



OPEN

## Geochemical inversion study of potassium and phosphorus in soil based on neural network and ZY1-02D hyperspectral data

Ziyang Li<sup>1,6</sup>, Junxu Chen<sup>2,3,6</sup>, Zhifang Zhao<sup>2,3,4</sup>✉, Xiaotong Su<sup>3,4,5</sup>, Shuanglan Yang<sup>1</sup>, Xinle Zhang<sup>1</sup>, Gaoqiang Xiao<sup>1</sup>, Tao Fu<sup>1</sup> & Lei Niu<sup>1</sup>

In response to the agricultural demand for improving the quality and efficiency of the unique agricultural product “Zhefang Gongmi” in Yingjiang County, Yunnan Province, this study aims to uncover the relationship between soil potassium (K) and phosphorus (P) content and hyperspectral data, and to develop a precise inversion model based on hyperspectral remote sensing. The study innovatively uses AHSI hyperspectral data (166 bands, 400–2500 nm) from the ZY1-02D satellite, combined with Y1-02D satellite, combined with geochemical data from 856 soil sampling points. Through Savitzky-Golay filtering, Minimum Noise Fraction (MNF) transformation, continuum removal, and third-order differential transformation to enhance spectral features, inversion models for K/P elements using Extreme Learning Machine (ELM) are constructed separately for vegetation-covered and bare soil areas. The key findings of the study are as follows: (1) The correlation of potassium content was significantly higher in the vegetated area compared to the bare area, reaching up to 0.55. After continuum removal, significant correlations were observed in the vegetated area at 979 nm, 1031 nm, 1929 nm, and 2334 nm, all with correlation coefficients above 0.50. In contrast, the bare area showed significant correlations in the third-order differential spectrum at 1014 nm, 1677 nm, 1880 nm, and 2216 nm, with a maximum correlation of 0.47. Phosphorus showed a higher correlation in the bare area than in the vegetated area. (2) The optimal prediction models for potassium and phosphorus in both the vegetated and bare areas were based on the ELM model. In the vegetated area, the coefficient of determination for potassium was 0.654, with a mean square error of 22.686 g/kg; in the bare area, the model for potassium yielded a coefficient of determination of 0.617 and a mean square error of 9.102 g/kg. (3) A novel method has been proposed for analyzing the geochemical element content of soil, designed to accurately assess potassium geochemical information and provide a basis for predicting phosphorus content. The “Vegetation - Bare Land” zonal inversion paradigm proposed in this study achieves high-precision inversion of soil potassium (K) content in the highland agricultural areas. The inversion paradigm proposed in this study achieves high-precision inversion of soil K content in the highland agricultural areas, providing an expandable technological pathway for improving the quality of Yingjiang rice and enhancing soil fertility. This approach offers a theoretical foundation for precision agricultural fertilization management.

**Keywords** ZY1-02D hyperspectral data, BP neural network, ELM

Superior crop quality and yield are intricately linked to the geochemical background of ecological and geological environments. In agricultural practice, high-quality crops often degrade across different regions even with identical crop varieties, cultivation techniques, climate, vegetation, topography, and companion crops<sup>1–5</sup>. Rare earth element (REE) distribution in plants is influenced by soil composition and biogeochemical properties<sup>6–11</sup>.

<sup>1</sup>Yunnan Institute of Geological Sciences, Kunming 650011, China. <sup>2</sup>School of Earth Sciences, Yunnan University, Kunming 650500, China. <sup>3</sup>Research Center of Domestic High-resolution Satellite Remote Sensing Geological Engineering, Universities in Yunnan Province, Kunming 650500, China. <sup>4</sup>Yunnan International Joint Laboratory of China-Laos-Bangladesh-Myanmar Natural Resources Remote Sensing Monitoring, Kunming 650500, China. <sup>5</sup>Institute of International Rivers and Eco-Security, Yunnan University, 650500 Kunming, China. <sup>6</sup>Ziyang Li and Junxu Chen contributed equally to this work. ✉email: zhaozhifang@ynu.edu.cn

The study of soil geochemistry aids in identifying the most suitable crops for various soil genetic types and recommends soil enhancement measures through the supplementation of specific trace elements and rock fertilizers to improve crop yields, thereby fundamentally enhancing the quality of arable land<sup>12</sup>.

Phosphorus promotes seed development and starch accumulation in tubers and roots<sup>13</sup>. It is critical for root development; phosphorus deficiency causes dull green leaves, stunted growth, and impairs reproductive organ formation and fruit development<sup>14</sup>. Potassium is essential for carbohydrate synthesis and translocation, indirectly influencing photosynthesis. It also enhances nitrogen metabolism, regulates plant protoplasm colloidal states, and improves resistance to cold, drought, lodging, and pests<sup>15</sup>. As soil potassium and phosphorus levels directly affect crop quality and yield, monitoring these elements is vital for understanding soil chemistry, tracking their geochemical behavior, and supporting agricultural cultivation and planning<sup>16-21</sup>.

Geochemical soil survey is a systematic method to quantify elemental content and other geochemical properties in soil<sup>22-24</sup>. Studies have explored how geomorphology, landscape features, climate, soil genesis, and elemental migration mechanisms influence the efficacy of this method<sup>25</sup>. Residual soil measurement is a mature and effective approach in chemical exploration. The effectiveness of soil measurement in the transport layer depends on the conditions of the measurement area. In wind-formed sand areas, soil measurement has advanced through sampling size interception tests; in organic soil areas, it has been enhanced by partial extraction techniques. Compared with traditional geochemical methods, remote sensing inversion for soil geochemistry offers advantages such as wide coverage, abundant data, advanced technology, rapid data acquisition, frequent updates, and dynamic monitoring.

The ZY1-02D satellite, launched in September 2019, is equipped with an embedded multispectral camera and hyperspectral imager (AHSI). It can acquire 9 band multispectral data over a 115 km swath and 166 bands hyperspectral data over a 60 km swath, covering the spectral range from the visible (400 nm) to the short - wave infrared (2,500 nm). The full chromatic band has a resolution of up to 2.5 m, and the multispectral bands have a resolution of 10 m. The AHSI can capture the spectral features of elements across 166 narrow bands, which is crucial for the inversion and analysis of soil chemical elements.

Shilan Felegari et al. utilized multi-temporal imagery, employing Support Vector Regression (SVR), Partial Least Squares Regression (PLSR), and Artificial Neural Networks (ANN) to estimate the concentration of Cd. The results indicated that the SVR model, using the original imagery as input, provided the most accurate estimation of Cd concentration in the region, ranging from 8 to 26 mg/kg. This method, however, relies on multispectral data with fewer bands, making it difficult to effectively extract the characteristic bands of Cd<sup>26</sup>. Mohammad Esmaeili et al. proposed the ResMorCNN model, which integrates 3D CNN with morphological feature residual injection to extract spatial-spectral features from hyperspectral images. On datasets such as Indian Pines, it achieved an Overall Accuracy (OA) of 99.71%, significantly outperforming traditional CNN and attention models, validating the enhancement provided by morphological features for classification. However, this approach relies on Principal Component Analysis (PCA) for dimensionality reduction (using the first three principal components), which may lead to the loss of crucial spectral information, thus affecting the differentiation of complex land cover<sup>27</sup>. Saeideh Marzvan et al. analyzed the expansion trend of *Azolla filiculoides* in the Anzali Lagoon, Iran, using Landsat time series (1988–2018) and Spectral Angle Mapper (SAM). The use of multispectral data, however, hindered fine-scale (SAM). The use of multispectral data, however, hindered fine-scale resolution<sup>28</sup>. Gomez Cécile explored remote sensing monitoring of soil geochemical elements, focusing on the continuum removal method and partial least squares regression (PLSR) for predicting soil clay and calcium carbonate content using visible and near-infrared (VNIR, 400–1200 nm) and short-wave infrared (SWIR, 1200–2500 nm). Results showed PLSR's advantage with Hymap hyperspectral data, especially when soil spectral features are weak<sup>29</sup>. Hummel John W applied multiple regression techniques, including MLSR, ANN, random forest, and BP neural networks, to invert the content of soil elements like As, Fe, Hg, Cu, Mo, Zn, and Pb in mine soils, confirming the effectiveness of remote sensing spectroscopy in detecting these elements<sup>29-31</sup>. Additionally, research on soil rare earth element inversion in Anxin County achieved an  $R^2$  of 0.982 using partial least squares, random forest, and BP neural networks<sup>32</sup>. Yuehan Qin, using GF-5 data, measured Au content in the Chahuazhai gold mining area, demonstrating that the geographically weighted regression (GWR) model, combined with S-G filtering, provided better fitting for Au content, though requiring a uniform sample distribution<sup>33</sup>. Mehrdad Daviran developed hybrid models combining Particle Swarm Optimization (PSO) with Support Vector Machine (SVM) and Random Forest (RF) to predict copper mineralization in southeastern Iran, showing superior performance over traditional methods, despite high computational costs and kernel selection issues<sup>34</sup>.

Seyed Mahdi Mirhoseini Nejad et al. proposed the ConvLSTM-ViT model, which integrates 3D-CNN, ConvLSTM, and Vision Transformer (ViT) to predict soybean yield using multispectral remote sensing data. Experiments showed that the model outperformed traditional methods in terms of Root Mean Square Error (RMSE) and correlation coefficients. This method, however, requires long-term, high-quality multispectral data and suffers from an abundance of model parameters, which complicates parameter setting<sup>35</sup>. Arezou Akhtaranesh et al. improved the UNet model by incorporating attention blocks in the decoder and addressing class imbalance in the DeepGlobe dataset through data augmentation (slicing, rotation). The model achieved an accuracy of 98.33% in road extraction. However, based on the attention mechanism, large data volumes demand significant computational power, making it resource-intensive<sup>36</sup>. Hadi Mahdipour et al. utilized multiple satellite images and adopted an “ultrafusion” method for land cover segmentation. Compared to the latest similar methods, the overall accuracy, Kappa coefficient, and F1 score improved by approximately 0.86%, 0.52%, and 1.03%, respectively. The method faces challenges related to multi-source data matching and high computational requirements<sup>37</sup>. Mohammad Mahdi Safari et al. combined the U-Net architecture with backbone networks such as VGG and DenseNet, using Sea Surface Temperature (SST) data to identify mesoscale vortices in the Atlantic. Using sparse classification cross-entropy loss, the model achieved an accuracy of 99.37% on multiple datasets, demonstrating the potential of deep learning in ocean dynamics research. However, the method is limited

by small image resolution<sup>37</sup>. Nizom Farmonov et al. proposed the HypsLiDNet framework, which integrates 3D/2D CNN and morphological attention mechanisms, combining hyperspectral (HSI) and LiDAR data for crop classification. Morphological operators are used to extract geometric features, and the attention mechanism optimizes feature fusion. Experiments on the DESIS and Houston datasets achieved high classification accuracy (OA of 98.67%), outperforming traditional machine learning and deep learning methods. The method requires synchronized acquisition of LiDAR and HSI data, which presents significant challenges. The model is also complex and processing is slow<sup>38</sup>. Alireza Vafaeinejad et al. utilized the Segmentation Anything Model (SAM) for high-precision segmentation, achieving a significant Intersection over Union (IoU) score of 92%, which notably surpasses traditional methods. However, this method struggles with the recognition of small-area features and lacks sufficient accuracy in complex scenes. Despite GPU acceleration, large-area computations still require considerable processing power<sup>39</sup>. Alireza Sharifi et al. developed a deep learning model based on Transformer architecture to enhance the spatial resolution of Sentinel-2 images. The model outperforms advanced methods such as ResNet, Swin Transformer, and ViT. Despite its lightweight design, the model's long training times restrict its application to large-scale areas<sup>40</sup>.

Against the backdrop of advancements in remote sensing technology and machine learning for environmental monitoring, this study focuses on using hyperspectral data for soil geochemical analysis. While previous research has explored remote sensing applications across diverse environmental domains, significant gaps remain in high-precision soil geochemical inversion—particularly for potassium (K) and phosphorus (P) content. Traditional methods such as support vector regression (SVR) and partial least squares (PLS) have achieved notable results in solving complex problems like pattern recognition, regression prediction, and time-series analysis. However, these approaches often require extensive feature selection, manual parameter tuning, or complex kernel functions, which can limit model efficiency and scalability. Additionally, although deep learning techniques are powerful, they often incur high computational costs in certain scenarios and are prone to overfitting. In contrast, the Extreme Learning Machine (ELM) has emerged as a promising alternative due to its ability to capture nonlinear data features through random feature mapping. This unique characteristic of ELM not only reduces the complexity of feature engineering but also enhances computational efficiency, offering faster training speeds and simpler parameter optimization.

Therefore, this study aims to achieve high-precision inversion of soil geochemical potassium content by exploring the relationship between AHSI hyperspectral data from the ZY1-02D satellite and soil K/P geochemical contents. A novel processing approach is proposed, integrating Savitzky-Golay filtering, Minimum Noise Fraction (MNF) transformation, spectral differentiation, and continuum removal to enhance spectral features. A “vegetation-bare soil” zonal inversion paradigm is established, using the ELM algorithm to invert soil K/P geochemical contents. Model reliability is evaluated using four parameters—R<sup>2</sup>, MSE, RMSE, and training time—to explore the potential of hyperspectral data in high-precision soil geochemical inversion and propose a scalable framework for regional geochemical mapping. The research findings not only provide a new method for soil nutrient status analysis but also offer critical support for precision agriculture practices and cultivated land quality improvement in plateau regions.

## Materials and methods

### Study area

The study area (97°31' and 98°16' E longitude and 24°24' and 25°20' N latitude) is located in Yingjiang County, Dehong Dai Jingpo Autonomous Prefecture, Yunnan Province, with a total study area of 100 km<sup>2</sup> and it is 3.8 km from Yingjiang County. The region experiences a southern subtropical monsoon climate, characterized by an average annual temperature of 19.3 °C and an average annual precipitation of 1464 m.

The regional stratigraphy of the Yingjiang area is mainly composed of the Paleozoic Gaoligongshan Group and the Holocene and Pleistocene. The rock formations consist of black cloud schist, dolomite schist, black cloud diorite meta granite, and black cloud dio-rite gneiss. The study area is predominantly granite, with significant mica content in the soils, particularly in the rice-growing regions. The soils are chiefly composed of brick red loam, red loam, yellow loam, yellow-brown loam, and brown loam. The majority of soils in the county are phosphorus-deficient, acidic, and exhibit an imbalanced nutrient ratio. The soil types and their altitudinal distribution are as follows: brick red loam at 210–600 m, red loam at 600–1400 m and 1400–2000 m, yellow loam at 2000–2300 m, yellow-brown loam at 2300–2800 m, and brown loam at 2800–3400 m. Notably, the brown loam is found at altitudes around 2800 m (Fig. 1).

In 2021, the grain cultivation area in the study area is 117.63 million hectares, in which the rice cultivation area is 29.92 million hectares<sup>41,42</sup>.

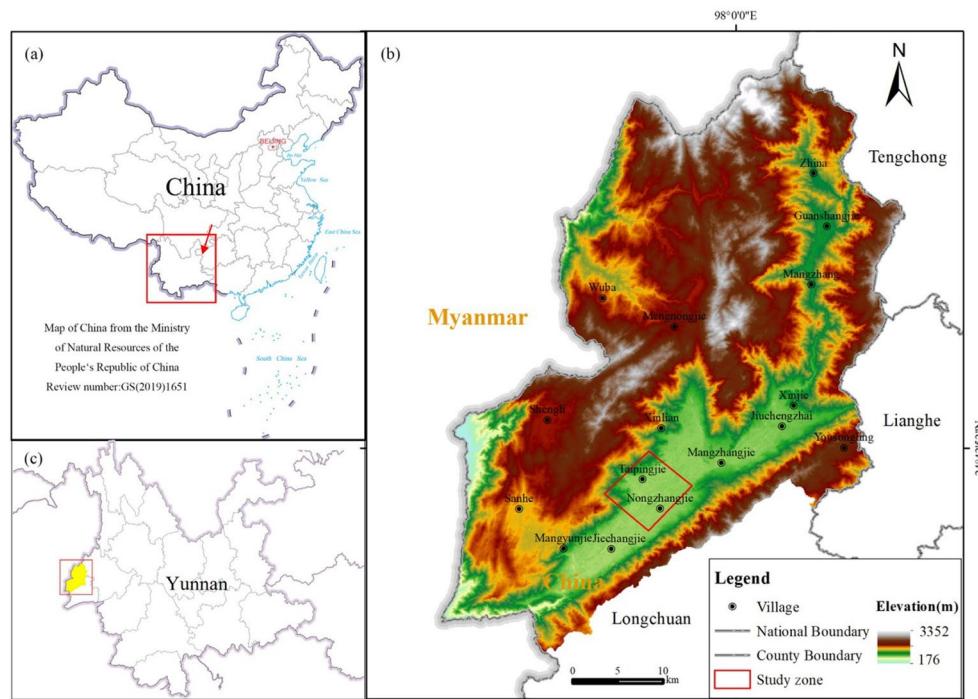
### Data

#### ZY1-02D hyperspectral data

The study primarily utilized ZY1-02D hyperspectral data from December 14, 2020, purchased from the Yunnan Remote Sensing Center, with the data ID “ZY1E\_AHSI\_E97.85\_N24.61\_20201214\_006588\_L1A0000208280.” The surface reflectance of the L1A data was obtained through radiometric calibration and atmospheric correction (Table 1). Subsequently, noise removal, spectral smoothing, and data enhancement were performed on the hyperspectral data using Savitzky-Golay filtering and spectral transformation techniques, including differential processing and continuum removal.

#### Geochemical data

A total of 856 soil geochemical samples were collected in the study area. Sampling density ranged from 4 to 16 points per square kilometer, with most samples collected from cultivated land. Other land use types were sampled at 4 points per square kilometer to ensure comprehensive coverage and avoid gaps. Soil samples from



The DEM data set is provided by Geospatial Data Cloud site, Computer Network Information Center, Chinese Academy of Sciences. (<http://www.gscloud.cn>) and the map was created by Ziyang Li.

**Fig. 1.** (a) Map of China; (b) Map of Yunnan province; (c) Location of the study area.

Spectral range	400–2500 nm
Number of bands	166
Spatial resolution	30 m
Spectral resolution	VNIR 10 nm
	SWIR 20 nm
Swath width	60 km
Side swing angle	26°
Revisit cycle	3 days
Orbital period	55 days

**Table 1.** The parameters of the advanced hyperspectral imager (AHSI) data of the ZY1-02D satellite.

arable and forest lands were collected at a depth of 0–20 cm. Using each GPS positioning point as the center, 4–6 subsamples were taken within a 50–100 m radius to form a mixed sample of equal volume.

Collected samples were fully air-dried and sieved through a 10-mesh screen. A 200-g subsample was weighed and sent for analysis, while at least 300-g subsamples were stored in clean plastic bottles and transferred to a sample repository. Analytical procedures strictly followed the *Specification for Multi-target Regional Geochemical Survey* (DZ/T 0295–2016) and *Technical Requirements for Analysis of Ecological Geochemical Evaluation Samples (Trial)* (DD2005-03). Potassium content was measured by inductively coupled plasma atomic emission spectrometry (ICP-AES), and phosphorus content was determined using the alkali fusion-molybdenum antimony anti-spectrophotometric method.

Analytical quality was ensured through a combination of external quality control (EQC) and internal quality monitoring (IQM). EQC involved inserting password-coded standard control samples; accuracy and precision of external standards ranged from 97 to 100%, correlation coefficients ( $r$ ) were 0.966–0.999, and F-test values for two-sample ANOVA were 1.02–1.09 (all < one-tailed F critical values). IQM tracked quality parameters including accuracy/precision of national standard substance analyses, reporting rates, repeatability tests, and anomaly checks. For IQM: accuracy/precision pass rates for all indices were 100%; reporting rates ranged 98.2–100%; repeatability test pass rates were 95.8–100%; and anomaly repeatability test pass rates were 95.3–100% (Table 2).

The average potassium content is 28.42 g/kg, falling within the medium range for agricultural land (20–30 g/kg), but with significant spatial heterogeneity (standard deviation 6.82, variance 46.50). The range is as wide as 45.05 g/kg (3.49–48.54), indicating a phenomenon of both local enrichment and depletion. The left-skewed

	Statistics of raw data	
	K	P
Mean	28.42	0.78
Std. deviation	6.82	0.26
Variance	46.50	0.07
Skewness	-0.62	1.37
Kurtosis	1.37	3.17
Minimum	3.49	0.19
Maximum	48.54	2.04

**Table 2.** The statistical treatments of Raw values of 2 analyzed elements.

distribution (skewness -0.62) and low kurtosis (1.37) suggest that the data is concentrated in the lower value range and has a relatively flat distribution, which may be related to the differences in mountain parent materials and the mobile nature of potassium. Agricultural management should focus on regions with potassium depletion (where low-value areas dominate) and hotspots of enrichment (where the maximum value exceeds the mean by 1.7 times).

The average phosphorus content is 0.78 g/kg, which is below the critical value for effective phosphorus in farmland (1.0 g/kg). Overall, it shows the characteristic of “general deficiency, local excess.” The standard deviation of 0.26 (33% of the mean) and the strong right-skewed distribution (skewness 1.37) indicate that phosphorus is deficient in most areas, but there are a few high-value outliers (the maximum value is 2.04 g/kg, 2.6 times the mean). The peaked distribution (kurtosis 3.17) shows that the data is highly concentrated in the lower value range. Combined with a range of 1.85 g/kg (0.19–2.04), this reflects an imbalance in the spatial application of phosphorus fertilizers.

Using ZY1-02D hyperspectral remote sensing data, vegetation index normalization categorized sampling points in the study area into 235 within vegetated areas and 621 in bare areas for chemical analysis (Fig. 2).

### Research process

The research flowchart is shown in Fig. 3. This study primarily includes three parts: (1) Hyperspectral data preprocessing: Noise is addressed via S-G filtering and MNF methods, followed by differentiation and continuum removal transformations to eliminate variations, thereby enhancing absorption and reflection features in spectral curves; (2) Selection of characteristic spectral bands with significant correlations to corresponding soil geochemical elements; (3) Development of predictive models for the two elements using Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR), BP, and ELM regression analysis, with the optimal model selected for soil element content inversion in the Yingjiang region.

#### Data preprocessing

First, the raw ZY1-02D hyperspectral data are pre-processed, which includes radiometric calibration, atmospheric correction, and orthorectification. Specifically, radiometric calibration converts the digital number (DN) of the image to radiometric brightness, reflectance, or surface temperature. Atmospheric correction compensates for atmospheric effects using the MODTRAN radiative transfer model. Additionally, orthorectification corrects the image for tilt and projection distortions.

After removing atmospheric and geometric distortion effects, the differences between the spectral curves after atmospheric correction and the USGS standard spectral library are compared in the selected broad-leaved woodland area (Fig. 4).

#### Savitzky-Golay filter

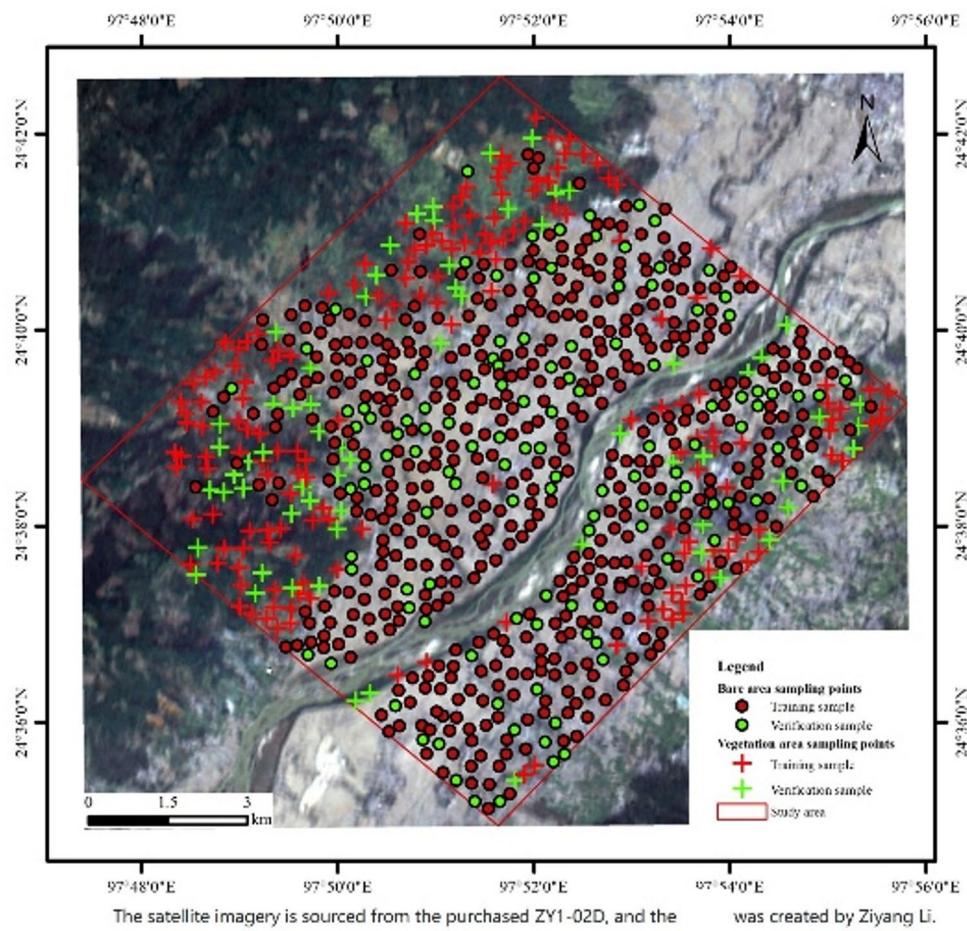
Due to the influence of imaging instruments and environmental factors, the hyperspectral data contain varying degrees of noise in both the imaging and pre-processing stages. This noise affects both the spatial and spectral domains, reducing the advantages of hyperspectral resolution. To address this issue, the pre-processed ZY1-02D hyperspectral data were filtered using the Savitzky-Golay (S-G) filtering method<sup>43</sup>.

The SG filtering method is based on local polynomials in the time domain using least squares fitting. It efficiently reduces noise compared to traditional methods by directly moving a “window” in the time domain. This approach preserves spectral characteristics such as extrema and width while effectively eliminating noise.

$$R'_i = \frac{1}{H} \sum_{j=-k}^k R_{i+j} * h_j \quad (1)$$

where  $R'_i$  is the fitted value,  $R_{i+j}$  is the original value of the pixel,  $\frac{h_j}{H}$  is the smoothing coefficient, which is obtained by fitting a polynomial by the least squares method,  $k$  is the number of unilateral bands to be fitted.

After SG filtering, the spectral curve becomes smoother compared to the atmospherically corrected result. The absorption and reflection intervals of the spectral curve are clearly visible (Fig. 5).



**Fig. 2.** The distribution of the geochemical soil samples.

#### Minimum noise fraction rotation(MNF)

Minimum Noise Fraction (MNF) is commonly used in hyperspectral data processing, particularly in the field of remote sensing. Its fundamental concept involves applying a linear transformation to hyperspectral data in order to separate the noise components from the valid signal components, thereby enhancing the effectiveness of subsequent processing. MNF mainly involves two principal component transformations. The first transformation uses the noise covariance matrix to reorder and separate the noise in the data (Fig. 6); the second transformation is mainly the standard principal component change of the noise-whitened data, and it alleviates the impact of noise on image quality compared to principal component analysis (PCA)<sup>44</sup>.

$$P = U * D_N^{-\frac{1}{2}} \quad (2)$$

where  $D_N$  is the diagonal matrix of  $C_N$  eigenvalues in descending order,  $I$  is the unit matrix, and  $P$  is the transform matrix. When  $P$  is applied to the image data  $X$ , the original image is projected into the new space by applying the  $Y = PX$  transformation, and the noise in the resulting transformed data has unit variance and is not correlated between bands.

$$C_{D-adj} = P^T * C_D * P \quad (3)$$

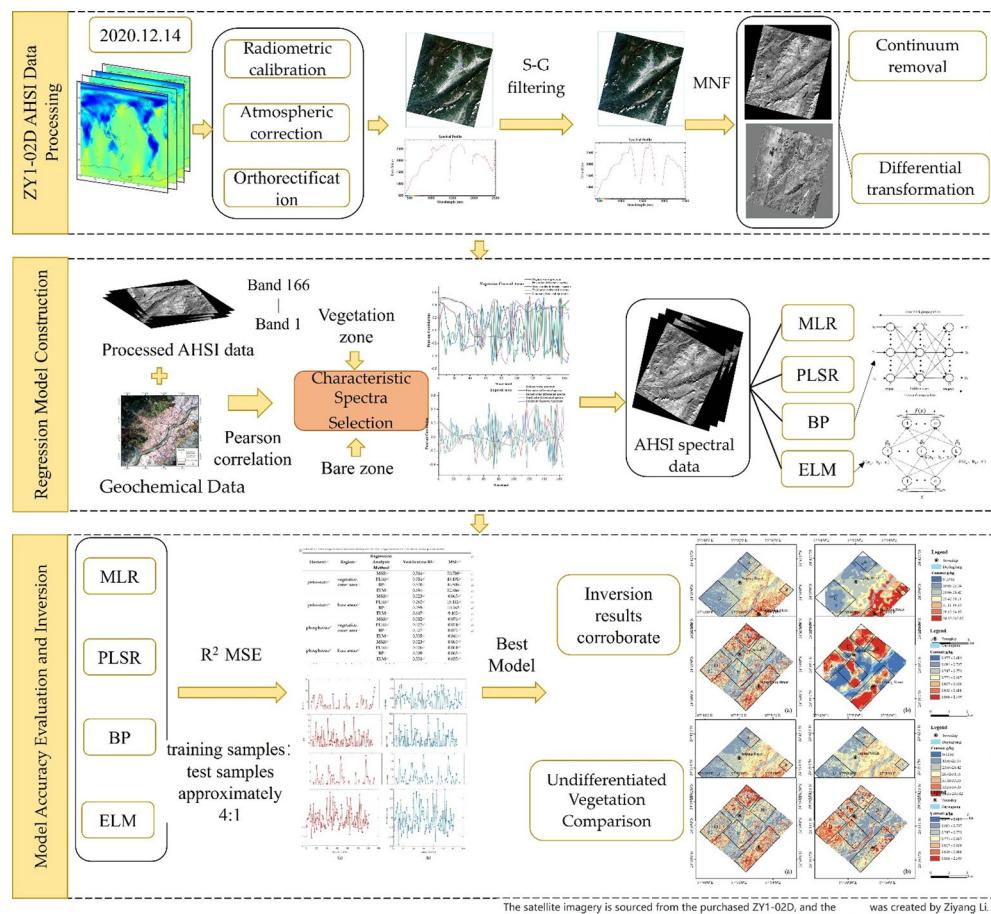
where  $C_D$  is the covariance matrix of image  $X$ ;  $C_{D-adj}$  is the matrix after P-transformation, and it is further diagonalized into a matrix  $D_{D-adj}$

$$D_{D-adj} = V^T * C_{D-adj} * V \quad (4)$$

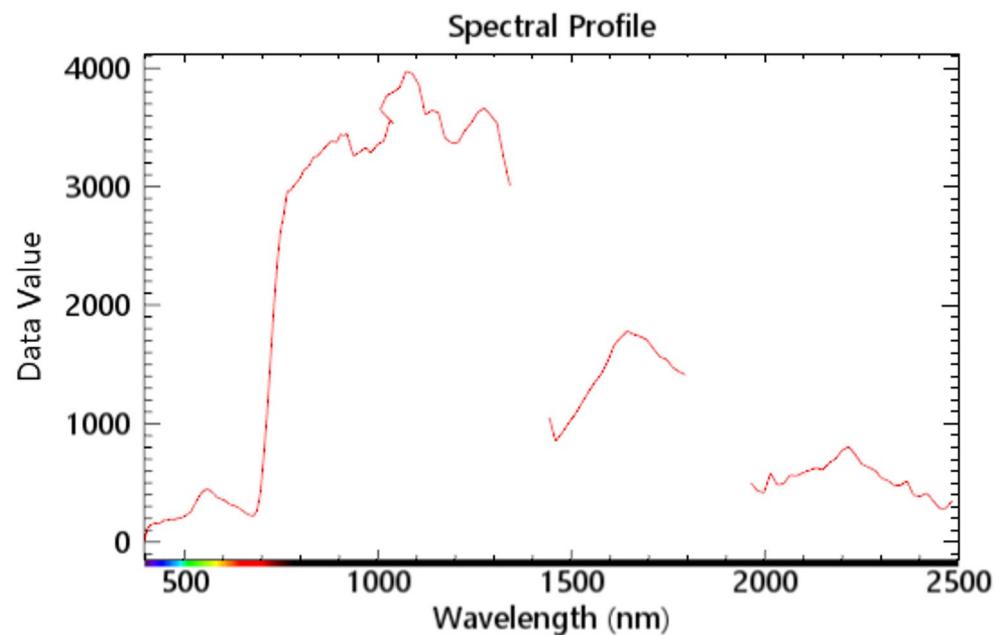
where,  $D_{D-adj}$  is the diagonal matrix of eigenvalues of  $C_{D-adj}$  in descending order;  $V$  is the orthogonal matrix composed of eigenvectors.

#### Continuum removal

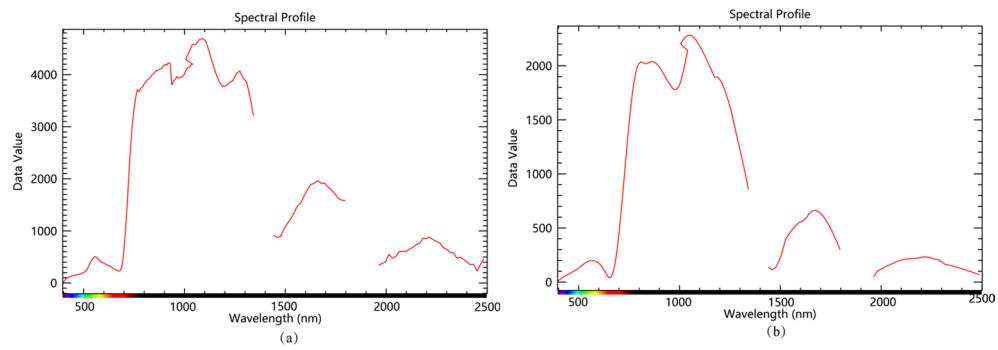
The continuum removal method is a commonly used spectral analysis method to compare feature values with other spectral curves, which can highlight the absorption and reflection features of the spectral curve by



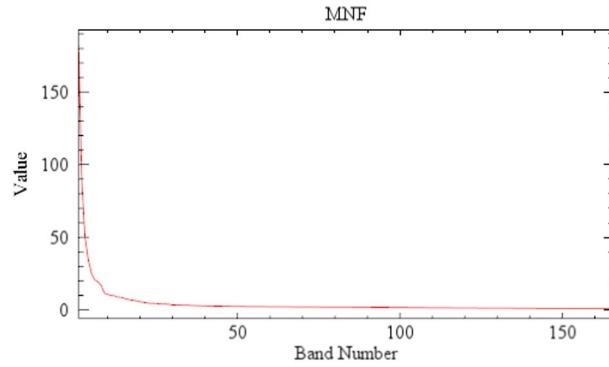
**Fig. 3.** The flow chart of the research process.



**Fig. 4.** The ZY1-02D spectral profile after radiometric calibration and FLAASH.



**Fig. 5.** The spectral profile after S-G filtering.



**Fig. 6.** The MNF analysis curve.

normalizing the absorption features of the spectrum to a consistent spectral background. The band depth is subtracted from the envelope spectrum by 1<sup>46</sup>, indicating the depth of the formed absorption valley due to the absorption spectral features of the soil chemical components at a certain wavelength point with lower reflectance than the neighboring bands. The calculation formula is calculated as follows:

$$R'_H(\lambda_i) = 1 - \frac{R(\lambda_i)}{R_C(\lambda_i)} \quad (5)$$

where  $\lambda_i$  is the spectral reflectance of the  $i$ -th band,  $R(\lambda_i)$  is the spectral reflectance of the corresponding wavelength,  $R_C(\lambda_i)$  is the band reflectance value on the envelope of the band,  $\frac{R(\lambda_i)}{R_C(\lambda_i)}$  is the envelope line, and  $R'_H(\lambda_i)$  is the depth of the band.

#### *First-order differential*

The spectral differential transform is a calculation of the spectral curve to increase the trend of the spectral curve and expand the spectral features while eliminating the noise effect of the spectral data, and the difference of the spectrum is generally used as a finite approximation of the differential in the actual calculation<sup>46</sup>.

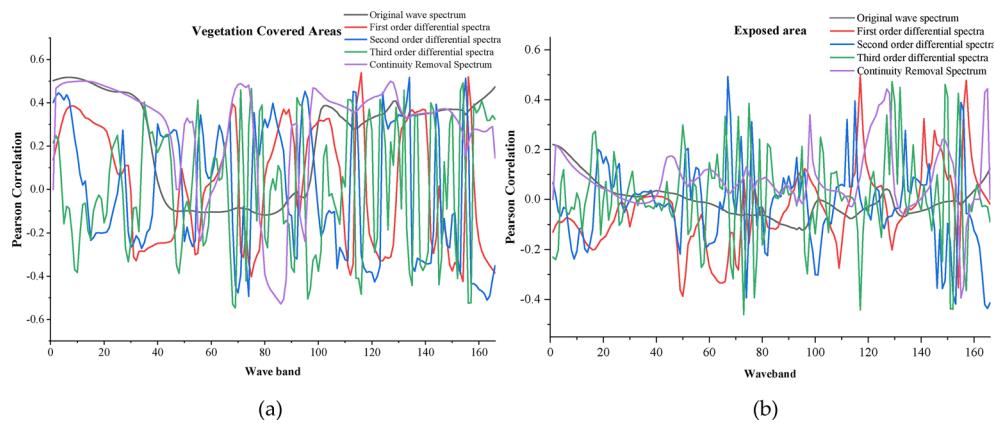
$$R'_H(\lambda_i) = \frac{[R(\lambda_{i+1}) - R(\lambda_{i-1})]}{(\lambda_{i+1} - \lambda_{i-1})} \quad (6)$$

where  $\lambda_{i+1}$ ,  $\lambda_i$ , and  $\lambda_{i-1}$  are adjacent wavelengths;  $R(\lambda_{i+1})$ ,  $R(\lambda_i)$ , and  $R(\lambda_{i-1})$  are the spectral reflectance of the corresponding wavelengths;  $R'_H(\lambda_i)$  is the first-order differential of the wavelength  $\lambda_i$ .

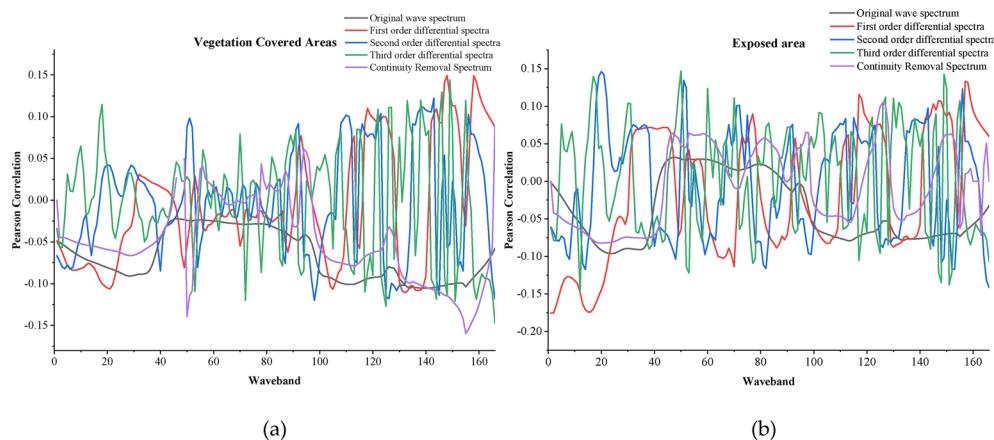
#### *Second-order differential*

$$R''(\lambda_i) = \frac{[R'(\lambda_{i+1}) - R'(\lambda_{i-1})]}{(\lambda_{i+1} - \lambda_{i-1})} \quad (7)$$

where  $\lambda_i$  is the adjacent wavelength,  $R'(\lambda_i)$  is the derivative of the corresponding wavelength, and  $R''(\lambda_i)$  is the second-order differentiation of the wavelength  $\lambda_i$ .



**Fig. 7.** (a) Correlation curves of soil geochemical potassium sample analysis data from vegetation cover areas to the removal of hyperspectral image bands from primary, differential, and continuum systems; (b) Correlation curves of soil geochemical potassium sample analysis data in bare areas to the removal of hyperspectral image bands by primary, differential, and continuum systems.



**Fig. 8.** (a) Correlation curves of soil geochemical phosphorus sampling and analysis data from vegetation cover areas to the removal of hyperspectral image bands from primary, differential, and continuum systems; (b) Correlation curves of soil geochemical phosphorus sampling and analysis data in bare areas to the removal of hyperspectral image bands by primary, differential, and continuum.

### Characteristic spectra selection

To investigate the relationship between geochemical element contents and the reflectance in soil, the original and the transformed spectra are correlated with soil geochemical elements, respectively. The correlation coefficients are calculated as follows.

$$R = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{S_X} \right) \left( \frac{Y_i - \bar{Y}}{S_Y} \right) \quad (8)$$

where  $X$  and  $Y$  are the spectral reflectance and the measured elemental content at the  $i$ -th wavelength, respectively;  $\bar{X}$  and  $\bar{Y}$  are the sample mean of the spectral reflectance and the measured elemental content, respectively;  $S_X$  and  $S_Y$  are the sample variance of the spectral reflectance and the measured elemental content, respectively;  $n$  is the number of samples;  $R$  is the correlation coefficient.

The images after S-G filtering are filtered by Pearson correlation analysis with the potassium and phosphorus elements of the measured chemical sampling sites, and the bands with a higher correlation of the respective elements are selected (Figs. 7 and 8).

Pearson correlation analysis revealed that the correlation of potassium content in vegetated areas is stronger compared to bare areas, with a significant correlation at the 0.01 significance level. The correlation coefficient curve for differential transformation in vegetated areas exhibits significant fluctuations, whereas the spectral curve of continuous unified removal transformation performs optimally in the ranges of 400–500 nm, 1100–1200 nm, and around 1650 nm. Following the third-order differential transformation, the correlation coefficients

for potassium in bare areas are pronounced, with stronger correlations observed in the ranges of 1000–1100 nm, 1650–1700 nm, 1850–1900 nm, and 2200–2300 nm, generally exceeding 0.4. The correlation between phosphorus elements and the spectra is not evident, with most bands showing no significant correlation, and the correlation coefficients remaining within 0.20. After first-order differential transformation, significant correlations were observed at the 0.05 significance level in the ranges of 1650–1700 nm, 2150–2250 nm, and 2350–2400 nm (Table 3).

Based on data collection and previous experience, the characteristic spectral bands in regions of spurious correlation were eliminated. Ultimately, the characteristic spectral bands necessary for model construction were selected within the near-infrared (NIR) band.

Following the third-order differential transformation of the spectra, the potassium (K) content in the vegetation zone exhibited significant correlations at 979 nm, 1031 nm, 1929 nm, and 2334 nm. The potassium in the bare soil area showed significant correlations at 1014 nm, 1677 nm, 1880 nm, and 2216 nm. The third-order differential transformed spectral data were selected as the characteristic bands for potassium in the vegetation zone. Phosphorus in vegetated areas was significantly correlated at 1812 nm, 2031 nm, 2216 nm, and 2501 nm. The phosphorus in the bare soil area was significantly correlated at 816 nm, 1005 nm, 2216 nm, and 2501 nm. The third-order differential transformed spectral data were selected as the characteristic bands for phosphorus in the vegetation area.

### Regression models

In this study, hyperspectral spectral data are modeled with measured soil geochemical data using multiple linear regression (MLR), partial least squares regression (PLSR), the BP neural network, and the ELM regression. Subsequently, the ZY1-02D hyperspectral spectral data obtained from the aforementioned correlation analysis are modeled with measured soil geochemical data for potassium (K) in the vegetation cover area, potassium in the bare area, phosphorus (P) in the bare area, and phosphorus in the bare area. Response models are established between the ZY1-02D hyperspectral spectral data and the soil geochemical measurements for potassium in the bare area and phosphorus in the vegetation cover area.

In the construction of the mathematical models, 70% of the total sample data are used for model construction and the rest for model validation. In the construction of the two neural network models, 170 and 65 pieces of sample data are involved in model training and model validation for the vegetation cover area, respectively. 500, and 121 pieces of sample data are involved in model training and validation for the bare area, respectively.

#### *Multiple linear regression (MLR)*

Multiple linear regression analysis predicts the dependent variable by establishing a regression equation between multiple independent variables and the dependent variable, providing greater accuracy than traditional univariate regression analysis models<sup>22</sup>

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_j X_j + \dots + \beta_n X_n + \epsilon \quad (9)$$

Spectral changes	K(Vegetation zone)		K(Bare area)		P(Vegetation zone)		P(Bare area)	
	Wave length(nm)	Coefficient of correlation	Wave length(nm)	Coefficient of correlation	Wave length(nm)	Coefficient of correlation	Wave length(nm)	Coefficient of correlation
Raw Spectrum	1627	0.408	1274	-0.120	1593	-0.101	842	0.029
	1880	0.408	1307	-0.122	2014	-0.105	894	0.030
	2467	0.447	2467	0.086	2132	-0.104	1610	-0.079
	2501	0.474	2501	0.125	2317	-0.104	1963	-0.079
First order spectrum	1031	-0.402	816	-0.387	1475	-0.107	954	-0.103
	1627	0.539	1677	0.497	1912	-0.111	988	-0.114
	2334	0.520	2300	-0.353	2199	0.149	1677	0.116
	2501.	-0.385	2350	0.477	2366	0.149	2350	0.133
Second order spectra	1023	-0.493	962	0.492	1357	-0.120	825	0.135
	1963	0.517	1023	-0.393	1828	-0.118	1089	-0.116
	2317	0.513	1644	0.395	2031	0.111	2334	0.123
	2450	-0.511	2484	-0.436	2115	0.122	2501	-0.142
Third order spectra	979	-0.547	1014	-0.460	1812	-0.127	816	0.147
	1031	0.466	1677	-0.442	2031	0.120	1005	-0.109
	1929	0.458	1880	0.473	2216	0.144	2216	0.144
	2334	-0.525	2216	0.461	2501	-0.147	2501	-0.107
Continuity Removal Spectrum	996	0.489	1845	0.443	816	-0.139	1627	-0.055
	1139	-0.511	2317	-0.394	1290	0.061	1828	0.105
	1156	-0.529	2467	0.429	2317	-0.160	2317	-0.063
	1845	0.500	2484	0.443	2501	0.090	2450	-0.072

**Table 3.** The ZY1-02D hyperspectral remote sensing data and element correlation.

where  $Y$  denotes the characteristic to be analyzed, which is the content of soil geochemical elemental abundance in this study;  $X_j$  denotes the  $j$ th independent variable, which is the extracted spectral feature in this study;  $\beta_j$  denotes the regression coefficient corresponding to the  $j$ th independent variable;  $\epsilon$  is the random error of the regression equation, and  $n$  is the number of independent variables used in modeling.

#### Partial least squares regression (PLSR)

Partial least squares regression (PLSR) is a novel multivariate statistical data analysis method that addresses the issue of multicollinearity among variables by performing a covariance test to identify the presence of covariance, determining the number of components, and establishing a partial least squares regression model. The model is straightforward to calculate, offers high predictive accuracy, and is easy to interpret qualitatively<sup>29</sup>.

$$X = TP^T + E \quad (10)$$

$$Y = UQ^T + F \quad (11)$$

where  $X$  is a matrix of predictors (element content), and  $Y$  is a matrix of responses (ZY1-02D spectra value);  $T$  and  $U$  are projections of  $X$  and  $Y$ , respectively;  $P$  and  $Q$  are orthogonal loading matrices; matrices  $E$  and  $F$  are error terms.  $X$  and  $Y$  are decomposed to maximize the covariance between  $T$  and  $U$ .

#### Back propagation (BP) neural network

The BP neural network is a multilayer feedforward network trained using the error backpropagation algorithm, capable of learning and storing numerous input-output pattern mapping relationships<sup>32</sup>. It consists of an input layer, a hidden layer, and an output layer (Fig. 9).

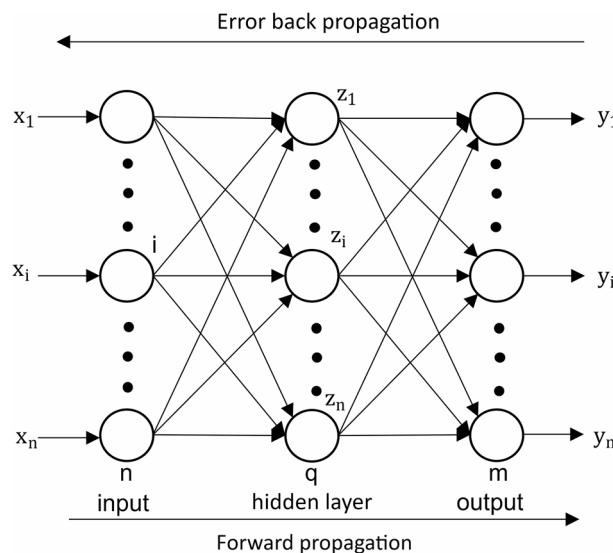
The core idea of the BP algorithm is to utilize gradient descent for searching the hypothesis space of possible weight vectors to find the optimal weight vector that best fits the samples. Specifically, using a loss function, the algorithm iteratively adjusts weights and biases in the direction of the negative gradient until the loss function reaches a minimum value. Additionally, the backpropagation algorithm computes gradients, the partial derivatives of the loss function with respect to weights and biases in each layer, updating initial weights and biases iteratively until either the loss function minimizes or a predefined number of iterations is completed. This approach is crucial for optimizing parameters within neural networks.

The parameter setting of the neural network is mainly the number of layers and neurons in each layer. After several experiments, the initial parameter setting of the BP neural network for samples in the vegetation area is: the network is set to contain three layers with 30, 10, and 30 neurons respectively, the maximum number of epochs of the training network is 1000, the learning rate is 0.01, the maximum number of verification checks is set to 10, and the TrainParam goal is set to 0.0000000004.

The initial parameter setting of the BP neural network for samples in the bare area is: the network is set to contain three layers with 20, 20, and 20 neurons respectively, the maximum epoch of the training network is 1000, the learning rate is 0.1, the maximum number of validation checks is set to 5, and the TrainParam goal is set to 0.0000000004.

#### Excess learning machine (ELM)

This study attempts to apply it to the construction of inversion models for soil geochemical content. ELM is a class of machine learning systems or methods constructed on the Feedforward-Neuron-Network (FNN) for supervised and unsupervised learning problems. ELM is regarded as a special class of FNN or an improvement



**Fig. 9.** The structure of the backpropagation neural network model.

of FNN and its backpropagation algorithm, and it has several characteristics: the weights of the nodes in the hidden layer are randomly or artificially given and do not need to be updated, and the learning process only calculates the output weights. The standard ELM adopts the Single Layer Feedforward neuron Network (SLFN) structure. Specifically, the SLFN is composed of an input layer, an implicit layer, and an output layer (Fig. 10).

ELM has found applications in various fields that require the processing of complex data, pattern recognition, regression prediction, and time series analysis. Whether in finance, healthcare, intelligent transportation, or in areas such as natural language processing and image processing, ELM has demonstrated exceptional performance. Furthermore, as research progresses, its areas of application continue to expand. Traditional mathematical models, such as Support Vector Regression (SVR) and polynomial regression, typically require manual feature selection or the use of complex kernel methods to enhance model capability. In contrast, ELM captures the nonlinear characteristics of data through random feature mapping, reducing the workload of feature engineering. By mapping to a high-dimensional space in a random manner, ELM can effectively address nonlinear problems, enabling it to outperform traditional models such as linear regression or SVM in nonlinear regression tasks. Additionally, compared to deep learning, it offers advantages such as faster training speeds, simpler parameters, and a reduced risk of overfitting.

ELM can randomly initialize the input weights and thresholds and obtain the corresponding hidden node outputs. In terms of the structure of the neural network, the transcendental learning machine is a simple single-layer forward neural network, which contains three layers: the input layer, the hidden layer, and the output layer. The hidden layer has  $L$  neurons,  $L$  is much smaller than  $N$ , and the output layer outputs a vector of  $m$  dimensions.

$$\sum_{i=1}^L \beta_i g(a_i * x_i + b_i) = f_L(x_j), j = 1, 2, 3, \dots, N \quad (12)$$

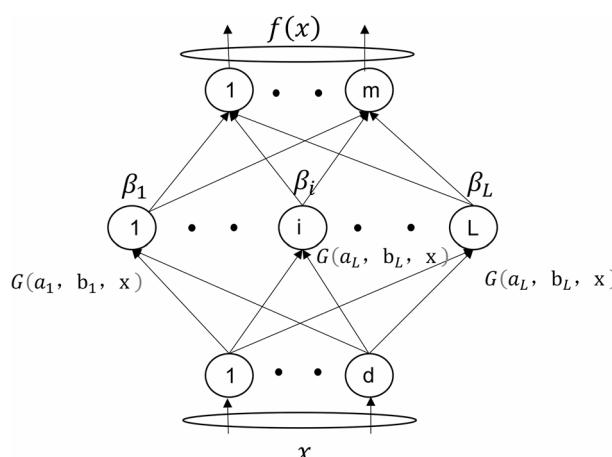
where  $g(x)$  is the activation function;  $a_i = (a_{i1}, a_{i2}, \dots, a_{in})^T$  is the input weight of the  $i$ th hidden unit;  $\beta_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{in})^T$  is the bias of the  $i$ th hidden unit;  $f_L(x_j)$  is the output weight of the  $i$ th hidden unit.

## Results

### Model accuracy evaluation

After distinguishing between vegetation-covered and bare areas, the vegetation cover area and bare areas, the vegetation cover area was modeled using 170 soil geochemical test data points, with model performance evaluated using 65 measured sample data points and the corresponding predicted values. For the bare area, the model was developed based on 500 measured analytical data points and assessed by comparing 121 measured data points with the predicted values. The correlation between the measured and predicted values was evaluated using the coefficient of determination ( $R^2$ ), where a higher  $R^2$  indicates a stronger correlation. The model was constructed by integrating the aforementioned parameters, and the models derived from the regression of two elements using different methods were compared and analyzed (Table 4).

For the K model in the vegetated area, the Partial Least Squares Regression (PLSR) and Multiple Stepwise Regression (MSR) models exhibit comparable accuracies, with an  $R^2$  value of approximately 0.31. The accuracy of the ELM model substantially surpasses those of the MSR, PLSR, and Backpropagation (BP) models. The ELM model achieves the highest accuracy, with an  $R^2$  of 0.654. While preserving the  $R^2$  value, this model significantly diminishes the prediction error, and the model training time is reduced to 0.023 s. For the K-element model in the bare area, when compared to the MSR model ( $R^2 = 0.023$ ), the PLSR model effectively enhances the inversion accuracy, with an  $R^2$  of 0.262. Both the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are also notably reduced. The BP model yields an  $R^2$  of 0.359, demonstrating approximately a 37% performance improvement relative to the PLSR model. It reduces the MSE and RMSE while maintaining model accuracy. The ELM model sustains a high  $R^2$ , and the model training time is 0.0032 s.



**Fig. 10.** The structure of the excess learning machine model.

Element	Region	Regression Analysis Method	Verification $R^2$	MSE	RMSE	Training time
K	vegetation cover area	MSR	0.314	50.788	7.127	-
		PLSR	0.314	49.491	7.035	-
		BP	0.570	46.805	6.841	1.1900s
		ELM	0.654	22.686	4.763	0.0052s
K	bare areas	MSR	0.023	65.134	8.071	-
		PLSR	0.262	18.441	4.294	-
		BP	0.359	10.262	3.203	1.1629s
		ELM	0.617	9.102	3.017	0.0032s
P	vegetation cover area	MSR	0.012	0.076	0.276	-
		PLSR	0.022	0.074	0.272	-
		BP	0.127	0.072	0.268	1.0943s
		ELM	0.335	0.041	0.202	0.0031s
P	bare areas	MSR	0.023	0.065	0.255	-
		PLSR	0.026	0.064	0.253	-
		BP	0.209	0.065	0.255	1.1150s
		ELM	0.331	0.035	0.187	0.0032s

**Table 4.** The regression model analysis of the vegetation cover area with potassium.

For the P models in both the vegetated and bare areas, the MSR and PLSR models have similar  $R^2$ , MSE, and RMSE values, indicating poor accuracy. Compared with traditional mathematical models and the BP model, the ELM model shows a huge improvement in accuracy, reaching approximately 0.33, and the model construction time is shortened to 0.003 s.

The performance stability of ELM in heterogeneous vegetation-bare land environments highlights its unique environmental adaptation mechanism. In vegetated areas, ELM's prediction accuracy for potassium ( $R^2 = 0.654$ ) was notably higher than that in bare land ( $R^2 = 0.617$ ), with both significantly surpassing traditional methods. This indicates that the randomly generated hidden-layer basis functions of ELM can autonomously separate the spectral masking effects of vegetation canopies (e.g., chlorophyll absorption valleys) from soil background spectral features—a dynamic decoupling unattainable by linear methods like PLSR due to their fixed basis functions (exhibiting an  $R^2$  difference of 0.355 in bare land). Furthermore, the MSE for phosphorus was lower in bare land (0.035) than in vegetated areas (0.041), suggesting that ELM is more robust to high-frequency surface noise (such as abrupt mineral reflectance variations). Additionally, during training, unlike the vanishing gradient problem encountered in BP for phosphorus prediction, ELM directly solves for output weights, significantly outperforming BP and enabling a much faster training process.

The table demonstrates that, for both potassium and phosphorus, neural network-based models significantly outperform traditional mathematical models in accuracy. Specifically, the ELM model for potassium in vegetated areas exhibits a 10% improvement over the conventional BP model, with an  $R^2$  of 0.654 and an MSE of 22.686. This suggests that the neural network regression model achieves optimal modeling accuracy. Considering both  $R^2$  and MSE, the ELM model was selected as the inversion model for potassium. The residual plot, histogram, and Q-Q plot of the potassium element model in the vegetation zone are presented below (Fig. 11).

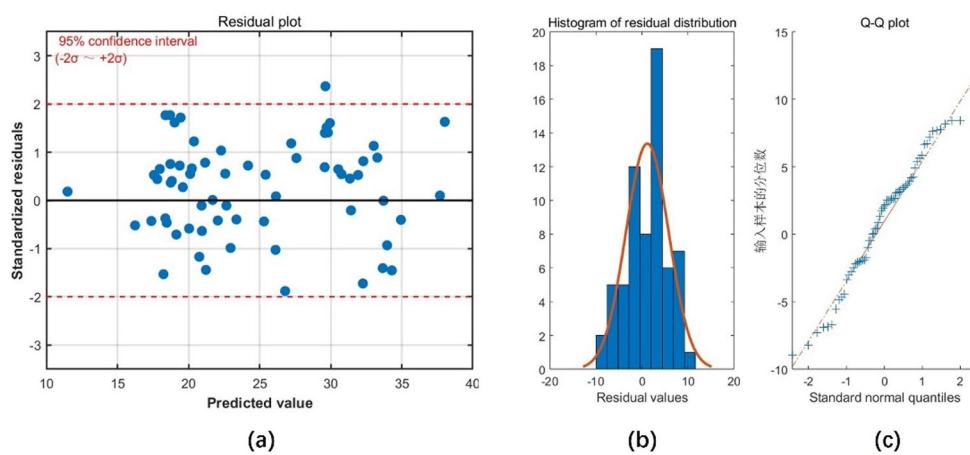
The regional ELM model shows a nearly 71% improvement over the BP model, with an  $R^2$  of 0.617 and an MSE of 9.102. The extreme learning machine regression model was selected as the inversion model for potassium in the bare area. The residual plot, histogram, and Q-Q plot of the potassium element model in the bare soil zone are presented below (Fig. 12).

For phosphorus in vegetated areas, among the four models, the neural network-based inversion model exhibits a substantial accuracy improvement, approximately reaching 400%. The ELM model has an  $R^2$  of 0.3354 and an MSE of 0.0414. Considering both  $R^2$  and MSE, the ELM model was selected as the inversion model for phosphorus in the vegetated area. The residual plot, histogram, and Q-Q plot of the phosphorus element model in the vegetation zone are presented below (Fig. 13).

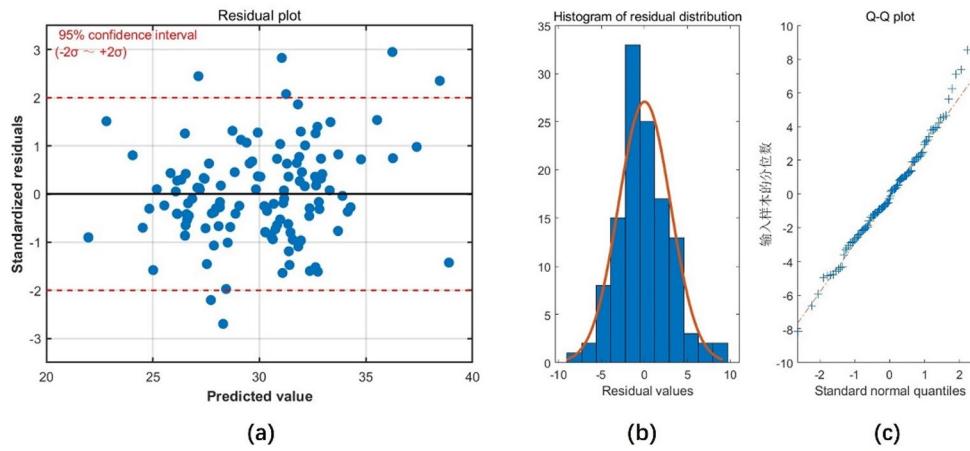
The inversion model accuracy for regional phosphorus has increased by 900%, with the ELM model showing an additional 58% improvement compared to the BP model. The  $R^2$  of the ELM was 0.3314 with an MSE of 0.0351, and the accuracy of the ELM model was higher, so the ELM model was used for the inversion model of the phosphorus in the bare area. The residual plot, histogram, and Q-Q plot of the phosphorus element model for the bare soil zone are shown below (Fig. 14).

### Element inversion

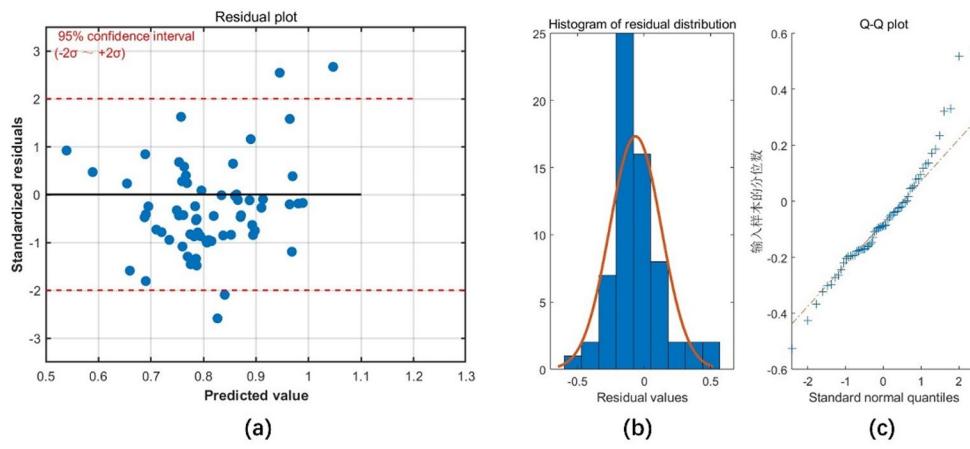
The optimal excess learning machine model obtained by training was inverted based on the obtained excess learning machine model for the potassium in the vegetation cover area, potassium in the bare area, phosphorus in the vegetation cover area, and phosphorus in the bare area (Figs. 15 and 16).



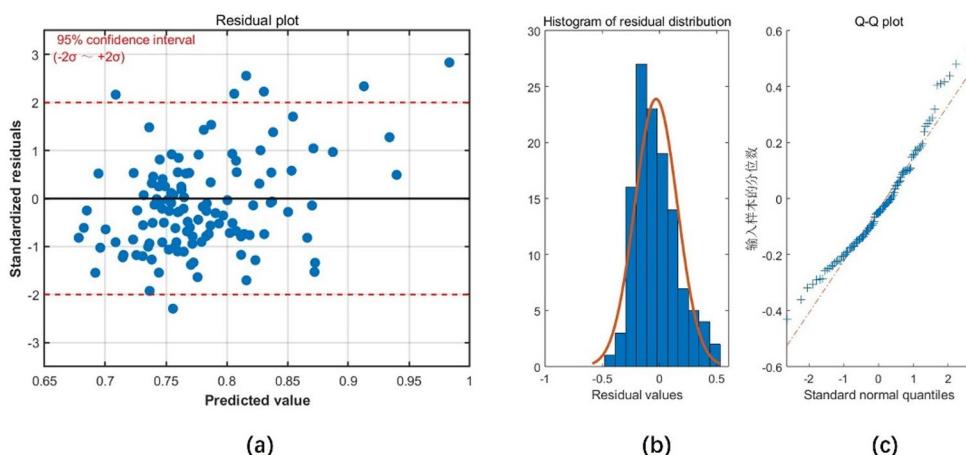
**Fig. 11.** (a) Plot of standardized residuals; (b) Histogram of residual distribution. (c) Q-Q plot.



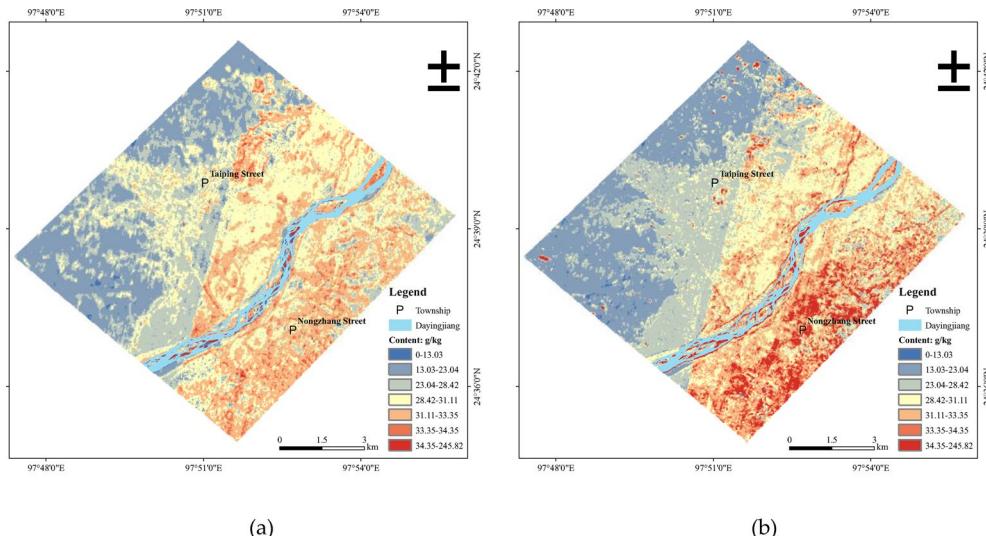
**Fig. 12.** (a) Plot of standardized residuals; (b) Histogram of residual distribution. (c) Q-Q plot.



**Fig. 13.** (a) Plot of standardized residuals; (b) Histogram of residual distribution. (c) Q-Q plot.



**Fig. 14.** (a) Plot of standardized residuals; (b) Histogram of residual distribution. (c) Q-Q plot.



**Fig. 15.** (a) The results of potassium inversion without differentiating between vegetation cover and bare areas. (b) The results of potassium inversion with differentiating between vegetation cover and bare areas.

## Discussion

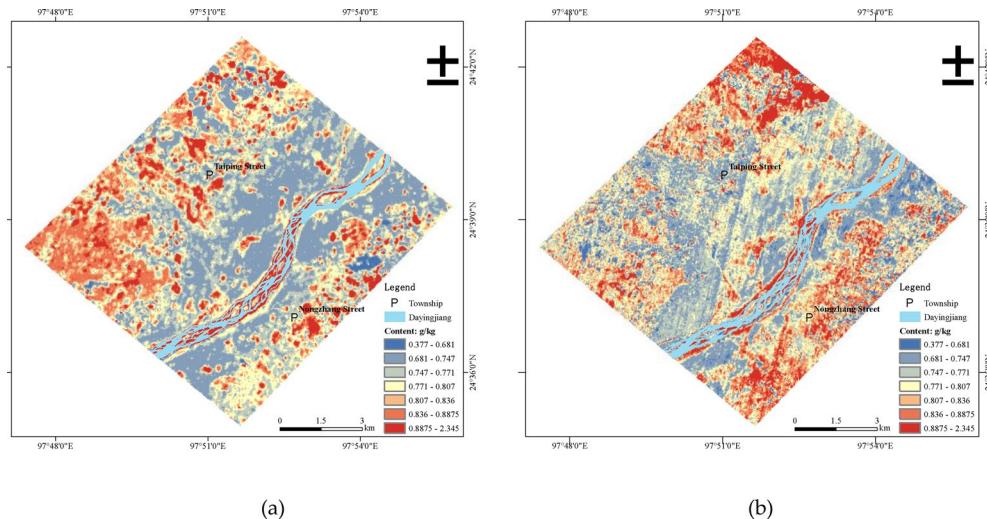
### Actual data comparison

The distribution maps of potassium and phosphorus elements derived from the Extreme Learning Machine regression model showed a strong correspondence with the measured sample data from Taiping Street and Nongzhang Street. The maximum potassium content detected was 5.321 mg/kg in bare areas and 3.891 mg/kg in vegetated areas. Correspondingly, the highest phosphorus content observed was 2.390 mg/kg in bare areas and 1.974 mg/kg in vegetated areas.

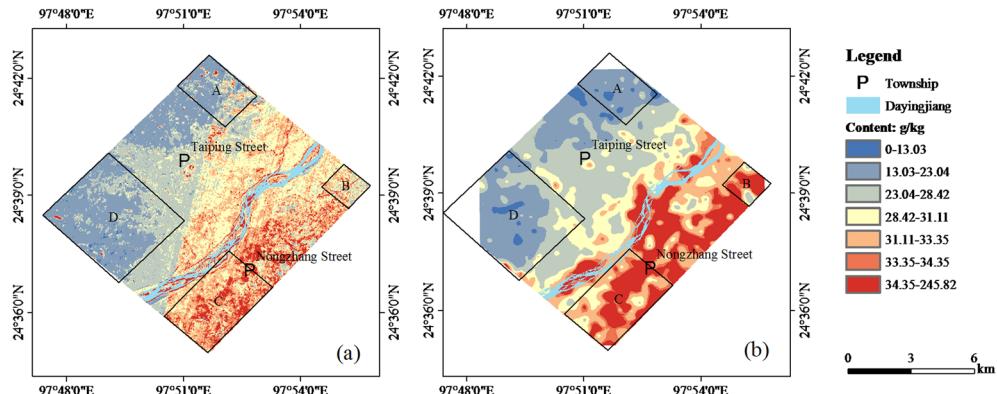
The predicted geochemical potassium anomalies exhibited a variation trend analogous to that of the field geochemical exploration sampling analysis data (Fig. 17). Following the differentiation between vegetated and bare areas, the inversion results revealed a strong correspondence between potassium concentrations in areas A, C, and D of the study area and the measured sample processing data. Notwithstanding, in areas B, despite its low vegetation coverage, the results remained inconsistent with the measured data.

The phosphorus inversion results after vegetation differentiation aligned with the high anomalies in the measured sample data for areas A, C, and E of the study area. Concurrently, phosphorus high-anomaly zones were concentrated in areas B, D, and F, though discrepancies with measured sample data were observed, with only some high-anomaly zones remaining consistent. The inversion results were indistinct in the critical vegetation-covered and bare soil areas of D and E (Fig. 18).

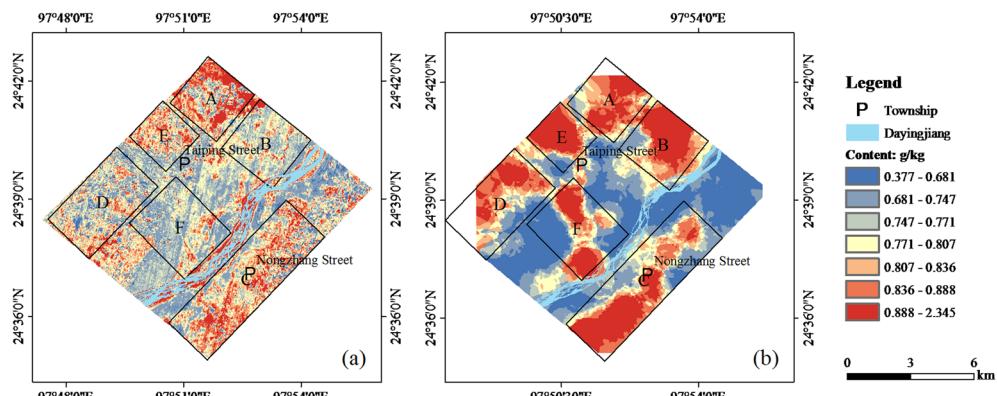
Through field verification and comparison with the potassium element inversion distribution, it was observed that at field points YJ01-R, YJ02-R, YJ03-R, YJ05-R, YJ06-R, YJ07-R, YJ08-R, YJ09-R, YJ11-R, YJ12-R, YJ13-R, YJ14-R, YJ15-R, YJ17-R, YJ18-R, YJ19-R, YJ20-R, and YJ25-R, the majority of soils were rich in mica. Mica



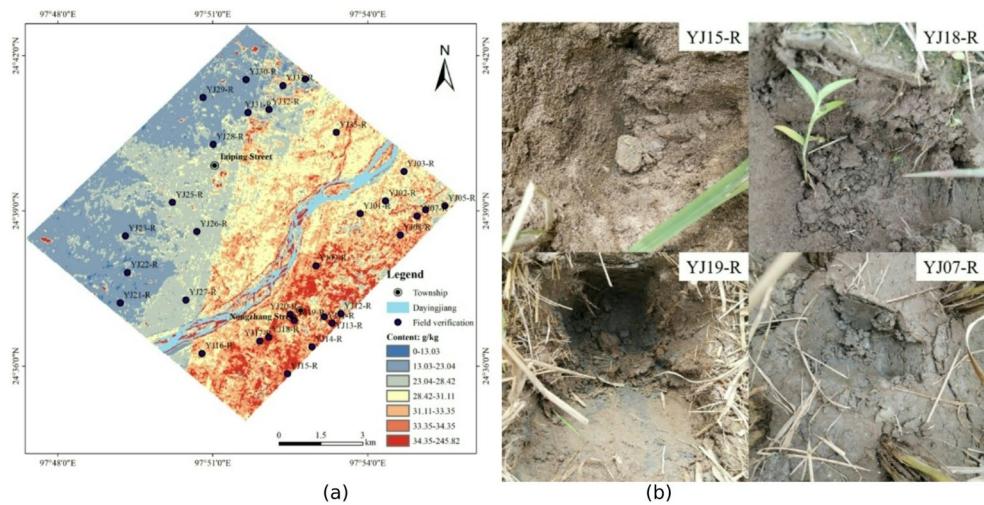
**Fig. 16.** (a) The results of phosphorus inversion without distinguishing vegetation cover and bare areas. (b) The results of potassium inversions in overlaid vegetation cover and bare areas.



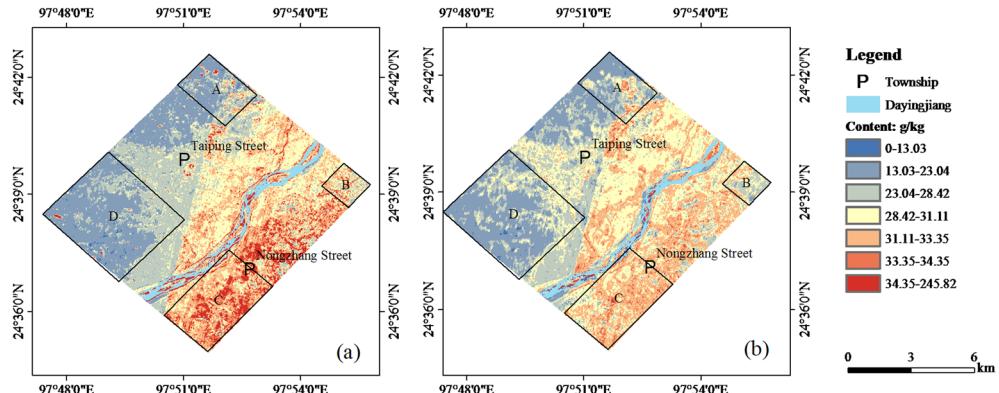
**Fig. 17.** (a) The inversion results of potassium content in differentiated vegetation cover bare areas; (b) The interpolated content data of potassium samples.



**Fig. 18.** (a) The inversion results of phosphorus content in bare areas with differentiated vegetation cover; (b) The interpolated phosphorus content data of the measured samples.



**Fig. 19.** (a) Field-based verification overlaid with the potassium element inversion distribution; (b) On-site verification of soil conditions.



**Fig. 20.** (a) The inversion results of potassium content after distinguishing vegetation (b) The inversion results of potassium content without distinguishing vegetation.

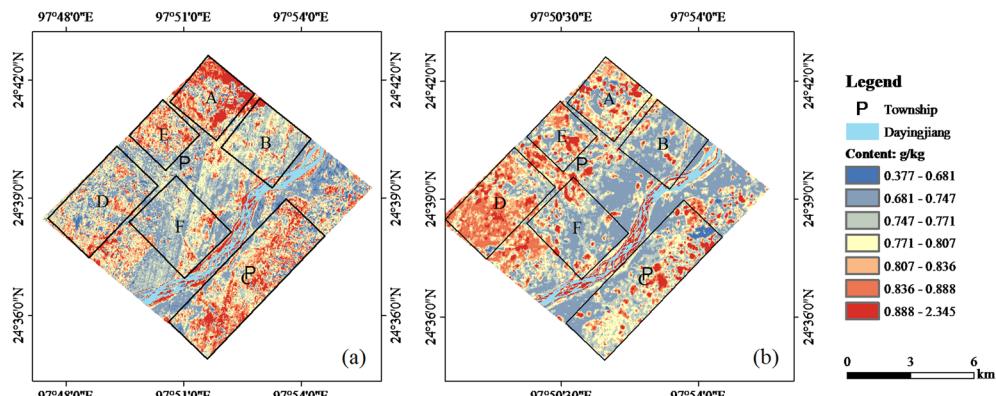
minerals contain abundant potassium, explaining why soils with higher mica content generally exhibit elevated potassium levels. This observation closely aligns with the potassium element inversion distribution (Fig. 19).

#### Undifferentiated vegetation comparison

After differentiating vegetation using normalized vegetation indices, the distribution of soil geochemical potassium elements in the study area was generally consistent with the inversion results from undifferentiated vegetation (see Fig. 20). Vegetation cover was categorized into four areas: A, B, C, and D. Among these, Zones A and D exhibited higher vegetation cover. The results showed scattered high anomalies of soil geochemical potassium elements in vegetated areas, with overall low potassium content in these regions. Both methods indicated that Zone C had the highest potassium content across the entire study area.

However, the ELM model was used to select distinct characteristic spectra for inverting bare areas within vegetated zones. Results indicated that this approach emphasized potassium high-anomaly areas in the southern Dayingjiang River region. Compared with the undifferentiated vegetation geochemical inversion method, the ELM model delineated the spatial distribution of potassium high anomalies across the entire study area while preserving the accuracy of low-potassium-content regions, effectively filtering high-value areas.

The inversions of phosphorus exhibited significant variations in vegetation areas A and D. In area A, the inversions showed a notable increase in the continuous distribution of high-value areas after vegetation differentiation, while the results in area B highlighted some high anomalous areas. The results for bare area C highlighted the high anomalous areas of phosphorus, which were consistent with the measured soil geochemical data, confirming the accuracy of extracting anomalies from the differentiated vegetation (Fig. 21).



**Fig. 21.** (a) The inversion results of phosphorus content after distinguishing vegetation (b) The inversion results of phosphorus content without distinguishing vegetation.

## Conclusions

This study presents a high-precision soil geochemical element inversion method by leveraging large-scale soil geochemical analysis data and ZY1-02D hyperspectral data. Through data transformations including denoising and enhancement, an ELM model is established to derive more accurate soil geochemical anomaly information for potassium. Experimental results show that this model currently exhibits higher operational efficiency and improved accuracy compared to the BP model.

The development of a high-precision soil geochemical survey method can, on one hand, effectively provide a scientific basis for regional agricultural planning, such as rice cultivation. By mapping the precise distribution of soil geochemical data, it offers recommendations for high-quality rice planting areas, thereby supporting the development of plateau-specific agriculture. On the other hand, this method can be extended to environmental pollution assessment by constructing relevant models for soil heavy metal concentrations, which are critical for engineering projects and human daily activities. This approach also serves as an important complement to existing technical methods. Compared with traditional soil geochemical surveys, which are time-consuming and labor-intensive, hyperspectral remote sensing technology enables more accurate prediction of soil geochemical content over larger areas.

The study is subject to certain limitations. The model construction focuses solely on the selected feature bands, without integrating the feature band selection with the model-building algorithm. In the future, methods for selecting feature bands will be incorporated to further streamline the model-building process, enhancing both its speed and accuracy, with the aim of achieving a more precise soil geochemical inversion model. Additionally, the number of sample points in vegetation-covered areas is comparatively smaller, which affects the model's accuracy. In practical applications, high-density vegetation and varying vegetation types in these areas can significantly influence model precision. Therefore, when constructing the model, factors such as the region's vegetation cover should be considered based on specific regional requirements. Moreover, other external factors may also impact the feature bands and model accuracy.

## Data availability

The data presented in this study are available upon request from the corresponding author.

Received: 23 December 2024; Accepted: 11 June 2025

Published online: 21 July 2025

## References

1. Fernando, D. R., Dennis, D., Mikkonen, H. G. & Reichman, S. M. A preliminary assessment of As and F uptake by plants growing on uncontaminated soils. *Water Air Soil Pollut.* **232**, 302 (2021).
2. Vinogradov, D. V. & Zubkova, T. V. Accumulation of Heavy Metals by Soil and Agricultural Plants in the Zone of Technogenic Impact. *IJARe* (2021). <https://doi.org/10.18805/IJARe.A-651>
3. Natasha, N. et al. Accumulation pattern and risk assessment of potentially toxic elements in selected wastewater-irrigated soils and plants in Vehari, Pakistan. *Environ. Res.* **214**, 114033 (2022).
4. Saleh, E. A. & Isinkaralar, Ö. Analysis of trace elements accumulation in some landscape plants as an indicator of pollution in an urban environment: case of Ankara. *Kastamonu Univ. J. Eng. Sci.* **8**, 1–5 (2022).
5. Connor, J. J. & Schaklette, H. T. Background geochemistry of some rocks, soils, plants, and vegetables in the conterminous United States. (1975).
6. Schachtman, D. P., Reid, R. J. & Ayling, S. M. Phosphorus uptake by plants: from soil to cell. *Plant Physiol.* **116**, 447–453 (1998).
7. Eriksson, H. Geochemistry of stream plants and its statistical relations to soil- and bedrock geology, slope directions and till geochemistry. *56*.
8. Miao, L. et al. Geochemistry and biogeochemistry of rare Earth elements in a surface environment (soil and plant) in South China. *Environ. Geol.* **56**, 225–235 (2008).
9. Harada, H. & Hatanaka, T. Natural background levels of trace elements in wild plants: variation and distribution in plant species. *Soil. Sci. Plant. Nutr.* **46**, 117–125 (2000).
10. Kronberg, B. I. & Nesbitt, H. W. Quantification of weathering, soil geochemistry and soil fertility. *J. Soil Sci.* **32**, 453–459 (1981).

11. Kulal, C., Padhi, R. K., Venkatraj, K., Satpathy, K. K. & Mallaya, S. H. Study on trace elements concentration in medicinal plants using EDXRF technique. *Biol. Trace Elem. Res.* **198**, 293–302 (2020).
12. Semenova, V. V. & Anatov, D. M. Accumulation of heavy metals in plants of the genus Achillea L. under arid conditions of the plain zone of Dagestan. *Arid Ecosyst.* **12**, 85–92 (2022).
13. Агбалян, Е. et al. *Environ. Dynamics Global Clim. Change* **12**, 27–42 (2021).
14. Wang, Z., Liu, X. & Qin, H. Bioconcentration and translocation of heavy metals in the soil-plants system in Machangqing copper mine, Yunnan province, China. *J. Geochem. Explor.* **200**, 159–166 (2019).
15. Harada, H. & Hatanaka, T. Natural background levels of trace elements in wild plants. *Soil. Sci. Plant. Nutr.* **44**, 443–452 (1998).
16. Shtangeeva, I. V. Behaviour of chemical elements in plants and soils. *Chem. Ecol.* **11**, 85–95 (1995).
17. Kryuchenko, N. O., Zhovinsky, E. Y. & Paparyga, P. S. Biogeochemical peculiarities of accumulation of chemical elements by plants of the sydovets Massif of the Ukrainian Carpathians. *J. Geol. Geogr. Geocology* **30**, 78–89 (2021).
18. DJINGOVA, R., KULEFF, I. & MARKERT, B. Chemical fingerprinting of plants. *Ecol. Res.* **19**, 3–11 (2004).
19. Zhukovskaya, N. V., Kavalchik, N. V. & Vlasov, B. P. Determination of Mn, Cu and Pb background level in higher aquatic plants within Belarusian reservoirs based on National monitoring system data. *Environ. Qual. Manage.* **32**, 279–288 (2022).
20. Koptsiuk, G. N., Koptsiuk, S. V., Smirnova, I. E. & Sinichkina, M. A. Effect of soil degradation and remediation in technogenic barrens on the uptake of nutrients and heavy metals by plants in the Kola Subarctic. *Eurasian Soil. Sc.* **54**, 1252–1264 (2021).
21. Ito, S., Yokoyama, T. & Asakura, K. Emissions of mercury and other trace elements from coal-fired power plants in Japan. *Sci. Total Environ.* **368**, 397–402 (2006).
22. Guo, H. et al. Mapping Soil Organic Matter Content Based on Feature Band Selection with ZY1-02D Hyperspectral Satellite Data in the Agricultural Region. *Agronomy* **12**, 2111 (2022).
23. Alinia-Ahandani, E., Nazem, H., Malekiran, A. & Fazilati, M. The safety evaluation of toxic elements in medicinal plants: a systematic review. *J. Human Environ. Health Promotion* **8**, 64–70 (2022).
24. Xu, Z. et al. Evaluating the capability of satellite hyperspectral imager, the ZY1-02D, for topsoil nitrogen content estimation and mapping of farmlands in black soil area, China. *Remote Sens.* **14**, 1008 (2022).
25. Calaway, M. J. Ice-cores, sediments and civilisation collapse: a cautionary Tale from lake Titicaca. *Antiquity* **79**, 778–790 (2005).
26. Felegari, S., Sharifi, A., Khosravi, M. & Sabanov, S. Using experimental models and multitemporal Landsat-9 images for cadmium concentration mapping. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–4 (2023).
27. Esmaeli, M., Abbasi-Moghadam, D., Sharifi, A., Tariq, A. & Li, Q. ResMorCNN model: hyperspectral images classification using Residual-Injection morphological features and 3DCNN layers. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **17**, 219–243 (2024).
28. Marzvan, S., Moravej, K., Felegari, S., Sharifi, A. & Askari, M. S. Risk assessment of alien *Azolla filiculoides* lam in Anzali lagoon using remote sensing imagery. *J. Indian Soc. Remote Sens.* **49**, 1801–1809 (2021).
29. Gomez, C., Lagacherie, P. & Coulouma, G. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma* **148**, 141–148 (2008).
30. Kemper, T. & Sommer, S. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. *Environ. Sci. Technol.* **36**, 2742–2747 (2002).
31. Liu, H. & Beaudoin, G. Geochemical signatures in native gold derived from Au-bearing ore deposits. *Ore Geol. Rev.* **132**, 104066 (2021).
32. Zhao, H. et al. Application of a fractional order differential to the hyperspectral inversion of soil Iron oxide. *Agriculture* **12**, 1163 (2022).
33. Qin, Y. et al. Coupling relationship analysis of gold content using GaoFen-5 (GF-5) satellite hyperspectral remote sensing data: A potential method in Chahuazhai gold mining area, Qiubei county, SW China. *Remote Sens.* **14**, 109 (2022).
34. Daviran, M., Maghsoudi, A., Ghezelbash, R. & Optimized, A. I. M. P. M. Application of PSO for tuning the hyperparameters of SVM and RF algorithms. *Comput. Geosci.* **195**, 105785 (2025).
35. Mirhoseini Nejad, S. M., Abbasi-Moghadam, D. & Sharifi, A. ConvLSTM-ViT: A deep neural network for crop yield prediction using Earth observations and remotely sensed data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **17**, 17489–17502 (2024).
36. Akhtarmanesh, A. et al. Road extraction from satellite images using Attention-Assisted UNet. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **17**, 1126–1136 (2024).
37. Mahdipour, H., Sharifi, A., Sookhak, M., Medrano, C. R. & Ultrafusion Optimal fuzzy fusion in Land-Cover segmentation using multiple panchromatic satellite images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **17**, 5721–5733 (2024).
38. Farmonov, N. et al. 3-D-2-D CNN model and Spatial-Spectral morphological attention for crop classification with DESIS and lidar data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **17**, 11969–11996 (2024).
39. Vafaeinejad, A., Alimohammadi, N., Sharifi, A., Safari, M. M. & Super-Resolution AI-Based approach for extracting agricultural cadastral maps: form and content validation. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **18**, 5204–5216 (2025).
40. Sharifi, A. & Safari, M. M. Enhancing the Spatial resolution of Sentinel-2 images through Super-Resolution using Transformer-Based Deep-Learning models. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **18**, 4805–4820 (2025).
41. Chen, J. et al. Alp-valley and elevation effects on the reference evapotranspiration and the dominant climate controls in red river basin, China: insights from geographical differentiation. *J. Hydrol.* **620**, 129397 (2023).
42. Peng, J. et al. The conflicts of agricultural water supply and demand under climate change in a typical arid land watershed of central Asia. *J. Hydrology: Reg. Stud.* **47**, 101384 (2023).
43. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *ACS Publications https://pubs.acs.org/doi/pdf/10.1021* (2002).
44. Nielsen, A. A. Kernel maximum autocorrelation factor and minimum noise fraction transformations. *IEEE Transactions on Image Processing* **20**, 612–624 (2010).
45. Huang, Z., Turner, B. J., Dury, S. J., Wallis, I. R. & Foley, W. J. Estimating foliage nitrogen concentration from HYMAP data using continuum removal analysis. *Remote Sensing of Environment* **93**, 18–29 (2004).
46. Hindmarsh, J. L. & Rose, R. M. A model of the nerve impulse using two first-order differential equations. *Nature* **296**, 162–164 (1982).

## Acknowledgements

The authors would like to thank all the reviewers who participated in the review and MJEditor ([www.mjeditor.com](http://www.mjeditor.com)) for its linguistic assistance during the preparation of this manuscript.

## Author contributions

Conceptualization, Z.Z., J.C. and Z.L.; methodology, Z.L.; software, Z.L. and X.Z.; formal analysis, Z.L.; investigation, Z.Z. and S.Y.; resources, G.X. and T.F.; data curation, Z.L.; writing—original draft preparation, Z.L., S.Y. and L.N.; writing—review and editing, Z.L.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

## Funding

This study was funded by the Science and Technology Major Project of Yunnan Province (Science and Technology Special Project of Southwest United Graduate School - Major Projects of Basic Research and Applied Basic Research) : Vegetation change monitoring and ecological restoration models in Jinsha River Basin mining area in Yunnan based on multi-modal remote sensing (Grant No. : 202302AO370003), Science and Technology Plan Project of Yunnan Province Science and Technology Department (Grant No. 202101BA070001-145), The Geological Survey Fund Project of Yunnan Province (D201711), Yunnan Province Project (Grant No. YNZZ202402-29), and Yunnan Province Project (Grant No. YNZZ202402-30).

## Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to Z.Z.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025