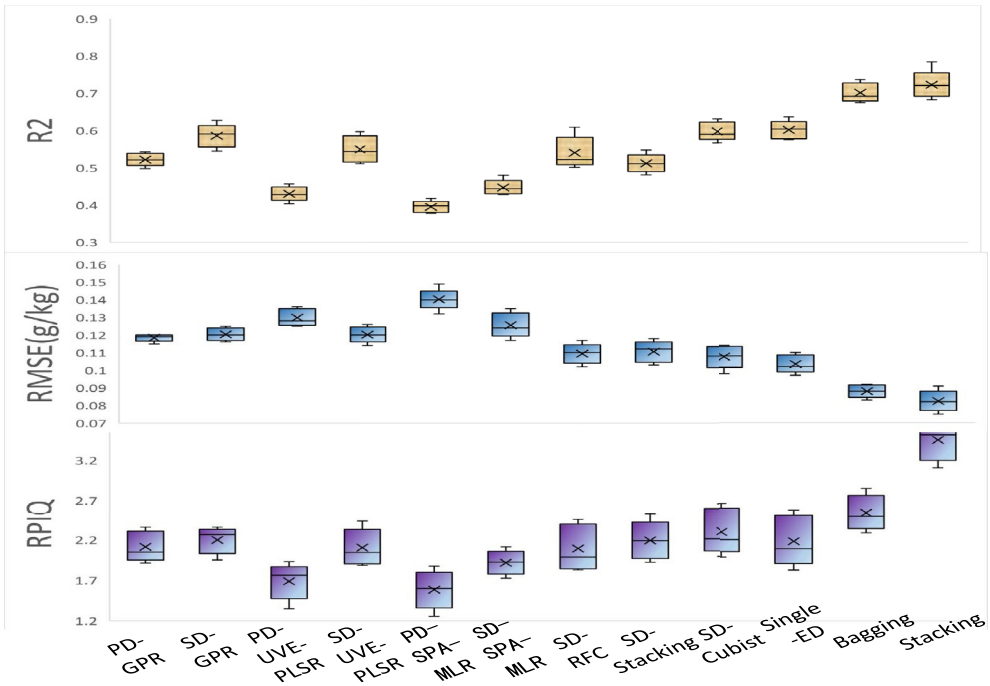


Fig. 7 Boxplots of estimation results of different model establishment methods. PD is an abbreviation for using pristine data to build models. SD is to use sieved data. ED refers to expanded data by soil particle decomposition and RF means random forest

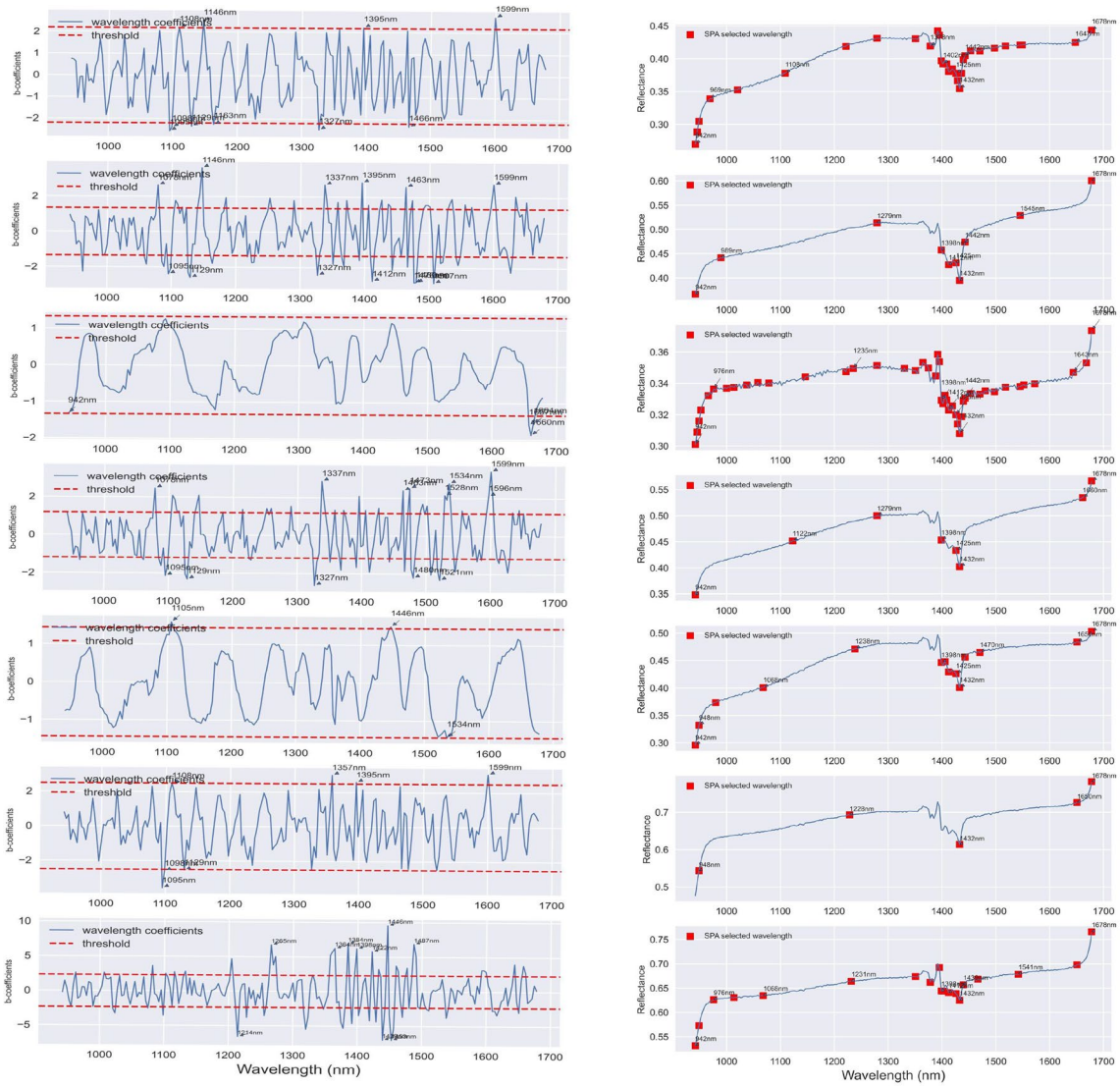
Important wavebands for predicting STN

To enhance the interpretability of the model, it is necessary to visualize the key band selection progress of model. In this study, UVE was employed to select the key bands, which is a method of variable filter based on stability analysis of the regression coefficient; those less than the cut-off threshold were regarded as uninformative and thus eliminated [31, 49]. The specifics of the key band selection for different particle sizes data are documented, and the distribution of the b-coefficients for important bands retained from the PLSR cross-validation for STN estimation is depicted in Fig. 8a. The selection results varied across different particle sizes; however, the overall distributions are similar, with approximate ranges of 992 -1078, 1352-1521 and 1589-1607nm. Basically, these selected bands demonstrate strong associations with the first N–H overtone (1500nm), the second N–H overtone (1000nm), and the first O–H overtone (1400 nm) [6]. Despite the drying of the soil in our experiments, significant O–H is still evident in the curves. The work of Xiao et al. [50] demonstrates that this may be influenced by the response of fundamental $Fe(OH)_3$ and $Al(OH)_3$ from latosols, as well as being affected by the high humidity in the environment.

The wavelength subset obtained after eliminating the spectral collinearity using SPA is shown in the right picture in Fig. 8b, where the marked red points are the selected wavelengths. Additionally, SPA algorithm shares a similar distribution of selected key wavelengths, although the particle sizes are distinct. In general, the UVE method obtains a larger range of wavelengths, larger number, and better results for modelling.

Limitations and future prospects

While our work appears practical, several potential limitations exist. Firstly, the climate of Hainan is an inevitable factor that may affect our experiments. The average annual relative humidity at our experimental site ranges from 75% to 86%, introducing noise at a wavelength of 1400 nm (Fig. 3b), which corresponds to the first O–H from the water in the air. Although S-G smoothing has been applied to mitigate interference in the raw spectral data, there might be an uncertain gap between the measured data and the true soil properties. Notably, it is very common to observe obvious noise at 1400nm, especially during the capture of hyperspectral data by drone-based spectrometers [51]. Thus, an intractable urgent should be viewed as. For future work, we intend to use a radiative transfer model [52] to theoretically calibrate our data. We hope to be able to make a sensible compromise between



(a) UVE (b) SPA

Fig. 8 Result diagrams of key bands for all groups of soil data. Bottom figures are the wavelength selection results of all particle decomposed soil data by using uninformative variables elimination and successive projection algorithm. The rest of the pictures correspond to the selection results of different particle size data. The left figure shows the b-coefficients associated with the partial least regression cross-validation models (k=10) for predicting the STN. The horizontal lines represent thresholds for the b-coefficients based on the standard deviations of the STN. The diagram on the right is the bands selection result of SPA, where red scatters represent the selected key bands

theoretical data and low-quality data captured in the field.

Secondly, our research area was limited, and the soil samples obtained were of a single type. The effectiveness of using our proposed method for rapid and accurate estimation of STN in other soil categories needs to be verified. In terms of this, we are actively seeking to cooperate with researchers in other regions to verify our results with data from multiple soil varieties. In this manner, the influence of geographical and environmental noise factors can be effectively eliminated. In addition, we will use deep learning methods in future experiments to build better performance estimation models.

Conclusion

Soil particle size decomposition was proposed to extract multi-scale spectral information, thereby improving the flexibility and reliability of applying NIRS technology to

quickly assess STN. The SPA-Stacking-MLR model proposed in this study was conducted to assess the total nitrogen value in Latosols soil. Compared with the establishment of a predictive model with limited pristine data, the $RMSEv$ decreased to 0.075 and R_v^2 increased to 0.784. As a result, the proposed approach demonstrated significant potential in improving the accuracy of STN estimation models and reducing uncertainty, particularly under conditions of limited sample sizes. Moreover, our study provided a new perspective for enhancing the performance of STN estimation models under small sample sizes without an intensive and expensive process of large-scale data construction. The results could be applicable to soils derived from other parent materials.

Acknowledgements
The work was supported by the High-level Talent Project of Natural Science Foundation of Hainan Province (No. 321RC468), Key R &D project of Hainan Province(ZDYF2022GXJS008), National Natural Science Foundation of China (No. 32060413) and Innovation Research Team Project of Natural Science Foundation of Hainan Province (No. 320CXTD431).

Author contributions
WT: conceptualization, methodology, writing original draft, writing python code, drafted the work. WH: data curation, investigation, resource, funding acquisition, formal analysis. CL: interpretation of data. JW: formal analysis and data collection. HL: approved the version to be published. CW: supervision. XL: approved the version to be published. RT: project administration, conceptualization, funding acquisition, revised it critically.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Mechanical and Electrical Engineering, Hainan University, Haikou 570228, China. [2]School of Electrical Engineering and Automation, Tianjin University, Tianjin 300072, China.

Received: 8 November 2023 Accepted: 20 February 2024
Published online: 07 March 2024

References

1. Powlson DS. Understanding the soil nitrogen cycle. Soil Use Manag. 1993;9(3):86–93. https://doi.org/10.1111/j.1475-2743.1993.tb00935.x.
2. Subbarao GV, Sahrawat KL, Nakahara K, Ishikawa T, Kishii M, Rao IM, Hash CT, George TS, Srinivasa Rao P, Nardi P, Bonnett D, Berry W, Suenaga K, Lata JC. Chapter six - biological nitrification inhibition-a novel strategy to regulate nitrificatio agricultural systems. In: Sparks DL (ed.) Advances in Agronomy. Advances in Agronomy, vol. 114, Academic Press; 2012; pp. 249–302. https://doi.org/10.1016/B978-0-12-394275-3.00001-8.
3. Schaefer CEGR, Fabris JD, Ker JC. Minerals in the clay fraction of Brazilian latosols (Oxisols): a review. Clay Miner. 2008;43(1):137–54. https://doi.org/10.1180/claymin.2008.043.1.11.
4. Reeves JB, Smith DB. The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America. Appl Geochem. 2009;24(8):1472–81. https://doi.org/10.1016/j.apgeochem.2009.04.017.
5. Yu X, Liu Q, Wang Y, Liu X, Liu X. Evaluation of MLSR and PLSR for estimating soil element contents using visible/near-infrared spectroscopy in apple orchards on the jiaodong peninsula. CATENA. 2016;137:340–9. https://doi.org/10.1016/j.catena.2015.09.024.
6. Wang Q, Zhang H, Li F, Gu C, Qiao Y, Huang S. Assessment of calibration methods for nitrogen estimation in wet and dry soil samples with different wavelength ranges using near-infrared spectroscopy. Comput Electron Agric. 2021;186: 106181. https://doi.org/10.1016/j.compag.2021.106181.
7. Peterson K, Sagan V, Sidike P, Hasenmueller E, Sloan J, Knouft J. Machine learning-based ensemble prediction of water-quality variables using feature-level and decision-level fusion with proximal remote sensing. Photogramm Eng Remote Sensing. 2019;85:269–80. https://doi.org/10.14358/PERS.85.4.269.
8. Shahshahani BM, Landgrebe DA. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Trans Geosci Remote Sens. 1994;32(5):1087–95. https://doi.org/10.1109/36.312897.
9. Zhang X, Lin T, Xu J, Luo X, Ying Y. Deepspectra: an end-to-end deep learning approach for quantitative spectral analysis. Anal Chim Acta. 2019;1058:48–57. https://doi.org/10.1016/j.aca.2019.01.002.
10. Hong Y, Liu Y, Chen Y, Liu Y, Yu L, Liu Y, Cheng H. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. Geoderma. 2019;337:758–69. https://doi.org/10.1016/j.geoderma.2018.10.025.
11. Ng W, Minasny B, Mendes WDS, Demattê JAM. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. SOIL. 2020;6(2):565–78. https://doi.org/10.5194/soil-6-565-2020.
12. Yang J, Xu J, Zhang X, Wu C, Lin T, Ying Y. Deep learning for vibrational spectral analysis: recent progress and a practical guide. Anal Chim Acta. 2019;1081:6–17. https://doi.org/10.1016/j.aca.2019.06.012.
13. Okin GS, Painter TH. Effect of grain size on remotely sensed spectral reflectance of sandy desert surfaces. Remote Sens Environ. 2004;89(3):272–80. https://doi.org/10.1016/j.rse.2003.10.008.
14. Xie S, Li Y, Wang X, Liu Z, Ma K, Ding L. Research on estimation models of the spectral characteristics of soil organic matter based on the soil particle size. Spectrochim Acta Part A Mol Biomol Spectrosc. 2021;260: 119963. https://doi.org/10.1016/j.saa.2021.119963.
15. Wu C, Zheng Y, Yang H, Yang Y, Wu Z. Effects of different particle sizes on the spectral prediction of soil organic matter. CATENA. 2021;196: 104933. https://doi.org/10.1016/j.catena.2020.104933.
16. Wang J, Ding J, Yu D, Teng D, He B, Chen X, Ge X, Zhang Z, Wang Y, Yang X, Shi T, Su F. Machine learning-based detection of soil salinity in an arid desert region, northwest china: a comparison between landsat-8 oli and sentinel-2 msi. Sci Total Environ. 2020;707: 136092. https://doi.org/10.1016/j.scitotenv.2019.136092.
17. Wang J, Shi T, Yu D, Teng D, Ge X, Zhang Z, Yang X, Wang H, Wu G. Ensemble machine-learning-based framework for estimating total nitrogen concentration in water using drone-borne hyperspectral imagery of emergent plants: A case study in an arid oasis, nw china. Environ Pollut. 2020;266: 115412. https://doi.org/10.1016/j.envpol.2020.115412.
18. Birch H. The effect of soil drying on humus decomposition and nitrogen availability. Plant Soil. 1958;10:9–31. https://doi.org/10.1007/BF01343734.
19. Ben Dor E, Ong C, Lau IC. Reflectance measurements of soils in the laboratory: standards and protocols. Geoderma. 2015;245–246:112–24. https://doi.org/10.1016/j.geoderma.2015.01.002.
20. Braak CJF. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology. 1986;67(5):1167–79. https://doi.org/10.2307/1938672.
21. Wang H-F, Huo Z-G, Zhou G-S, Liao Q-H, Feng H-K, Wu L. Estimating leaf spad values of freeze-damaged winter wheat using continuous wavelet analysis. Plant Physiol Biochem. 2016;98:39–45. https://doi.org/10.1016/j.plaphy.2015.10.032.
22. Li S, Luo H, Hu M, Zhang M, Feng J, Liu Y, Dong Q, Liu B. Optical non-destructive techniques for small berry fruits: a review. Artif Intell Agric. 2019;2:85–98. https://doi.org/10.1016/j.aiia.2019.07.002.

23. Basile T, Marsico AD, Perniola R. Nir analysis of intact grape berries: chemical and physical properties prediction using multivariate analysis. Foods. 2021. https://doi.org/10.3390/foods10010113.
24. Altieri G, Genovese F, Tauriello A, Di Renzo GC. Models to improve the non-destructive analysis of persimmon fruit properties by VIS/NIR spectrometry. J Sci Food Agric. 2017;97(15):5302–10. https://doi.org/10.1002/jsfa.8416.
25. Munawar AA, Yunus Y, Devianti, Satriyo P. Calibration models database of near infrared spectroscopy to predict agricultural soil fertility properties. Data Brief. 2020;30: 105469. https://doi.org/10.1016/j.dib.2020.105469.
26. Ren G, Sun Y, Li M, Ning J, Zhang Z. Cognitive spectroscopy for evaluating Chinese black tea grades (Camellia sinensis): near-infrared spectroscopy and evolutionary algorithms. J Sci Food Agric. 2020;100(10):3950–9. https://doi.org/10.1002/jsfa.10439.
27. Khan ZA, Adil M, Javaid N, Saqib MN, Shafiq M, Choi J-G. Electricity theft detection using supervised learning techniques on smart meter data. Sustainability. 2020. https://doi.org/10.3390/su12198023.
28. Schafer RW. What is a savitzky-golay filter? IEEE Signal Process Mag [Lecture Notes]. 2011;28(4):111–7. https://doi.org/10.1109/MSP.2011.941097.
29. Isaksson T, Næs T. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. Appl Spectrosc. 1988;42(7):1273–84.
30. Araújo MCU, Saldanha TCB, Galvão RKH, Yoneyama T, Chame HC, Visani V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemom Intell Lab Syst. 2001;57(2):65–73. https://doi.org/10.1016/S0169-7439(01)00119-8.
31. Cai W, Li Y, Shao X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemom Intell Lab Syst. 2008;90(2):188–94. https://doi.org/10.1016/j.chemolab.2007.10.001.
32. Bremner JM. Determination of nitrogen in soil by the kjeldahl method. J Agric Sci. 1960;55(1):11–33. https://doi.org/10.1017/S0021859600021572.
33. Tian T, Wang J, Wang H, Cui J, Shi X, Song J, Li T, Li W, Zhong M. Synergistic use of spectral features of leaf nitrogen and physiological indices improves the estimation accuracy of nitrogen concentration in rapeseed. Int J Remote Sens. 2022;43(8):2755–76. https://doi.org/10.1080/01431161.2022.2068359.
34. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test. 2018;2(3):249–62. https://doi.org/10.1007/s41664-018-0068-2.
35. Yuan R, Liu G, He J, Ma C, Cheng L, Fan N, Ban J, Li Y, Sun Y. Determination of metmyoglobin in cooked tan mutton using VIS/NIR hyperspectral imaging system. J Food Sci. 2020;85(5):1403–10. https://doi.org/10.1111/1750-3841.15137.
36. Ge X, Ding J, Teng D, Xie B, Zhang X, Wang J, Han L, Bao Q, Wang J. Exploring the capability of Gaofen-5 hyperspectral data for assessing soil salinity risks. Int J Appl Earth Obs Geoinf. 2022;112: 102969. https://doi.org/10.1016/j.jag.2022.102969.
37. Rossel RAV. Robust modelling of soil diffuse reflectance spectra by "bagging-partial least squares regression". J Near Infrared Spectrosc. 2007;15(1):39–47. https://doi.org/10.1255/jnirs.694.
38. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59. https://doi.org/10.1016/S0893-6080(05)80023-1.
39. Han G, Chen S, Wang X, Wang J, Wang H, Zhao Z. Noninvasive blood glucose sensing by near-infrared spectroscopy based on PLSR combines SAE deep neural network approach. Infrared Phys Technol. 2021;113: 103620. https://doi.org/10.1016/j.infrared.2020.103620.
40. de Santana FB, de Giuseppe LO, de Souza AM, Poppi RJ. Removing the moisture effect in soil organic matter determination using NIR spectroscopy and PLSR with external parameter orthogonalization. Microchem J. 2019;145:1094–101. https://doi.org/10.1016/j.microc.2018.12.027.
41. Houborg R, McCabe MF. A hybrid training approach for leaf area index estimation via cubist and random forests machine-learning. ISPRS J Photogramm Remote Sens. 2018;135:173–88. https://doi.org/10.1016/j.isprsjprs.2017.10.004.
42. Verrelst J, Rivera JP, Gitelson A, Delegido J, Moreno J, Camps-Valls G. Spectral band selection for vegetation properties retrieval using gaussian processes regression. Int J Appl Earth Obs Geoinf. 2016;52:554–67. https://doi.org/10.1016/j.jag.2016.07.016.
43. Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, Roger J-M, McBratney A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by nir spectroscopy. TrAC, Trends Anal Chem. 2010;29(9):1073–81. https://doi.org/10.1016/j.trac.2010.05.006.
44. Peng J, Biswas A, Jiang Q, Zhao R, Hu J, Hu B, Shi Z. Estimating soil salinity from remote sensing and terrain data in Southern Xinjiang Province, China. Geoderma. 2019;337:1309–19. https://doi.org/10.1016/j.geoderma.2018.08.006.
45. An X, Li M, Zheng L, Sun H. Eliminating the interference of soil moisture and particle size on predicting soil total nitrogen content using a NIRS-based portable detector. Comput Electron Agric. 2015;112:47–53. https://doi.org/10.1016/j.compag.2014.11.003.
46. Xu S, Zhao Y, Wang M, Shi X. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by vis-nir spectroscopy. Geoderma. 2018;310:29–43. https://doi.org/10.1016/j.geoderma.2017.09.013.
47. van der Meer F. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. Int J Appl Earth Obs Geoinf. 2006;8(1):3–17. https://doi.org/10.1016/j.jag.2005.06.001.
48. Wang L, Zhou X, Zhu X, Dong Z, Guo W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. Crop J. 2016;4(3):212–9. https://doi.org/10.1016/j.cj.2016.01.008.
49. Chen X, Wu D, He Y, Liu S. Detecting the quality of glycerol monolaurate: A method for using fourier transform infrared spectroscopy with wavelet transform and modified uninformative variable elimination. Anal Chim Acta. 2009;638(1):16–22. https://doi.org/10.1016/j.aca.2009.02.002.
50. Xiao S, He Y, Dong T, Nie P. Spectral analysis and sensitive waveband determination based on nitrogen detection of different soil types using near infrared sensors. Sensors. 2018. https://doi.org/10.3390/s18020523.
51. Cao C, Wang T, Gao M, Li Y, Li D, Zhang H. Hyperspectral inversion of nitrogen content in maize leaves based on different dimensionality reduction algorithms. Comput Electron Agric. 2021;190: 106461. https://doi.org/10.1016/j.compag.2021.106461.
52. Jacquemoud S, Bacour C, Poilvé H, Frangi J-P. Comparison of four radiative transfer models to simulate plant canopies reflectance: direct and inverse mode. Remote Sens Environ. 2000;74(3):471–81. https://doi.org/10.1016/S0034-4257(00)00139-5.

Publisher's Note