RESEARCH ARTICLE

ATR-FTIR Spectral Characteristics and Their Application to the Classification of Cultivated Karst Soils in Shilin County

Xing Yang[1], Bixia Jiang[1], Yanjie He[1], Rui Yue[1], Ruiqian Bao[1], Gaoqiang Lv[2,3,*], Fei Ma[4] and Changwen Du[4]

[1] Kunming Branch, Yunnan Tobacco Company, Kunming 650000, China
[2] College of Agricultural Engineering, Jiangsu University, Zhenjiang 212013, China
[3] Key Laboratory of Crop Production and Smart Agriculture in Yunnan Province, Kunming 650000, China
[4] Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210000, China

Abstract

To enable rapid, non-destructive analysis and functional classification of karst calcareous soils in Shilin County, Kunming, this study acquired mid-infrared attenuated total reflectance Fourier-transform infrared (ATR-FTIR) spectra in the 4000-900 cm^{-1} range from 409 soil samples. After standardized preprocessing, the samples were preliminarily partitioned into four spectral groups by K-means clustering, and a partial least squares discriminant analysis (PLS-DA) model was constructed. Variable importance in projection (VIP) scores were further used to identify key spectral features distinguishing the classes. The spectra were dominated by peaks at 3620, 1650, and 1020 cm^{-1}, and no obvious calcite absorption at ~1430 cm^{-1} was observed, suggesting strong carbonate leaching (decalcification weathering). K-means clustering divided the samples into four groups (G1–G4), interpreted from their spectral features as organic matter-mineral impoverished, mineral-dominated, organic matter-mineral enriched, and organic matter impoverished types, respectively, possibly reflecting spectral differentiation along a rocky desertification gradient. The PLS-DA model achieved an overall accuracy of 93.9% under nested cross-validation, with precision and recall exceeding 92% for G1–G3. VIP analysis identified the silicate skeleton region, the organic matter/water region, and the hydroxyl region as key bands, suggesting that gradients in clay mineral, organic matter, and bound water contents contribute most to spectral differentiation. These findings indicate that ATR-FTIR spectroscopy has potential for rapid, non-destructive classification of karst calcareous soils and can provide a technical pathway for soil survey and rocky desertification monitoring.

Submitted: 11 July 2026
Revised: 15 September 2026
Accepted: 17 September 2026
Published: 24 September 2026

Vol. 2, **No.** 3, 2026.
10.62762/DIA.2026.518539

***Corresponding author:**
✉ Gaoqiang Lv
gqlv@ujs.edu.cn

Citation
Yang, X., Jiang, B., He, Y., Yue, R., Bao, R., Lv, G., Ma, F., & Du, C. (2026). ATR-FTIR Spectral Characteristics and Their Application to the Classification of Cultivated Karst Soils in Shilin County. *Digital Intelligence in Agriculture*, 2(3), 179–190.

Keywords: Karst calcareous soil, spectral characteristics, soil classification, ATR-FTIR, PLS-DA.

1 Introduction

Karst rocky desertification (KRD) is a major global ecological and environmental problem [1]. The karst region of southwest China is one of the world's three largest contiguous karst regions, where rocky desertification poses a severe threat [2, 3]. Accordingly, rapid and accurate assessment of soil type differentiation and quality in cultivated land of this region is of great importance for rocky desertification monitoring, soil resource management, and ecological restoration decision-making. Karst calcareous soil is a distinctive soil type of this landform. Its mineral composition is dominated by phyllosilicate clay minerals (e.g., kaolinite and illite) and quartz, while organic matter content and its occurrence state are jointly regulated by biological accumulation and leaching, giving rise to high spatial heterogeneity. Conventional soil classification and property determination rely on field profile surveys combined with laboratory physicochemical analyses, such as organic carbon determination by the potassium dichromate oxidation method and mineral composition identification by X-ray diffraction (XRD) [4]. Although these methods are accurate, they are labor-intensive, time-consuming, and costly, and cannot meet the demand for rapid screening and dynamic monitoring of soil resources over large areas. Therefore, the development of efficient, non-destructive, and low-cost techniques for rapid soil analysis has become a research focus in soil science and ecology, with Fourier-transform infrared (FTIR) spectroscopy emerging as a promising alternative for rapid soil constituent characterization [5].

Mid-infrared (MIR) spectroscopy provides a powerful molecular fingerprinting tool for rapid soil analysis. The MIR spectrum of soil (4000–400 cm^{-1}) reflects the combined molecular response of organic functional groups and inorganic mineral constituents. Because soil types differ in pedogenic conditions, they exhibit systematic differentiation in the positions and intensities of characteristic absorption peaks, forming unique "spectral fingerprints" [6]. Over the past two decades, the application of infrared spectroscopy in soil analysis has expanded into several important directions. For soil organic matter analysis, FTIR spectroscopy can effectively characterize the functional group composition and structural features of organic fractions such as humus and fulvic acid. The broad hydroxyl absorption band near 3400 cm^{-1}, the Amide I band at 1650 cm^{-1}, and the Amide II band at 1550 cm^{-1} collectively reveal the humification degree and molecular structural stability of soil organic matter [7, 8]. For soil mineral analysis, MIR spectroscopy is highly sensitive to phyllosilicate clay minerals; kaolinite, illite, and montmorillonite display characteristic absorption differences at 3620 cm^{-1} (structural hydroxyl), 1020 cm^{-1} (Si-O-Si skeleton), and in the 800–1200 cm^{-1} range, and have been successfully used for qualitative identification and semi-quantitative estimation of soil clay minerals [9, 10]. For soil classification and type discrimination, full-spectrum fingerprint MIR spectroscopy combined with chemometric models can distinguish soil types, with reported overall accuracies of about 62–89% depending on the spectral range and classifier [11].

Because soil spectral data are characterized by high dimensionality, strong collinearity, and high background noise, accurate discrimination by visual spectral interpretation alone is difficult. The introduction of chemometric methods offers an effective means for in-depth mining of spectral fingerprint information: unsupervised clustering algorithms (e.g., K-means) can reveal natural differentiation patterns of soil samples based on spectral similarity, whereas supervised discriminant models (e.g., partial least squares discriminant analysis, PLS-DA) discriminate soil types by projecting the spectra onto latent variables that maximize the covariance between the spectral matrix and the class-membership matrix; their variable importance in projection (VIP) analysis can further identify the key spectral response regions driving classification [12, 13]. For soil quality assessment and degradation monitoring, the integrated strategy of "fingerprint FTIR spectroscopy + chemometrics" has been successfully applied to assess soil ecosystem disturbance and recovery trajectories caused by forest harvesting and land-use changes, by tracking spectral dynamics of soil organic matter and reactive mineral components [14]. These studies demonstrate that MIR spectroscopy is not merely a tool for soil constituent analysis but also a systematic diagnostic means that can comprehensively reflect soil development status and degradation degree.

Karst calcareous soils exhibit distinctive regional spectral characteristics. Their mineral assemblage is dominated by quartz and secondary clay minerals, with kaolinite and illite as the principal clay minerals, which produce characteristic spectral

signals such as a sharp structural hydroxyl peak of clay minerals at 3620 cm^{-1} and a strong silicate skeleton absorption at 1020 cm^{-1} [4, 15]. Meanwhile, soils at different rocky desertification degrees differ significantly in organic matter and mineral abundance, providing a material basis for classification by MIR spectroscopy [16]. However, existing soil spectral classification studies have primarily drawn on globally compiled spectral libraries and targeted zonal soils in non-karst regions [17], with site-specific ATR-FTIR discrimination efforts concentrated in temperate and agricultural settings [18–20]; systematic spectral classification of calcareous soils in the karst region of southwest China therefore remains relatively scarce. Therefore, this study took cultivated soils in the Shilin area of Yunnan Province as the research object, employed attenuated total reflectance Fourier-transform infrared (ATR-FTIR) spectroscopy combined with chemometric methods to systematically analyze soil spectral fingerprints, and explored their application in soil classification. The aim is to provide a technical pathway for rapid, non-destructive detection of karst calcareous soils, as well as a scientific basis for regional soil quality survey and rocky desertification ecological monitoring.

2 Materials and methods

2.1 Study Area

The study area is located in Shilin Yi Autonomous County, Kunming City, Yunnan Province (24°30′N–25°03′N, 103°10′E–104°40′E), situated in the core karst landform distribution zone of the Yunnan-Guizhou Plateau, at an elevation of 1500–2100 m. The region has a subtropical monsoon climate, with an annual mean temperature of 16.3 °C and annual mean precipitation of approximately 960 mm, with distinct dry and wet seasons. The parent rocks are mainly Permian limestone and dolomite, and the dominant soil type is calcareous soil. The soil layer is shallow (mostly <50 cm), the texture ranges from clay loam to clay, and organic matter content varies widely in space. Cultivated land in the study area consists mainly of sloping and depression farmland. Affected by topographic relief, rocky desertification gradients, and human cultivation activities, soil physicochemical properties exhibit pronounced spatial heterogeneity.

2.2 Soil Sampling and Preparation

A total of 409 cultivated soil samples were collected from the tillage layer (0–20 cm) of 7 townships or regions, including Banqiao (22 cases), Changhu (102 cases), Dake (37 cases), Guishan (87 cases), Lufu (38 cases), Shilin (48 cases), and Xijiekou (75 cases) in 2024. At each sampling site, approximately 1 kg of soil was collected by the five-point composite method, sealed in plastic bags, and transported to the laboratory. After air-drying, plant residues, gravel, and other impurities were removed; the samples were ground and passed through a 0.25 mm (60-mesh) sieve, thoroughly mixed, and stored in a desiccator until use. This preparation protocol follows established procedures for mid-infrared diffuse reflectance spectroscopic analysis of agricultural soils, which have demonstrated that fine grinding and sieving to ≤0.25 mm are critical for obtaining reproducible mid-infrared spectra [21].

2.3 ATR-FTIR Spectral Measurement

Mid-infrared spectra of soil samples were acquired using a Fourier-transform infrared spectrometer equipped with an attenuated total reflectance (ATR) accessory. The instrument parameters were set as follows: scanning range 4000–900 cm^{-1}, resolution 4 cm^{-1}, 64 scans, with automatic background subtraction against air. During measurement, an appropriate amount of sieved soil was placed directly on the diamond ATR crystal surface and pressed into intimate contact with the crystal using a pressure device, and spectral data were recorded. Each sample was measured in triplicate, and the average spectrum was used for subsequent analysis [22, 23]. All spectral acquisitions were performed under constant temperature (25 °C) and humidity (relative humidity <50%) conditions to minimize the influence of environmental fluctuations on the spectral baseline.

2.4 Data Preprocessing

The raw spectra were first smoothed by the Savitzky-Golay method (window width 15 points, polynomial order 2) to remove high-frequency noise, followed by standard normal variate (SNV) transformation to eliminate scattering effects caused by differences in particle size and optical path length [24, 25]. To facilitate comparison of absorption intensities among samples, the preprocessed spectra were subjected to vector normalization. All preprocessing steps and subsequent data analyses were performed in MATLAB R2020b (MathWorks, Natick, MA, USA).

2.5 Chemometric Analysis

2.5.1 K-means Clustering Analysis

Using the preprocessed full-spectrum data (4000–900 cm^{-1}) as input variables, the K-means algorithm was applied to perform unsupervised clustering of the 409 soil samples. Prior to clustering, principal component analysis (PCA) was applied to the spectral data for dimensionality reduction. Principal components with cumulative variance contribution >95% were extracted as inputs for clustering to mitigate collinearity interference [26]. The optimal number of clusters k was determined by the Elbow Method combined with the Silhouette Score, and the samples were ultimately partitioned into four spectral groups (G1–G4). The mean spectral curve of each group was calculated, and the physicochemical composition characteristics of each group were interpreted in conjunction with the vibrational assignments of infrared functional groups. One-way ANOVA confirmed that absorption intensities at the key diagnostic bands (3620, 1650, and 1020 cm^{-1}) differed significantly among the four clusters ($P < 0.05$, Tukey's post hoc test).

2.5.2 Partial Least Squares Discriminant Analysis (PLS-DA)

A supervised PLS-DA model was constructed using K-means-derived soil class labels (G1, G2, G3, G4) as response variables and preprocessed full-spectrum absorbance data as predictor variables. A nested cross-validation strategy was applied to avoid data leakage in the selection of the number of latent variables (LVs), following variable optimization practices established for MIR soil spectral modeling [27]. In the outer loop, leave-one-out cross-validation (LOOCV) over the full dataset ($n = 409$) served for final model evaluation. For each outer iteration, one sample was left out for prediction; the remaining samples formed an inner training subset, where 10-fold cross-validation (CV) was used to optimize the number of LVs by minimizing cross-validated classification error. The model trained on the inner subset predicted the held-out sample. After completing all outer-loop iterations, classification accuracy, precision, recall, and the confusion matrix were aggregated from predictions of all samples.

2.5.3 Variable Importance in Projection (VIP) Analysis

Based on the optimal PLS-DA model, the VIP value of each spectral band was calculated as

$$\mathrm{VIP}_j = \sqrt{\frac{p \sum_{a=1}^{A} \mathrm{SSY}_a \left(w_{ja}/\|\mathbf{w}_a\|\right)^2}{\sum_{a=1}^{A} \mathrm{SSY}_a}}, \quad (1)$$

where p is the number of spectral variables (wavenumbers), A is the number of latent variables, SSY_a is the sum of squares of the response explained by the a-th latent variable, and w_{ja} is the weight of variable j in the a-th latent variable. Bands with VIP > 1.0 were selected as key bands contributing significantly to classification [28]. The soil constituent information corresponding to high-VIP bands was interpreted by reference to infrared functional group vibrational assignments, to elucidate the core spectral response regions driving soil type differentiation.

2.6 Statistical Analysis

Spectral differences among K-means clusters were tested for significance by one-way analysis of variance (ANOVA) followed by Tukey's post hoc test ($P < 0.05$). The classification performance metrics of the PLS-DA model were calculated as follows:

$$\mathrm{Accuracy} = \frac{\sum_{i=1}^{K} \mathrm{TP}_i}{N}, \quad (2)$$

$$\mathrm{Precision}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}, \quad (3)$$

$$\mathrm{Recall}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}, \quad (4)$$

where $K = 4$ is the number of classes, $N = 409$ is the total number of samples, and TP_i, FP_i, and FN_i denote the numbers of true positive, false positive, and false negative samples of class i (one-versus-rest), respectively [29, 30].

3 Results and Discussion

3.1 Spectral Characteristics of Cultivated Soils in the Shilin Area

As shown in Figure 1, the attenuated total reflectance Fourier-transform infrared (ATR-FTIR) spectra of the 409 soils from the Shilin area cover the MIR range of 4000-900 cm^{-1}. All samples exhibited highly consistent spectral peak shapes and characteristic peak positions, with only continuous gradient variations in absorption intensity, indicating that the functional groups and mineral composition types of soils in the study area were qualitatively similar, and that

only the relative contents of constituents exhibited spatial or profile differentiation. Overall, characteristic absorptions were concentrated in three regions: (1) a broad hydroxyl absorption band at 3700–3000 cm^{-1}, in which a sharp peak at 3620 cm^{-1}, a main peak at 3434 cm^{-1}, a shoulder at 3369 cm^{-1}, and a weak absorption at 3190 cm^{-1} were distributed sequentially from high to low wavenumber; (2) a medium-intensity absorption peak at 1650 cm^{-1} and a weak absorption peak at 1550 cm^{-1}; and (3) the strongest absorption peak of the entire spectrum at 1020 cm^{-1}, representing the core spectral response of the soil solid-phase skeleton constituents.

The characteristic peaks were assigned to specific substances based on classical vibrational mode assignments in soil infrared spectroscopy and previous studies: the sharp peak at 3620 cm^{-1} is attributed to the stretching vibration of inner structural hydroxyl (O-H) groups in phyllosilicate clay minerals such as kaolinite and illite, and can serve as a qualitative indicator of secondary clay minerals in soil [9, 10]; the continuous broad absorption band formed by 3434, 3369, and 3190 cm^{-1} results from the superposition of multiple hydrogen-containing group vibrations, corresponding respectively to the asymmetric O-H stretching of adsorbed water on soil particle surfaces, the stretching of interlayer bound water in clay minerals and amino (N-H) groups in organic matter, and the vibrations of strongly hydrogen-bonded carboxylic hydroxyl groups in humus and mineral hydration hydroxyl groups, collectively reflecting the occurrence state of bound water and oxygen-/nitrogen-containing organic matter in soil [7, 8]; the peak at 1650 cm^{-1} arises from the superposition of the H-O-H bending vibration of adsorbed water and the Amide I band (C=O stretching vibration) of soil organic matter, while the peak at 1550 cm^{-1} originates mainly from the Amide II (N-H) bending vibration [31, 32]; the strong peak at 1020 cm^{-1} is attributed to the asymmetric stretching vibration of the Si-O-Si skeleton in silicate minerals, indicating that clay minerals and quartz constitute the main skeleton of the soil solid phase [33, 34].

It should be noted that the spectral scanning range of this study was 900–4000 cm^{-1}, which only covered the ν_3 stretching band (~1430 cm^{-1}) of calcite. The diagnostic ν_2 (~875 cm^{-1}) and ν_4 (~713 cm^{-1}) bending vibration bands of calcite fell outside the detection window. No obvious characteristic absorption of calcite at 1430 cm^{-1} was observed in all soil samples. Since the absence of a spectral peak is only indirect evidence, which may be affected by the instrument detection limit and spectral peak overlap, this result cannot independently confirm the complete disappearance of carbonate minerals. Nevertheless, combining the dominant spectral responses of the silicate skeleton (1020 cm^{-1}) and clay-mineral structural hydroxyl groups (3620 cm^{-1}), the spectral features suggested strong carbonate leaching and high-degree decalcification weathering of local calcareous soils, which matches the long-term karst corrosion background in the Shilin area [1, 2].

Moreover, the sharp hydroxyl peak of clay minerals at 3620 cm^{-1} reflects that the soils are at an intermediate weathering stage, with kaolinite and illite as the probable dominant clay minerals (crystallinity was not assessed), which is consistent with reports that residual calcium and magnesium in calcareous soils inhibit the desilication and allitization process [4, 15]. In addition, the broadening and multi-peak superposition of the hydroxyl absorption band in the 3434–3190 cm^{-1} range indirectly reflect the ternary interaction among minerals, water, and organic matter in soil, which may be related to the strong organic matter retention capacity in the shallow karst soil layer [16]. Overall, ATR-FTIR technology could rapidly and non-destructively characterize the mineral and organic matter composition of cultivated soils in the Shilin area, providing an important basis for subsequent soil classification analysis.

3.2 K-means Clustering of Soil Spectra and Characterization of Clusters

To distinguish the differentiation patterns of soil samples, K-means clustering was applied to the ATR-FTIR full-spectrum data of the 409 cultivated soils from the Shilin area, partitioning them into four groups: G1 (97 samples), G2 (177 samples), G3 (87 samples), and G4 (48 samples). The scatter plot in the factor space defined by the first two factor variables (FV1, FV2) (Figure 2(a)) shows that the four groups exhibited a differentiation pattern in the factor space: G1 and G4 were both distributed on the positive half-axis of FV1, occupying the mid-to-high and low value ranges of FV2, respectively; G2 was concentrated in the negative half-axis region of FV1; G3 clustered in the mid-value range of FV1 with the highest FV2 scores. The cluster centers of the four groups were clearly separated in space, although partial overlap between G1, G2, and G4 was visible, indicating that K-means clustering achieved a reasonable partition of soil spectra, with pronounced within-group spectral similarity and

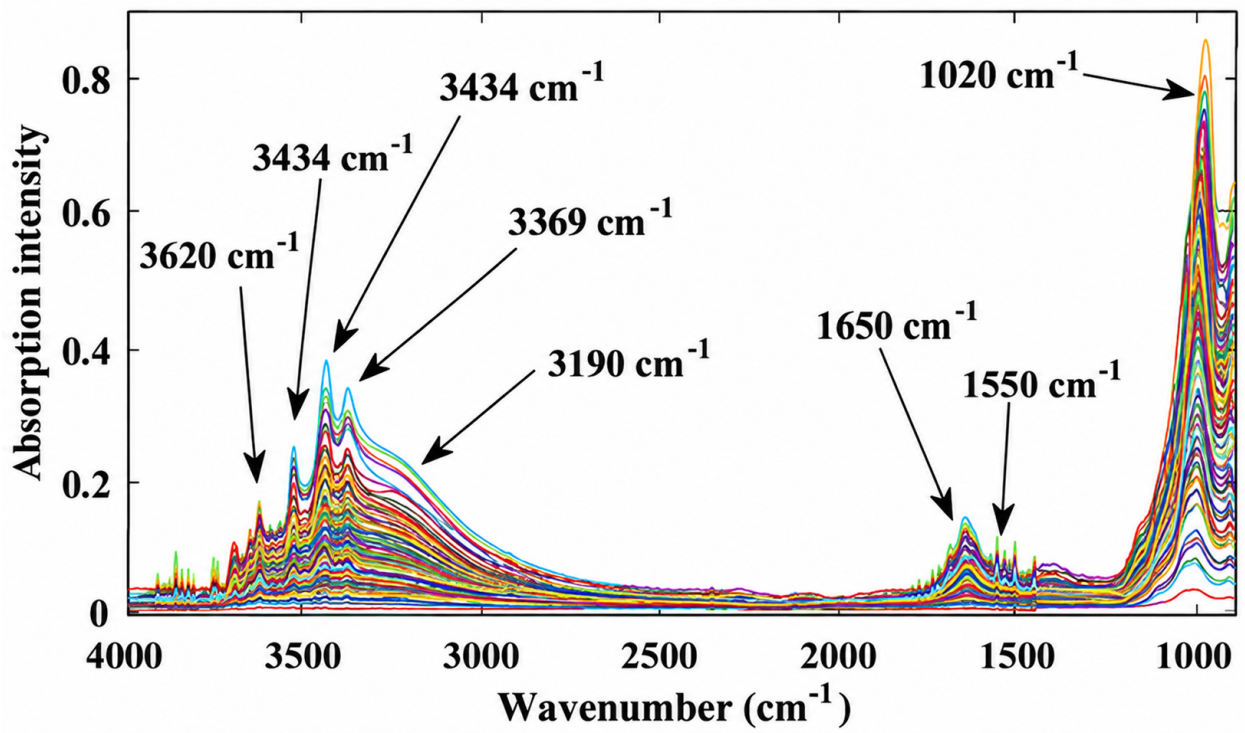

Figure 1. ATR-FTIR spectra of 409 soil samples from the Shilin area in the 4000–900 cm^{-1} range, with major absorption peaks indicated by arrows.

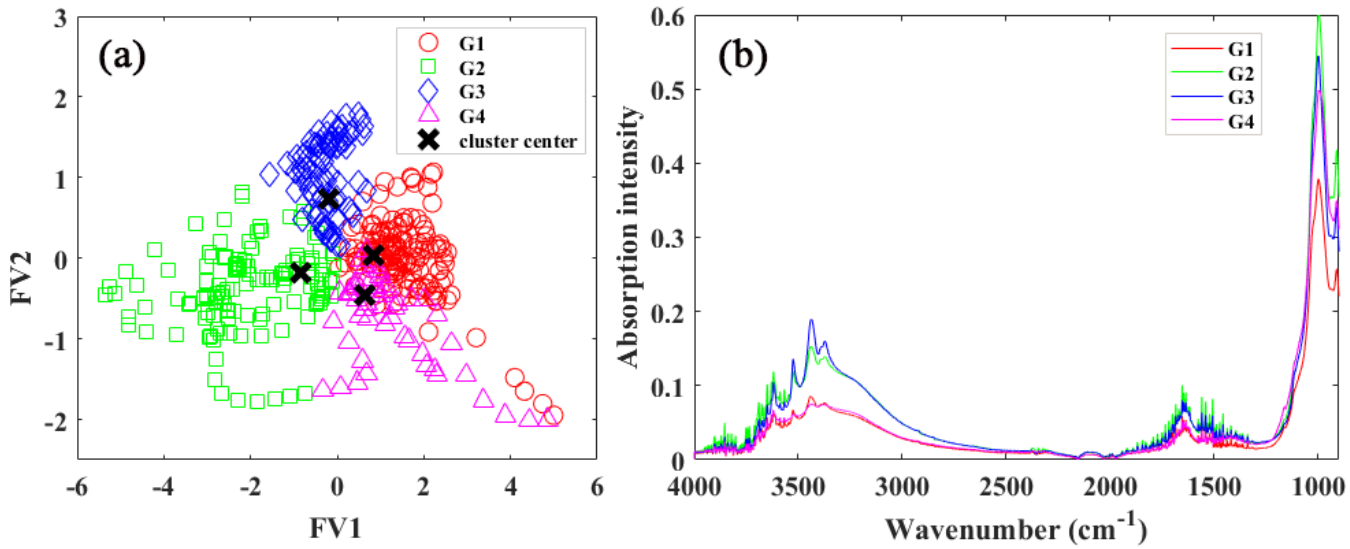

Figure 2. K-means clustering of 409 samples into four groups (a) and the mean spectrum of each cluster (G1–G4) (b).

between-group differences. The mean spectral curves of the four groups (Figure 2(b)) show that all groups had identical characteristic peak positions, exhibiting the typical broad hydroxyl absorption band at 3700–3000 cm^{-1}, the water deformation peak at 1650 cm^{-1}, and the strong silicate absorption peak at 1020 cm^{-1}. However, the absorption intensities differed markedly among groups, and the intensity variation patterns across bands were not entirely synchronized, reflecting that the four soil types possess distinct characteristics in material composition and relative proportions.

Further interpretation based on infrared functional group vibrational assignments revealed clear differentiation in the physicochemical composition of the four soil types: G3 exhibited the highest absorption intensities at both the clay mineral structural hydroxyl peak (3620 cm^{-1}) and the adsorbed water/organic matter hydroxyl peak (3434 cm^{-1}) among the four groups, suggesting the highest clay mineral content,

bound water, and organic functional group abundance, corresponding to clayey, organic-matter-rich soils; G2 showed the highest Si-O-Si skeleton vibration peak intensity at 1020 cm^{-1} and the second-highest hydroxyl region absorption, suggesting that silicate clay minerals are the dominant constituent, with the highest mineral skeleton proportion and intermediate organic matter and water contents; G1 exhibited overall low absorption across the full spectrum, with weak absorptions in both the hydroxyl region and the silicate characteristic peak, reflecting low overall contents of both clay and organic matter; G4 showed the lowest hydroxyl region absorption among the four groups, but its 1020 cm^{-1} absorption was higher than that of G1, suggesting extremely scarce organic matter and bound water while silicate minerals still maintain a certain proportion, presenting a composition characterized by high mineral proportion and organic matter impoverishment. From the perspective of information carried by the factor dimensions, the FV1 axis mainly reflects the gradient variation in the relative content of silicate minerals, while the FV2 axis mainly corresponds to the integrated information of soil hydroxyl groups, water, and organic functional groups; together they constitute the core dimensions of spectral variation in karst calcareous soils.

The karst region of Shilin, Yunnan, features large topographic relief and a pronounced rocky desertification gradient, and the four spectral groups may correspond to different soil development stages and ecological degradation degrees. G3 most likely corresponds to calcareous soils in depressions or forestland with good vegetation cover, where biological accumulation is intense and both organic matter and clay contents are high; G2 may correspond to leached calcareous soils on footslopes, where clay minerals have migrated and accumulated via surface processes and the mineral component proportion is prominent; G1 and G4 may correspond to shallow soils under different degrees of rocky desertification, with thin soil layers and severe organic matter loss, among which G4 more closely resembles the characteristics of initial soils in the early stage of rocky desertification [35–37]. These attributions were inferential and were not validated against measured site attributes such as land use, slope position, or rocky desertification grade. K-means clustering combined with MIR ATR-FTIR spectroscopy may therefore enable rapid soil type classification and degradation degree identification in karst areas, allowing attribute discrimination of large batches of samples without cumbersome chemical determinations, and providing an efficient, low-cost technical pathway for large-scale rocky desertification monitoring and soil quality survey [38]. Meanwhile, the categorical differentiation characteristics of soil spectra also suggest that in subsequent spectral quantitative inversion of soil properties, introducing classification variables for group-specific modeling can effectively reduce the interference of between-group compositional differences and further improve prediction accuracy [39, 40].

3.3 Performance of the PLS-DA Classification Model

To accurately discriminate soil spectral types, a PLS-DA discriminant model was constructed based on the MIR ATR-FTIR spectra using the four soil classes (G1, G2, G3, G4) derived from K-means clustering as labels; the results are shown in Figure 3. The model achieved an overall classification accuracy of 93.9%, with a misclassification rate of only 6.1%. Misclassified samples were mainly concentrated in the G4 predicted group: 10 of the 56 samples predicted as G4 (17.9%) were misclassified, being actually G1 (1 sample) and G2 (9 samples); the misclassification rates of the other predicted classes were 3.2%, 3.6%, and 6.6% for G1, G2, and G3, respectively, all below 7%.

Among all 409 samples, the numbers of correctly classified samples in G1, G2, G3, and G4 were 90, 163, 85, and 46, respectively. The per-class precision values were 96.8% (G1), 96.4% (G2), 93.4% (G3), and 82.1% (G4); the per-class recall values were 92.8% (G1), 92.1% (G2), 97.7% (G3), and 95.8% (G4). Precision reflects the reliability of model predictions. The precision of both G1 and G2 exceeded 96%, indicating that samples classified as G1 or G2 by the model were rarely misclassified, and the infrared spectral fingerprint features of these two soil types were highly distinguishable. The precision of G3 decreased slightly to 93.4%, with a small number of samples predicted as G3 actually belonging to G1 or G2, suggesting slight band information overlap between G3 and the G1/G2 classes. The precision of G4 was only 82.1%, the lowest among the four groups; the vast majority of misclassified samples (9 cases) were confused with G2, and only 1 case came from G1. It is worth noting that among the 48 samples in G4, only 2 were misclassified as G2. G4 and G2 were the clusters with the lowest (48) and highest (177) sample counts, respectively. Because 9 G2 samples were assigned to the smaller G4 class, spectral overlap between G2 and G4 (Figure 2(a))

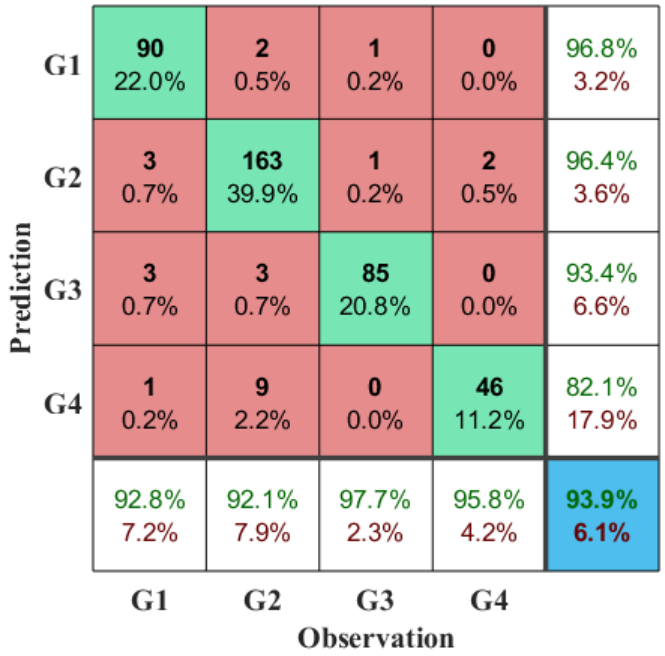

Figure 3. Classification performance of the PLS-DA model for the four soil classes. Matrix rows represent model-predicted classes and columns represent true observed classes from clustering; diagonal green cells indicate correctly classified sample counts (with percentages of all samples), off-diagonal red cells indicate misclassified samples, the right-hand white cells indicate the precision and misclassification rate of each predicted class, the bottom white cells indicate the recall and miss rate of each observed class, and the bottom-right blue cell indicates the overall model accuracy and overall misclassification rate.

is the more likely reason for the low precision of G4, although the class-size imbalance may also have contributed to blurred classification boundaries [30].

In addition, recall characterizes the model's ability to capture samples of a true class. G3 had the highest recall (97.7%) among the four groups, with almost all true G3 samples correctly identified, probably attributable to its higher absorption intensity in the hydroxyl functional group bands compared with the other three groups, giving it outstanding feature distinctiveness. The recall values of G1, G2, and G4 all exceeded 92%, indicating that the model could stably capture the spectral features of each of the four soil types, with very few true samples being missed. Overall, the 93.9% overall accuracy of the PLS-DA model demonstrated that MIR ATR-FTIR spectroscopy combined with partial least squares discriminant analysis could reproduce the K-means-derived cluster types of karst calcareous soils rapidly and non-destructively, which is suited to the automated classification needs of large batches of soil samples [11, 12]. Future work could employ second-derivative spectral preprocessing to amplify subtle differences in hydroxyl bands, expand the G4 sample library, or construct a hierarchical two-stage PLS-DA model to reduce the spectral information overlap between G2 and G4 and improve overall model performance. Additionally, integrating particle size information alongside spectral data, as demonstrated by mid-infrared ATR spectroscopy studies targeting soil carbon and texture simultaneously, may further enhance classification robustness [41].

It should be noted that this study has limitations regarding the coverage of rocky desertification gradients. Constrained by the accessibility of field sampling, moderate and severe rocky desertification plots within the study area are mostly distributed in core protected areas and on steep slopes, which renders systematic sampling impracticable. Accordingly, this study only incorporated samples from three rocky desertification gradients: non-obvious, potential, and mild. This also indicates that the ATR-FTIR-PLS-DA soil classification model developed in this study is currently only applicable to the discrimination of karst calcareous soils within the aforementioned gradient ranges, and direct extrapolation to soils affected by moderate or severe rocky desertification carries uncertainty. In subsequent work, we will carry out targeted supplementary sampling in areas with moderate and severe rocky desertification to improve the coverage of sample gradients, so as to further expand the application scope and enhance the generalization performance of the model. Meanwhile, the core discriminant wavenumber bands and the methodological framework of spectral classification revealed in this study can provide a fundamental basis for future research on full-gradient spectral classification of rocky desertification soils.

3.4 Variable Importance Analysis of the PLS-DA Classification Model

To investigate the spectral features that play a key role in classification, the VIP value of each spectral band in the PLS-DA model was calculated, and the contribution rate was analyzed using VIP $>$ 1 as the threshold; the results are shown in Figure 4. VIP values exhibited significant regional aggregation across the full spectrum, with high-contribution bands concentrated in three independent regions: (1) the high-wavenumber hydroxyl vibration region at 3800–3500 cm^{-1}, where multiple sharp peaks with

VIP > 1 were distributed, with a maximum value of approximately 1.4; (2) the organic functional group and water deformation vibration region at 1800–1500 cm^{-1}, where VIP values were generally elevated to form a continuous high-value band, with a peak reaching approximately 1.8; and (3) the silicate skeleton vibration region at 1200–1000 cm^{-1}, where VIP values rose rapidly and formed the highest peak in the full spectrum, at approximately 1.8 (comparable to the peak in the 1800–1500 cm^{-1} region). VIP values in the remaining bands were generally below 1, with values of only 0.2–0.5 in the 3000–2500 cm^{-1} and 2200–1800 cm^{-1} ranges, contributing weakly to classification.

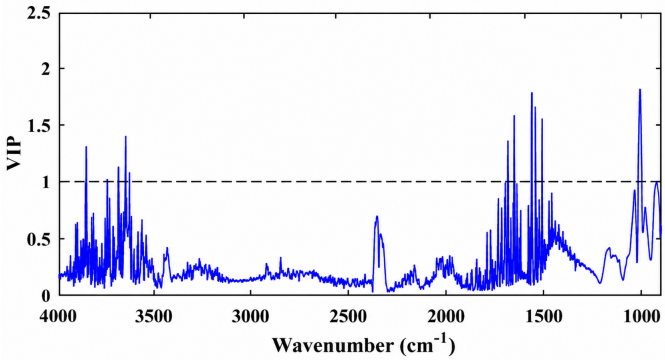

Figure 4. VIP value analysis results from the PLS-DA model.

Interpretation based on infrared functional group vibrational assignments showed that the three high-VIP regions correspond well to the core characteristics of compositional differences among the four soil types, consistent with the results from the mean spectra and confusion matrix. The Si-O-Si skeleton vibration region near 1000 cm^{-1} had the highest VIP values in the full spectrum, indicating that the relative content of silicate clay minerals contributes the most to the model's classification decisions, corresponding to the overall gradient variation in clay content among the four soil types. The 1800–1500 cm^{-1} region had comparably high VIP values, encompassing the H-O-H bending vibration of water at 1650 cm^{-1}, the C=O stretching vibration of the organic matter Amide I band, and the Amide II band and carboxylate vibration signals near 1500 cm^{-1}; the continuous high VIP values in this region indicate that gradient differences in soil water and organic functional group composition constitute the core discriminant basis for the model to distinguish soils at different development degrees. The 3800–3500 cm^{-1} region had the weakest VIP values among the three, covering the inner structural hydroxyl groups of clay minerals (near 3620 cm^{-1}), soil adsorbed water, and hydroxyl stretching vibrations of organic matter; the fact that VIP values in this region exceeded the threshold indicates that differences in clay mineral abundance, bound water content, and organic hydroxyl functional groups form an important basis supporting soil class discrimination, corresponding to the between-group differentiation characteristics of high hydroxyl absorption in G3 and low hydroxyl absorption in G4. These findings are consistent with broader evidence that near-infrared and mid-infrared spectroscopy can reliably track soil organic matter quality and carbon pool dynamics across landscape gradients [42, 43]. Thus, ATR-FTIR analysis of cultivated soils in Shilin demonstrated clear differentiation in clay-mineral abundance, soil organic-matter content, and bound-water content across samples. These differentiation characteristics may provide a basis for subsequent soil management and support the feasibility of ATR-FTIR spectroscopy in soil classification applications.

4 Conclusions

In this study, 409 cultivated soil samples from Shilin, Yunnan, were used to investigate soil classification by ATR-FTIR spectroscopy combined with chemometric methods. The soil spectra were dominated by characteristic peaks at 3620 cm^{-1} (clay mineral hydroxyl), 1650 cm^{-1} (Amide I band and water bending), and 1020 cm^{-1} (silicate skeleton), and no obvious calcite absorption at ~1430 cm^{-1} was observed, suggesting strong decalcification weathering. K-means clustering partitioned the samples into four clusters (G1–G4), interpreted as organic-matter–mineral-impoverished, mineral-dominated, organic-matter–mineral-enriched, and organic-matter-impoverished types, respectively, which may reflect spectral differentiation along the rocky desertification gradient. The PLS-DA model achieved an overall accuracy of 93.9%, with precision and recall exceeding 92% for G1–G3; G4 showed lower precision (82.1%), possibly due to spectral overlap with G2 and its smaller sample size. VIP analysis identified the silicate skeleton region, the organic matter/water region, and the hydroxyl region as key bands, suggesting that gradients in clay mineral, organic matter, and bound water contents contribute most to the differentiation. This study provides technical support for soil survey and rocky desertification monitoring in karst regions.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported in part by the core research tasks of the "Construction and Application of Soil Spectral Information System in Shilin County, Kunming" project under Grant KMYC202403, as specified in the "Notice on Work Arrangements for the 2024 Science and Technology Plan Project of Yunnan Tobacco Company Kunming Branch Office" (Kunyan Office Comprehensive [2024] No. 30); the Open Project Fund of Yunnan Provincial Key Laboratory of Crop Production and Smart Agriculture under Grant 2024ZHNY03; and the Jiangsu University Student Innovation Training Program under Grant S202510299117.

Conflicts of Interest

Xing Yang, Bixia Jiang, Yanjie He, Rui Yue, and Ruiqian Bao are affiliated with the Kunming Branch, Yunnan Tobacco Company, Kunming 650000, China, a commercial enterprise, and this study was partly funded by a science and technology project of Yunnan Tobacco Company Kunming Branch (Project No. KMYC202403). The authors declare no other conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Jiang, Z., Lian, Y., & Qin, X. (2014). Rocky desertification in Southwest China: Impacts, causes, and restoration. *Earth-Science Reviews, 132*, 1-12. [CrossRef]

[2] Wu, Q., Zheng, W., Rao, C., Wang, E., & Yan, W. (2022). Soil quality assessment and management in karst rocky desertification ecosystem of Southwest China. *Forests, 13*(9), 1513. [CrossRef]

[3] Zhang, J. Y., Dai, M. H., Wang, L. C., Zeng, C. F., & Su, W. C. (2016). The challenge and future of rocky desertification control in karst areas in southwest China. *Solid Earth, 7*(1), 83-91. [CrossRef]

[4] Beckford, H. O., Chu, H., Song, C., Chang, C., & Ji, H. (2021). Geochemical characteristics and behaviour of elements during weathering and pedogenesis over karst area in Yunnan–Guizhou Plateau, southwestern China. *Environmental Earth Sciences, 80*(2), 61. [CrossRef]

[5] Margenot, A. J., Parikh, S. J., & Calderón, F. J. (2023). Fourier-transform infrared spectroscopy for soil organic matter analysis. *Soil Science Society of America Journal, 87*(6), 1503-1528. [CrossRef]

[6] Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied spectroscopy reviews, 49*(2), 139-186. [CrossRef]

[7] Tinti, A., Tugnoli, V., Bonora, S., & Francioso, O. (2015). Recent applications of vibrational mid-infrared (IR) spectroscopy for studying soil components: A review. *Journal of Central European Agriculture, 16*(1), 1-22. [CrossRef]

[8] Solomon, D., Lehmann, J., Kinyangi, J., Liang, B., & Schäfer, T. (2005). Carbon K-edge NEXAFS and FTIR-ATR spectroscopic investigation of organic carbon speciation in soils. *Soil Science Society of America Journal, 69*(1), 107-119. [CrossRef]

[9] Madejová, J. (2003). FTIR techniques in clay mineral studies. *Vibrational spectroscopy, 31*(1), 1-10. [CrossRef]

[10] Nguyen, T. T., Janik, L. J., & Raupach, M. (1991). Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in soil studies. *Australian Journal of Soil Research, 29*(1), 49-67. [CrossRef]

[11] Li, S., Shen, X., Shen, X., Cheng, J., Xu, D., Makar, R. S., ... & Shi, Z. (2025). Improving the accuracy of soil classification by using Vis–NIR, MIR, and their spectra fusion. *Remote Sensing, 17*(9), 1524. [CrossRef]

[12] Lee, L. C., Liong, C. Y., & Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *Analyst, 143*(15), 3526-3539. [CrossRef]

[13] Xie, H., Zhao, J., Wang, Q., Sui, Y., Wang, J., Yang, X., ... & Liang, C. (2015). Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis. *Scientific reports, 5*(1), 10930. [CrossRef]

[14] Maynard, J. J., & Johnson, M. G. (2018). Applying fingerprint FTIR spectroscopy and chemometrics to assess soil ecosystem disturbance and recovery. *Journal of Soil and Water Conservation, 73*(4), 443-451. [CrossRef]

[15] Madejova, J., & Komadel, P. (2001). Baseline studies of the clay minerals society source clays: infrared methods. *Clays and clay minerals, 49*(5), 410-432. [CrossRef]

[16] Wang, X., Huang, X., Xiong, K., Hu, J., Zhang, Z., & Zhang, J. (2021). Mechanism and evolution of soil

organic carbon coupling with rocky desertification in south China karst. *Forests, 13*(1), 28. [CrossRef]

[17] Rossel, R. V., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., ... & Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews, 155*, 198-230. [CrossRef]

[18] Sangwan, P., Nimi, C., Nain, T., Singh, R., & Sharma, N. (2024). Discrimination of soil samples collected from Haryana (India) using non-destructive ATR-FTIR spectroscopy coupled with multivariate statistical analysis. *Indian Journal of Science and Technology, 17*(11), 1087-1096. [CrossRef]

[19] Seybold, C. A., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Nester, B., ... & Thomas, P. (2019). Application of mid-infrared spectroscopy in soil survey. *Soil Science Society of America Journal, 83*(6), 1746-1759. [CrossRef]

[20] Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., ... & Wetterlind, J. (2015). Soil spectroscopy: an alternative to wet chemistry for soil monitoring. *Advances in agronomy, 132*, 139-159. [CrossRef]

[21] Reeves, J. B., McCarty, G. W., & Reeves, V. B. (2001). Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils. *Journal of agricultural and food chemistry, 49*(2), 766-772. [CrossRef]

[22] Linker, R., Shmulevich, I., Kenny, A., & Shaviv, A. (2005). Soil identification and chemometrics for direct determination of nitrate in soils using FTIR-ATR mid-infrared spectroscopy. *Chemosphere, 61*(5), 652-658. [CrossRef]

[23] Margenot, A. J., Parikh, S. J., & Calderón, F. J. (2019). Improving infrared spectroscopy characterization of soil organic matter with spectral subtractions. *JoVE (Journal of Visualized Experiments)*, (143), e57464. [CrossRef]

[24] Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry, 36*(8), 1627-1639. [CrossRef]

[25] Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy, 43*(5), 772-777. [CrossRef]

[26] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions. *Series A, Mathematical, physical, and engineering sciences, 374*(2065), 20150202. [CrossRef]

[27] Wang, J., Liu, T., Zhang, J., Yuan, H., & Acquah, G. E. (2022). Spectral variable selection for estimation of soil organic carbon content using mid-infrared spectroscopy. *European Journal of Soil Science, 73*(4), e13267. [CrossRef]

[28] Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems, 58*(2), 109-130. [CrossRef]

[29] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management, 45*(4), 427-437. [CrossRef]

[30] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering, 21*(9), 1263-1284. [CrossRef]

[31] Artz, R. R., Chapman, S. J., Robertson, A. J., Potts, J. M., Laggoun-Défarge, F., Gogo, S., ... & Francez, A. J. (2008). FTIR spectroscopy can be used as a screening tool for organic matter quality in regenerating cutover peatlands. *Soil Biology and Biochemistry, 40*(2), 515-527. [CrossRef]

[32] Haberhauer, G., Rafferty, B., Strebl, F., & Gerzabek, M. H. (1998). Comparison of the composition of forest soil litter derived from three different sites at various decompositional stages using FTIR spectroscopy. *Geoderma, 83*(3-4), 331-342. [CrossRef]

[33] Demyan, M. S., Rasche, F., Schulz, E., Breulmann, M., Müller, T., & Cadisch, G. (2012). Use of specific peaks obtained by diffuse reflectance Fourier transform mid-infrared spectroscopy to study the composition of organic matter in a Haplic Chernozem. *European Journal of Soil Science, 63*(2), 189-199. [CrossRef]

[34] Stuart, B. H. (2004). *Infrared Spectroscopy: Fundamentals and Applications*. Chichester, UK: John Wiley & Sons. [CrossRef]

[35] Lu, Y., Wang, Y., Chen, Y., Song, N., Wang, Q., Liu, M., & Guan, X. (2026). Nonlinear Responses of Vegetation and Soil Properties to Rock Desertification Gradients in Qingzhen, China. *Land, 15*(3), 499. [CrossRef]

[36] Huang, Y., Zhou, Z., Tang, Q., Huang, D., Li, B., & Luo, Y. (2026). Karst Rocky Desertification in Southwest China: Remote Sensing Progress, Critical Challenges, and Future Pathways. *Applied Sciences, 16*(13), 6459. [CrossRef]

[37] Dai, G., Sun, H., Wang, B., Huang, C., Wang, W., Yao, Y., ... & Zhang, Z. (2021). Assessment of karst rocky desertification from the local to regional scale based on unmanned aerial vehicle images: A case-study of Shilin County, Yunnan Province, China. *Land Degradation & Development, 32*(18), 5253-5266. [CrossRef]

[38] Wu, M., Huang, Y., Zhao, X., Jin, J., & Ruan, Y. (2024). Effects of different spectral processing methods on soil organic matter prediction based on VNIR-SWIR spectroscopy in karst areas, Southwest China. *Journal of Soils and Sediments, 24*(2), 914-927. [CrossRef]

[39] Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets.

Geoderma, 195, 268-279. [CrossRef]

[40] Miao, Y., Wang, H., Huang, X., Liu, K., Sun, Q., Meng, L., & Xu, D. (2024). Soil organic carbon prediction based on Vis–NIR spectral classification data using GWPCA–FCM algorithm. *Sensors, 24*(15), 4930. [CrossRef]

[41] Ge, Y., Thomasson, J. A., & Morgan, C. L. (2014). Mid-infrared attenuated total reflectance spectroscopy for soil carbon and particle size determination. *Geoderma, 213*, 57-63. [CrossRef]

[42] Cécillon, L., Barthès, B. G., Gomez, C., Ertlen, D., Génot, V., Hedde, M., ... & Brun, J. J. (2009). Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). *European Journal of Soil Science, 60*(5), 770-784. [CrossRef]

[43] Grinand, C., Barthès, B. G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., ... & Bernoux, M. (2012). Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). *European Journal of Soil Science, 63*(2), 141-151. [CrossRef]

Xing Yang received the Bachelor's degree and is an engineer. His research areas include crop cultivation, crop physiology, smart agriculture, and intelligent tobacco production. (Email: 50082649@qq.com)

Bixia Jiang received the Bachelor's degree and is an engineer. His research areas include crop cultivation, crop physiology, smart agriculture, and intelligent tobacco production. (Email: 2274320416@qq.com)

Yanjie He received the Bachelor's degree and is an engineer. His research areas include crop cultivation, crop physiology, smart agriculture, and intelligent tobacco production. (Email: 1456039769@qq.com)

Rui Yue received the Bachelor's degree and is an engineer. His research areas include crop cultivation, crop physiology, smart agriculture, and intelligent tobacco production. (Email: 2360461558@qq.com)

Ruiqian Bao received the Bachelor's degree and is an engineer. His research areas include crop cultivation, crop physiology, smart agriculture, and intelligent tobacco production. (Email: 2273279916@qq.com)

Gaoqiang Lv received the Doctor's degree and is a lecturer. His research areas include (1) plant physiology and molecular ecology: physiological changes and gene expression and regulation of crops in response to different nutritional and disease states in complex environments; and (2) agricultural information perception: perception and diagnosis of agricultural information such as crop nutrition and pathological status under complex environmental influences based on various spectral technologies and big data of agricultural resources and environment. (Email: gqlv@ujs.edu.cn)

Fei Ma received the Doctor's degree and is an Associate Researcher. Her research areas include soil fertility and soil proximity sensing/spectroscopy technology. (Email: fma@issas.ac.cn)

Changwen Du received the Doctor's degree and is a researcher. His research areas include soil fertility, soil spectroscopy, and research and application of new fertilizers. (Email: chwdu@issas.ac.cn)