

An improved slime mould algorithm applied to hyperspectral feature selection in soil fertility indicators

Lijun Tian^{a,b,c,1}, Chao Niu^{a,b,c,1}, Xue Wang^{a,b,c},
Yayu Yang^{a,b,c}, Kun Tan^{a,b,c,*}, Xiaojun Xu^d, Wen Zhang^e

^a School of GeoAI and Hindon STAI Institute, East China Normal University, Shanghai 200241, China

^b Key Laboratory of Geographic Information Science (Ministry of Education), East China normal University, Shanghai 200241, China

^c Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources, East China Normal University, Shanghai 200241, China

^d Shanghai Environmental Monitoring Center, Shanghai 200003, China

^e Shanghai Municipal Institute of Surveying and Mapping, Shanghai 200063, China

ARTICLE INFO

Keywords:

Hyperspectral imagery
Spectral feature selection
Improved slime mould algorithm
Soil fertility
Machine learning

ABSTRACT

High-precision soil quality monitoring is vital for precision agriculture. Recently, swarm intelligence (SI) algorithms have become essential for identifying hyperspectral features. Aiming at the challenges of high dimensionality in soil fertility indicator retrieval, this paper integrates ZY1-02E satellite and HyMap airborne data to conduct feature selection and fertility estimation using an improved slime mould algorithm (ISMA). ISMA integrates three synergistic stages: (1) spectrally-informed initialization and pre-screening for redundancy reduction, (2) a multi-level location update strategy for balancing exploration and exploitation, and (3) a complementary stagnation control strategy for late stage search refinement. A customized fitness function is designed to optimize the estimation of soil fertility indicators. The effectiveness of the proposed ISMA was first validated on ZY1-02E satellite imagery. Experimental results demonstrated that the ISMA-XGBoost model yielded the best optimization capability, achieving the test R^2 of 0.752 for soil organic matter (SOM) on ZY1-02E data, and its multi-indicator applicability was confirmed with R^2 values of 0.675 (total phosphorus), 0.670 (total potassium), and 0.678 (total nitrogen). The model's reliability was validated on independent HyMap data ($R^2 = 0.757$), a significant improvement over conventional algorithm ($R^2 = 0.730$). These findings provide critical technical support for regional precision agriculture.

1. Introduction

As a pivotal determinant of soil health, the accurate quantification of fertility status is a vital prerequisite for effective agricultural production and achieving key Sustainable Development Goals (SDGs) (Kok et al., 2024; Ji et al., 2025; Zhang et al., 2025). Given the importance, soil management urgently requires tools for rapid, non-destructive estimation. Visible and near-infrared (vis-NIR) spectroscopy enables a time-efficient and economically viable alternative to quantitatively assess these soil fertility parameters (Shin et al., 2025).

The spectral response of soil fertility indicators is not uniformly distributed across the wavelength domain but is concentrated in specific physicochemically meaningful regions. As a vital soil component, soil

organic matter (SOM) exerts a strong inverse influence on spectral reflectance (Ladoni et al., 2010). Galvao and Vitorello (1998) identified SOM-related spectral features within the 550–700 nm region. More recently, Wu et al. (2023) highlighted the high predictive weight of bands at 678, 755, 767, and 831 nm in SOM estimation. Other studies investigated total nitrogen (TN)-sensitive features in the 560–600 nm region (acting as indicators via soil color) and the 893–975 nm region (sensitive to N–H, C–H, and O–H groups) (Maarouf et al., 2022; Diaz-Gonzalez et al., 2022). Compared to the extensive research on SOM and TN, Li et al. (2023) identified features for total phosphorus (TP) in the 250–750 nm region. For total potassium (TK), the spectral features have been identified in the 1765–2357 nm region, primarily attributed to the characteristic of potassium hydrates and ionic forms (Liu et al.,

* Corresponding author.

E-mail address: tankuncu@gmail.com (K. Tan).

¹ Lijun Tian and Chao Niu contributed equally to this work, and should be considered co-first authors.

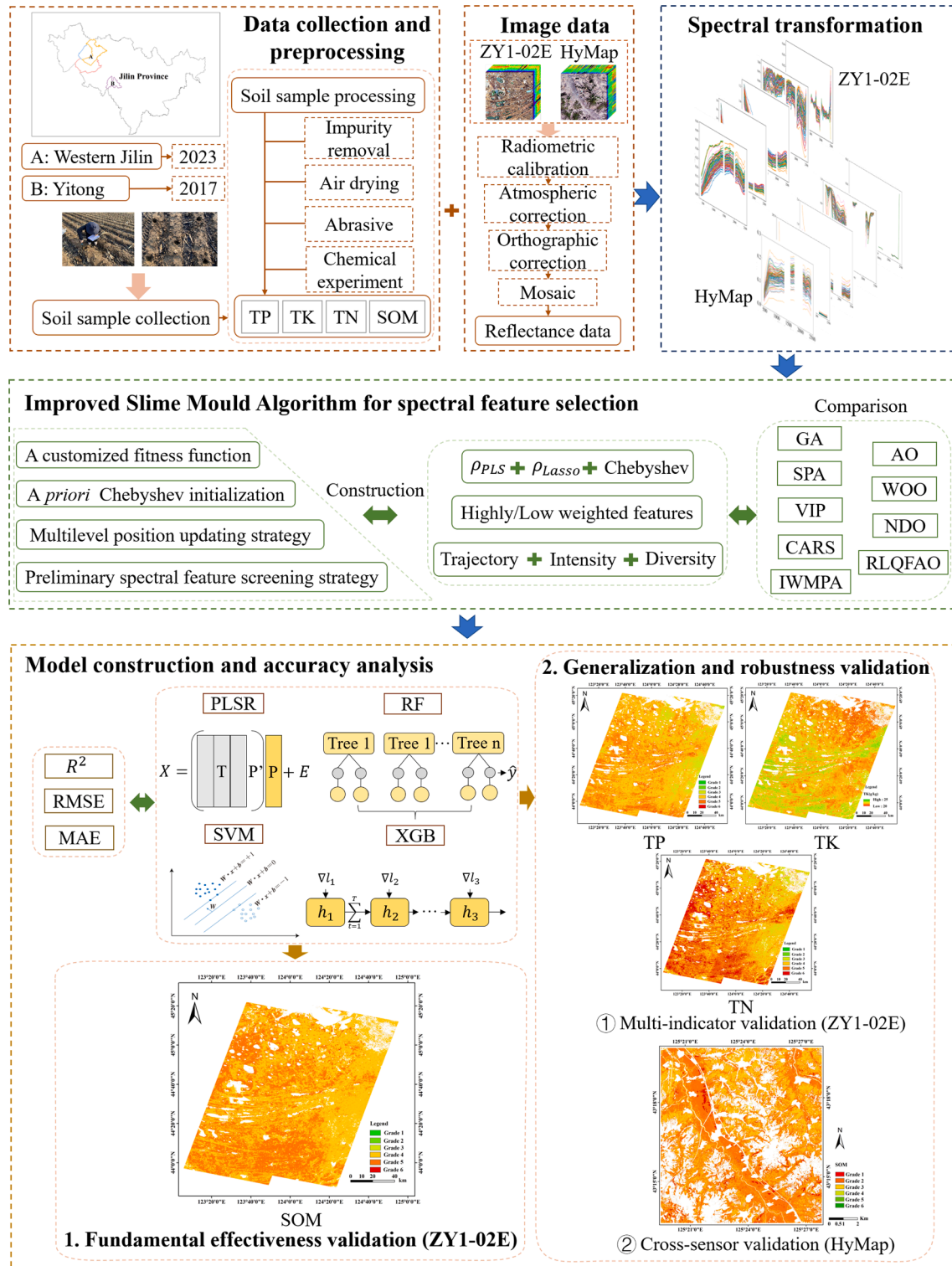


Fig. 1. Flowchart of this research.

2025).

However, modeling of soil properties via spectral analysis is often impeded by the high dimensionality and intrinsic complexity of hyperspectral datasets. Therefore, spectral feature dimensionality reduction methods must be applied to optimize feature representation and accentuate the spectral characteristics of soil (Asadzadeh and De Souza Filho, 2016). In hyperspectral remote sensing, feature dimensionality reduction is principally achieved via two strategies—feature selection

and feature extraction—both of which have seen widespread application (Esmaeili et al., 2023; Bai et al., 2024; Liu et al., 2024). Traditional statistical methods and exhaustive search strategies become computationally prohibitive or get easily trapped in local optima when dealing with such vast combinatorial spaces. Consequently, there is an urgent need for more robust, intelligent search mechanisms that can traverse these complex feature spaces more efficiently. Accordingly, metaheuristic-based optimization strategies offer a promising approach,

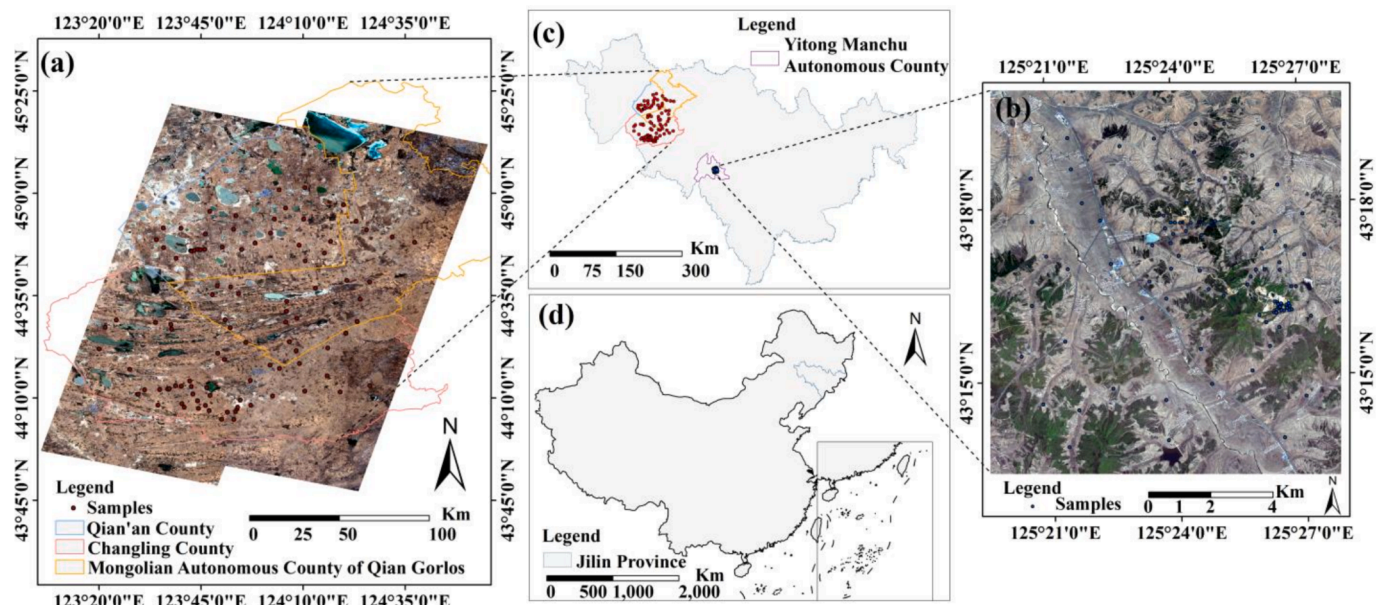


Fig. 2. Overview of the study region.

Table 1
Specifications of HyMap-C imaging spectrometer.

Parameter	Parameter description
Spectral range	450–2500 nm
Number of bands	144
Spectral resolution	15 nm @ 450–905 nm, 880–1440 nm 18 nm @ 1400 nm–1960 nm, 1950–2500 nm
Field of view	60°
Radiance accuracy (in flight)	95%
Scan rate	5–25 lines/s Continuous adjustable

Table 2
Load parameters of ZY1-02E AHSI.

Parameter	Parameter description
Spectral range	400–2500 nm
Number of bands	76 (VNIR) 90 (SWIR)
Spectral resolution	10 nm (VNIR) 20 nm (SWIR)
Spatial resolution	30 m
Swath width	60 km
Revisit period	55 days

Table 3
Statistical characteristics of the soil covariate content.

Year	Data source	Soil fertility indicators	Min (g/kg)	Max (g/kg)	Mean (g/kg)	SD (g/kg)	CV (%)
2017	HyMap	SOM	19.760	49.840	30.010	6.380	21
2023	ZY1-02E	TP	0.240	0.804	0.496	0.122	25
		TK	19.200	24.700	21.630	1.196	6
		TN	0.227	2.300	0.990	0.337	34
		SOM	4.300	24.300	13.680	4.133	30

using global search mechanisms to effectively balance feature relevance and redundancy reduction (Su, 2022).

In recent years, various SI algorithms, such as Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), and Genetic Algorithm (GA) (Maroufpoor et al., 2019; Dada et al., 2025; Qu et al., 2025), have

Table 4
Calculation formulae for the spectral transformations.

Spectral transformation	Formula
Original reflectance (OR)	$R(\lambda)$
Reciprocal reflectance (RR)	$RR(\lambda) = \frac{1}{R(\lambda)}$
Logarithmic reciprocal reflectance (LgRR)	$LgRR(\lambda) = \lg RR(\lambda)$
Continuum removal (CR)	$CR(\lambda) = \frac{R(\lambda)}{R'(\lambda)}$
First-order derivative reflectance (FDR)	$FDR_{(i)} = \frac{R_{\lambda_{i+1}} - R_{\lambda_i}}{\lambda_{i+1} - \lambda_i}$
Ratio index (RI)	$RI_{(ij)} = \frac{R_{\lambda_j}}{R_{\lambda_i}}$

been adapted for soil property estimation. However, the efficacy of these methods is often constrained by a fundamental mismatch between algorithmic mechanisms and hyperspectral complexity. For instance, while PSO and GWO are computationally efficient, their search trajectories are prone to stagnation in high-dimensional spaces. In PSO, the rapid collapse of population diversity frequently leads to premature convergence in suboptimal spectral regions (Amini and Hu, 2021; Abasabadi et al., 2022). Similarly, newer metaheuristics like the Dung Beetle Optimizer (DBO) and Whale Optimization Algorithm (WOA) have introduced adaptive boundaries and multi-swarm structures to mitigate stagnation (Miao et al., 2025; Xia et al., 2025). Yet, these improvements primarily address exploration–exploitation balance on standard benchmarks and often fail to account for the spectral noise in soil datasets or to maintain predictive robustness across diverse soil matrices (Got et al., 2021; Jiang et al., 2024; Tan et al., 2025).

Taken together, the current limitations of SI-based hyperspectral feature selection can be summarized in three aspects. First, most algorithms rely on a uniform search mechanism, which treats all spectral bands with equal initial importance. However, soil organic matter (SOM) signals are often concentrated in narrow, physically meaningful absorption bands (e.g., C–H or O–H overtones), which are easily obscured by redundant spectral noise during a purely stochastic search. Moreover, traditional SI algorithms rely on purely stochastic search updates, which lack the domain-aware guidance needed to prioritize physically significant wavelength regions (Gui et al., 2019). Without a mechanism to amplify high-contribution features, the search process remains susceptible to human-induced bias and random noise. Existing

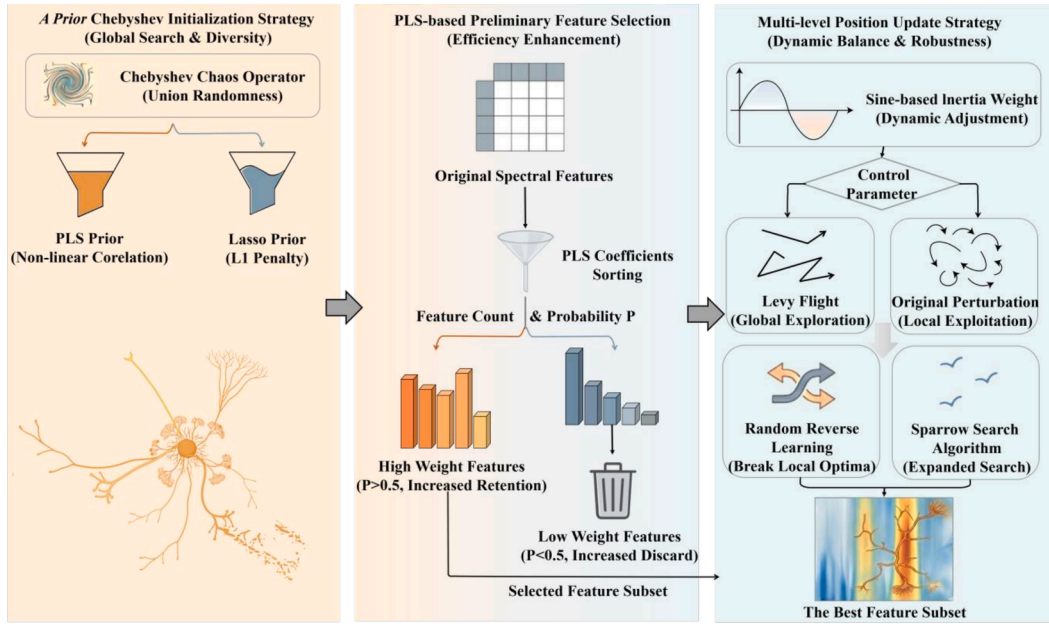


Fig. 3. Framework of ISMA.

Table 5
Pseudo-code of the ISMA.

Algorithm 1 ISMA**Input:** L : training set Max_Iter : maximum iteration limit N : total number of the population D : dimensionality of the feature space $Threshold$: minimum value required to select a feature ω_1, ω_2 , and ω_3 : the weight coefficients of the fitness function according to Eq. (17)**Begin**Initialize a population $\vec{P}$ according to Eq. (3)

Feature binarization according to Eq. (16)

Conduct a preliminary screening of the initialised $\vec{P}$ Calculate the initial $f(X_i)$ according to Eqs. (17)–(19)**for** $t \leftarrow 1$ **to** Max_Iter **do**Calculate weights W, vb , and vc Generate $rand^{1st}, rand^{2nd}$ **if** $rand^{2nd} < z$ **then**

Update position using Eq. (2)

else**if** $rand^{1st} < p$ **then**

Update position using sine-golden strategy according to the first expression in Eq. (4)

else

Update position using Lévy flight strategy using the second and third expressions in Eq. (4)

end if**end if**

Apply random reverse learning and sparrow search perturbations according to Eqs. (12)–(15)

Generate $rand$ and calculate p_v according to Eq. (14)**if** $rand < p_v$ **then**

Use Eq. (15) to determine the SSA perturbation form and update position using the first branch of Eq. (13)

else

Apply OBL and update position using the second branch of Eq. (13)

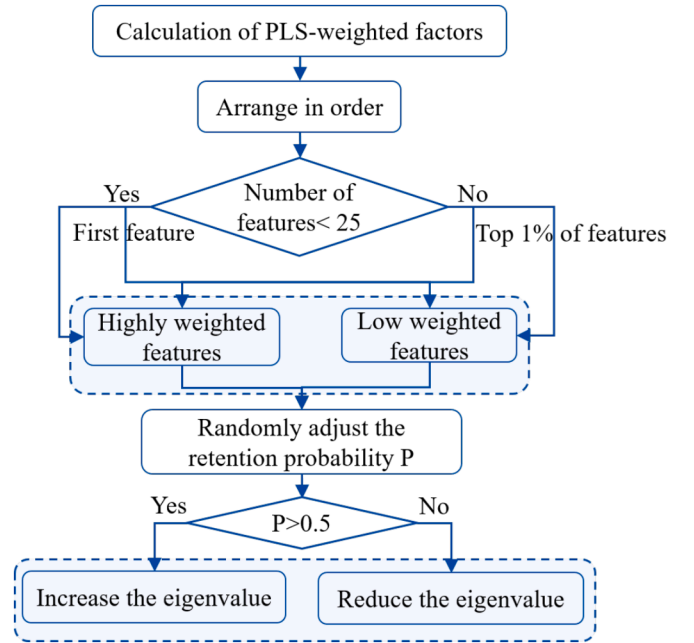
end ifUpdate the position of $\vec{P}$ Evaluate the solution by $f(X_i)$ according to Eqs. (17)–(19)**end for****return** the best solution X^{best} **End**

Fig. 4. The preliminary spectral feature screening strategy.

evaluation frameworks are often narrow in scope, focusing on a single indicator (Guo et al., 2024; Tan et al., 2024) or a single sensor platform (Shi et al., 2023; Saberioon et al., 2024; Wang et al., 2026). This leads to algorithms that are overfitted to specific spectral configurations, lacking the cross-platform transferability required for operational soil monitoring. Among recent SI methods, SMA has attracted attention because of its simple structure and competitive optimization capability (Li et al., 2020). However, the original SMA relies heavily on the current best individual and stochastic differential updates, which can lead to population aggregation around local optima, insufficient exploration of alternative spectral regions, and unstable band selection under repeated runs. These limitations become more pronounced in hyperspectral soil analysis, where the search space is highly dimensional, spectrally

Table 6
The different spectral feature selection strategies considered in this research.

Spectral feature selection method	Description
PCC	Measures linear correlation between spectral features.
CARS	Iteratively retains wavelengths with large absolute coefficients in PLS models.
VIP	Calculates feature importance based on regression coefficients and weight vectors.
SPA	Iteratively selects wavelength combinations with minimal redundancy.
GA	Performs stochastic global search for optimal feature subsets.
IWMPA	A meta-heuristic feature selection algorithm optimized for SOM estimation.
WOO	Employs diffraction and interference of optical waves for search expansion and information synergy.
AO	Simulates the hierarchical hunting strategies and predatory behaviors of the Aquila.
NDO	Hybridizes Newton-Downhill mathematical precision with meta-heuristic global exploration.
RLQFAO	Models four core predatory stages of the Aquila within a meta-heuristic search framework.
BASADMO	A two-stage hyperspectral feature selection method using hybrid Fisher-Relief and improved dwarf mongoose optimization.
GWOGA	A hybrid GWO-GA feature selection method that reduces bands via chaotic initialization, elite learning, and dynamic search balance.

Table 7
The parameters of ISMA.

Parameter	ZY1-02E				HyMap
	SOM	TP	TN	TK	SOM
L	75	75	75	75	69
Max_Iter	300	200	200	400	100
N	30	30	30	30	30
D	75	75	75	75	69
$Threshold$	0.55	0.5	0.5	0.6	0.6
$(\omega_1, \omega_2, \omega_3)$	(8,1,1)	(8,1,1)	(8,1,1)	(8,1,1)	(8,1,1)

collinear, and easily perturbed by local noise. Therefore, improving SMA for this task requires not only a better exploration–exploitation balance, but also a search framework that is better aligned with the structural properties of soil spectral data.

To overcome these shortcomings, this study aimed to: (1) develop an improved slime mould algorithm (ISMA) for hyperspectral feature

selection, which consists of three coordinated stages: spectrally informed initialization and pre-screening for redundancy reduction, a multi-level location update strategy for balanced exploration and exploitation, and a complementary stagnation control strategy for late stage search refinement, and an adaptive fitness function for effective guidance; (2) establish the fundamental effectiveness of the proposed ISMA by applying it to the estimation of SOM using ZY1-02E satellite data. Subsequently, we (3) validate its broader application by estimating three other soil fertility indicators (TP, TN, and TK) with the same data. Finally, to further (4) assess its cross-platform generalization and practical applicability, the framework is applied to an independent SOM estimation task using high-resolution HyMap airborne hyperspectral image dataset, intending to establish a theoretical foundation for environmental protection. Fig. 1 illustrates the flowchart of this paper.

2. Materials and methods

2.1. Study area

Two typical regions in Jilin Province, China, were selected for this investigation (Fig. 2). The first study area, situated in western Jilin (123.09°E–124.92°E, 43.78°N–45.37°N), encompasses three counties: (1) the Mongolian Autonomous County of Qian Gorlos, located in the southern Songnen Plain where chernozem and meadow soils are predominant; (2) Changling County, which displays significant nutrient variations; and (3) Qian'an County, where soils are generally slightly alkaline. The second study area is Yitong Manchu Autonomous County in Siping City (125.33°E–125.47°E, 43.22°N–43.33°N). This region is dominated by a topography of low mountains and rolling hills, creating a complex and varied landscape.

Table 8
Correlations of OR and RI with soil fertility indicators.¹

Data source	Soil fertility indicators	$ \rho _{\max}$	
		OR	RI
HyMap	SOM	0.511	0.633
ZY1-02E	TP	0.303	0.399
	TK	0.272	0.337
	TN	0.531	0.487
	SOM	0.551	0.494

¹ In Table 8, $|\rho|_{\max}$ denotes the maximum absolute correlation coefficient.

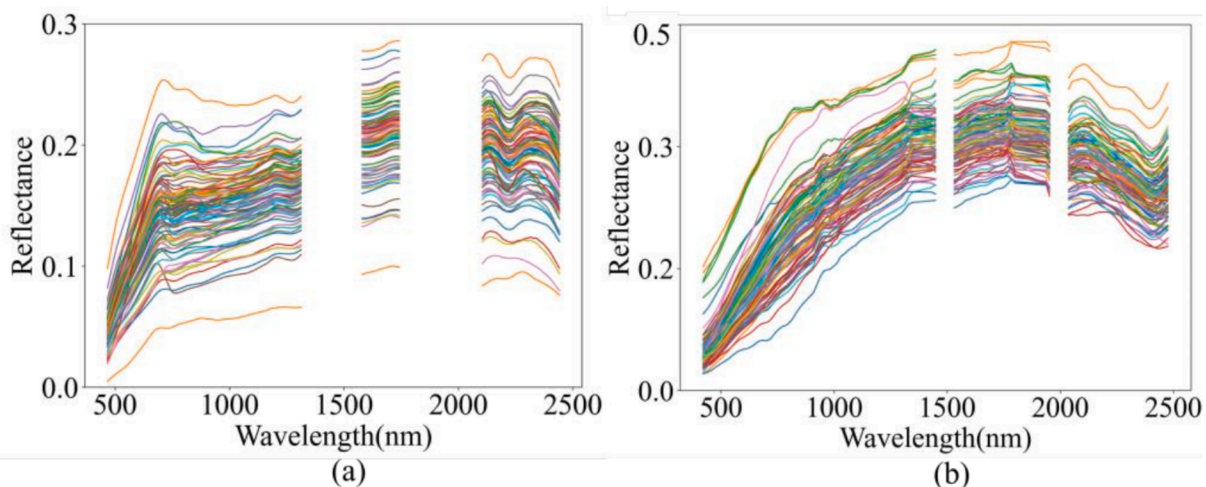


Fig. 5. Spectral reflectance characteristics corresponding to the selected sampling locations after preprocessing. (a) HyMap airborne hyperspectral imagery. (b) ZY1-02E satellite imagery.

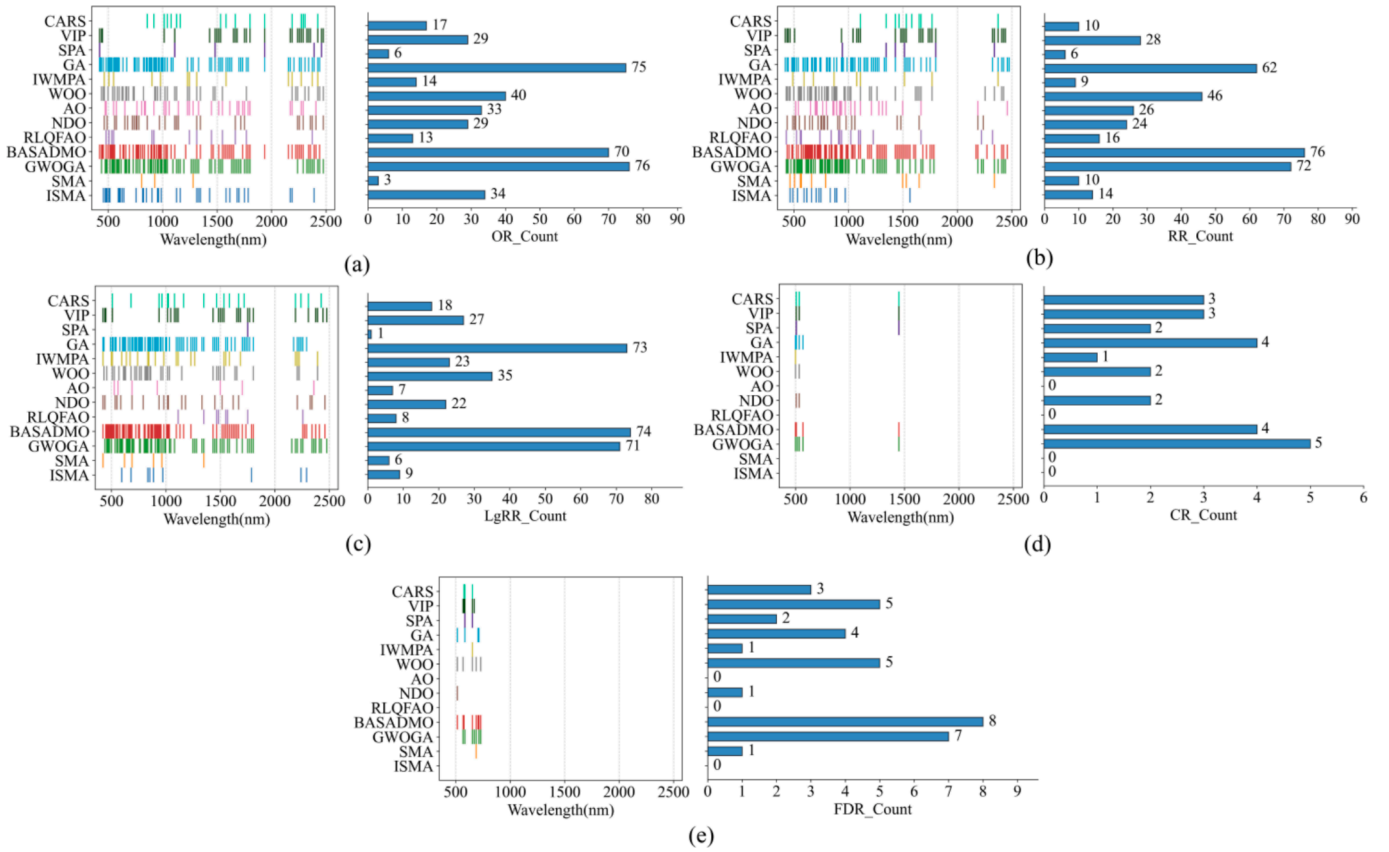


Fig. 6. Feature selection results for SOM corresponding to the different feature selection methods based on ZY1-02E satellite image dataset. (a) OR. (b) RR. (c) LgRR. (d) CR. (e) FDR.

2.2. Data acquisition

2.2.1. Hyperspectral data acquisition and preprocessing

HyMap hyperspectral data was collected in early May 2017, the specifications of which are detailed in Table 1. Following standard preprocessing, the final imagery features the spatial resolution of 4.5 m and the spectral resolution of 10–20 nm. Six hyperspectral images acquired by the ZY1-02E satellite on 29 March 2023 and 21 April 2023 were used in this study. The specifications of the Advanced Hyperspectral Imager (AHSI) payload are detailed in Table 2.

2.2.2. Acquisition of soil samples

Field soil sampling was conducted in two separate campaigns for the respective study areas. For the western Jilin study area, 100 samples were collected in mid-April 2023. For the Yitong Manchu Autonomous County study area, 93 samples were collected in May 2017. The spatial distributions of the sampling points for each area are depicted in Fig. 2 (a) and Fig. 2(b). The statistical characteristics of the soil properties are listed in Table 3.

2.2.3. Spectral transformation

As detailed in Table 4, we employed spectral scaling transformations to mitigate noise interference and enhance feature extractability. The resulting transformed spectra were subsequently combined with the ratio index (RI) to generate the input datasets required for the modeling algorithm.

where $R'(\lambda)$ represents corresponding calculated continuum values. R_{λ_j} and R_{λ_i} signify the spectral response at the specific wavelengths λ_j and λ_i , respectively, with $\lambda_i \neq \lambda_j$.

2.3. Algorithm and modeling

2.3.1. Slime mould algorithm

The core mechanism of ISMA relies on the propagation wave of a bio-oscillator to generate positive and negative feedback, represented by the weight W (1). While effective, the standard position update mechanism heavily relies on random parameters (z , $rand^{1st}$ and $rand^{2nd}$), which can lead to premature convergence and limited exploration capability (Li et al., 2020).

$$W = \begin{cases} 1 + rand \times \log\left(\frac{bF - Sort(i)}{bF - wF} + 1\right), & i < \frac{N}{2} \\ 1 - rand \times \log\left(\frac{bF - Sort(i)}{bF - wF} + 1\right), & i \geq \frac{N}{2} \end{cases} \quad (1)$$

where bF and wF denote the optimal and poorest fitness scores, respectively; $Sort(i)$ represent the ranked fitness scores; and $rand$ represents a random value in [0,1].

The mechanism for updating agent locations during spectral feature encircling can be formulated as:

$$X_i^{t+1} = rand \times (UB - LB) + LB, rand^{2nd} < z \quad (2)$$

where $rand^{2nd}$ is a random value in [0,1]. The algorithm performs best in searching for the optimal performance when $z = 0.03$; UB and LB correspond to the search space limits.

2.3.2. Improved slime mould algorithm for spectral feature selection

Building upon SMA, we developed ISMA, which is specifically optimized to address the feature selection challenges of soil fertility indicator content estimation. Fig. 3 illustrates the framework of ISMA, while Table 5 details its corresponding pseudocode.

1) Slime mould population initialization

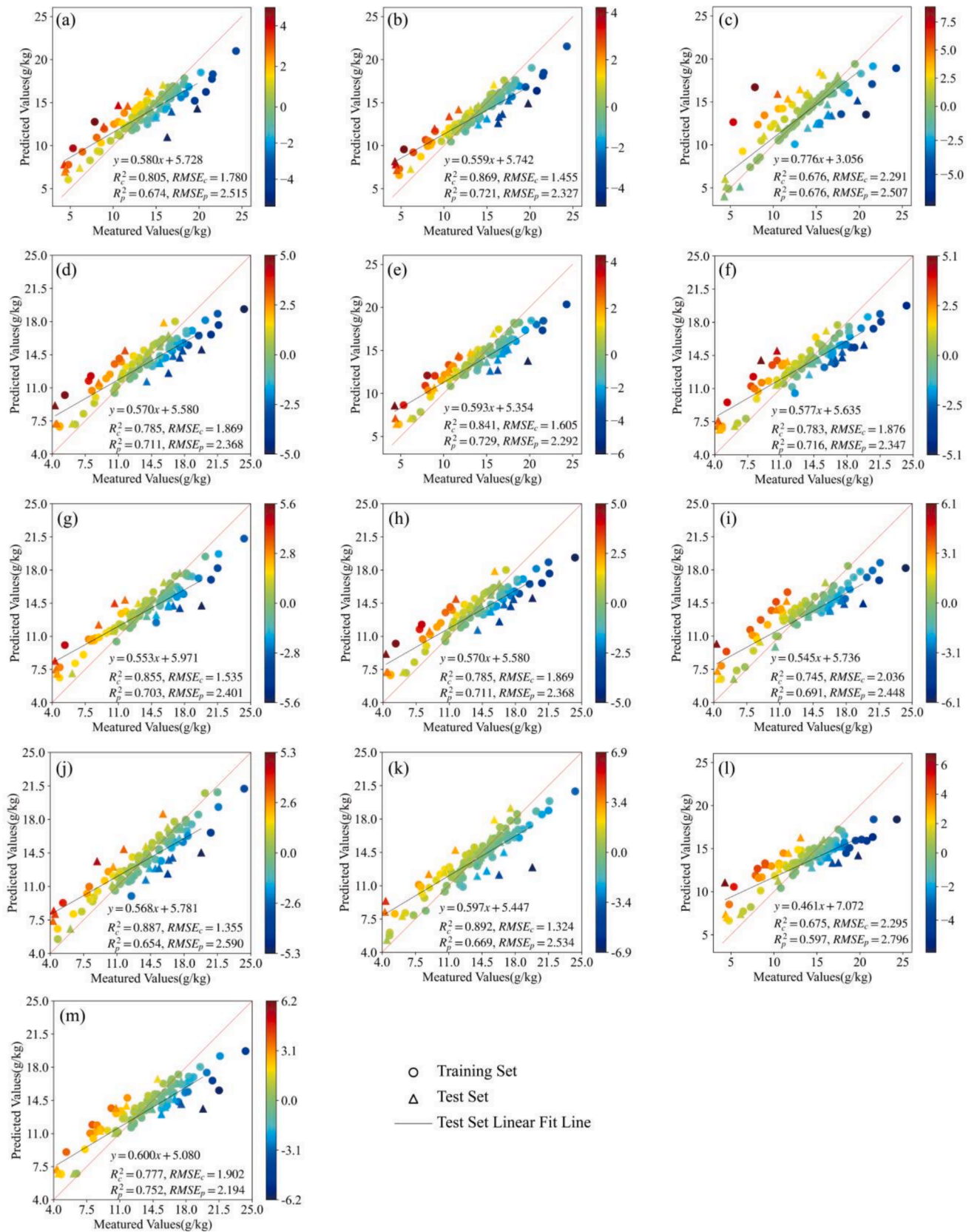


Fig. 7. Scatter plots of predictions for the optimal SOM models based on ZY1-02E satellite image dataset (The heatmap represents the residuals between the estimated and actual values). (a) SPA-RF. (b) VIP-RF. (c) GA-SVM. (d) CARS-XGBoost. (e) IWMPA-XGBoost. (f) WOO-SVM. (g) AO-XGBoost. (h) NDO-XGBoost. (i) RLQFAO-XGBoost. (j) BASADMO-RF. (k) GWOGA-RF. (l) SMA-XGBoost. (m) ISMA-XGBoost.

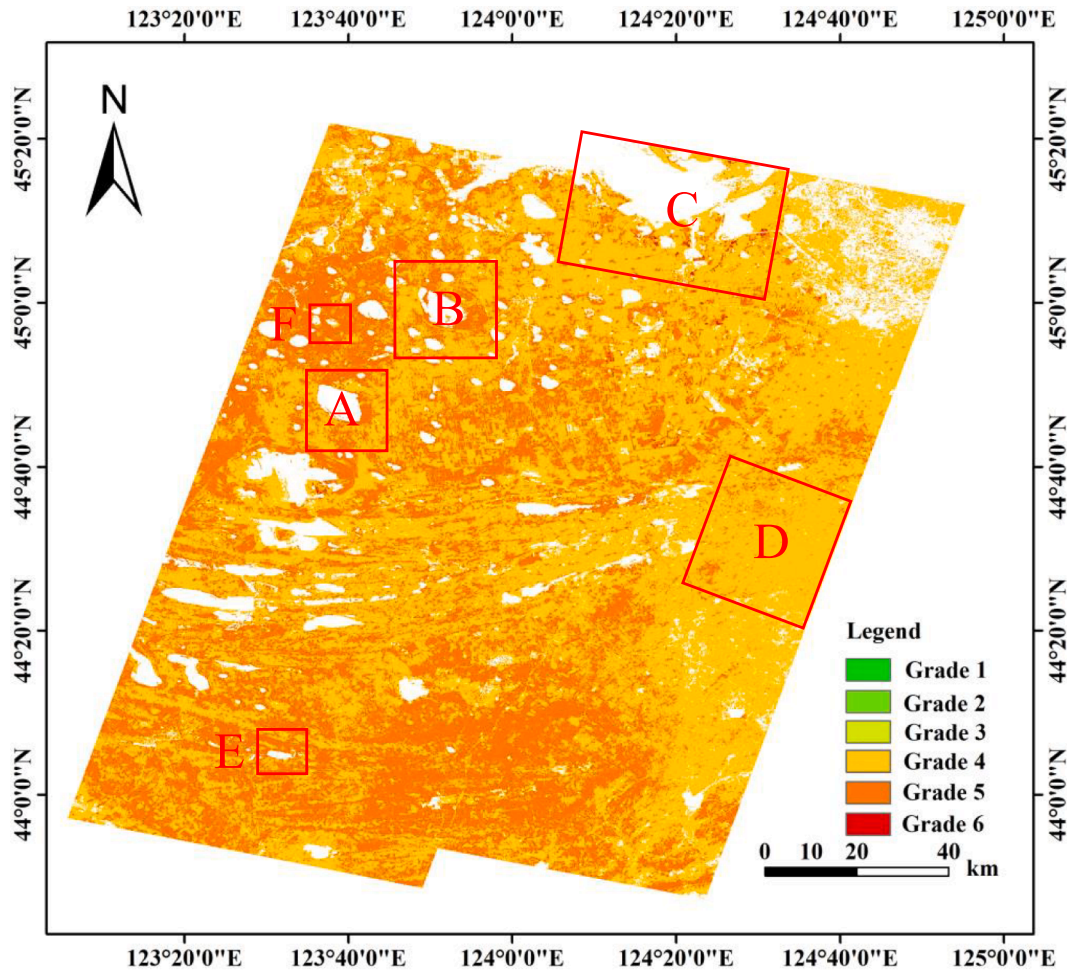


Fig. 8. Spatial mapping of SOM content based on ZY1-02E satellite image dataset.

An *a priori* Chebyshev initialization (Eq. (3)) integrating PLS and Lasso coefficients is employed to prioritize key features. While PLS coefficients are effective in capturing the intricate associations between spectral bands and soil attributes, the method is prone to incorporating uninformative or weakly correlated features. To counteract this, Lasso integration utilizes L1 regularization to induce sparsity. By shrinking the coefficients of redundant variables—which PLS might otherwise retain—to zero, the feature set is effectively streamlined, thereby enhancing both its parsimony and predictive generalizability.

$$X_{initial} = \text{Chebyshev}(X_{rand}) + \rho_{PLS} + \rho_{lasso} \quad (3)$$

2) Multi-level location update strategy

Although the initialization step improves the starting population, it cannot solve the main difficulties of hyperspectral feature selection by itself. During iteration, the standard SMA often converges too quickly around the current best solution. To address this problem, we introduce the multi-level update strategy in Eq. (4). Unlike the original SMA, this algorithm is better equipped to navigate the high-dimensional search space and identify superior feature subsets. When $\text{rand}^{2nd} \geq z$, the corresponding improved multi-level position update formula is as follows:

where $R_1 \in [0, 2\pi]$, $R_2 \in [0, \pi]$, $x_1 = -\pi + (1 - \tau) \cdot 2\pi$, $x_2 = -\pi + \tau \cdot 2\pi$, $\tau = \frac{\sqrt{5}-1}{2}$, $q = 0.9$; vb and vc represent interaction and retention factors; X_i^t , X_A^t , and X_B^t are the current and random individual positions; and $\text{rand}^{1st} \in [0, 1]$. In Eq. (4), ω controls the overall step magnitude and gradually adjusts the search intensity over iterations. R_1 and R_2 introduce sine-based modulation to diversify movement direction and amplitude, while vb preserves the interaction mechanism inherited from SMA. The golden-ratio-based variables x_1 and x_2 provide bounded coefficients for the oscillatory update, so that the search around the current promising region remains structured.

For intuitive interpretation, the three expressions in Eq. (4) correspond to three complementary search modes. The first mode performs a structured oscillatory search around the current promising region by combining the best solution with population difference information and sine modulation. In this mode, x_1 and x_2 , which are determined by the golden ratio τ , define the relative search interval and help prevent excessively random displacement. The second mode combines best-solution guidance, individual memory, and Lévy perturbation, thereby balancing exploration and exploitation during the intermediate search

$$X_i^{t+1} = \begin{cases} \omega \times X_{best}^t \times |\sin(R_1)| + \omega \times R_2 \times \sin(R_1) \times vb \times |W \times (x_1 X_A^t - x_2 X_B^t)| & \text{rand}^{1st} < p \\ X_{best}^t + vc \times X_i^t + \theta \times Levy & p \leq \text{rand}^{1st} < q \\ vc \times X_i^t + \theta \times Levy & q \leq \text{rand}^{1st} \end{cases} \quad (4)$$

Table 9
The nutrient classification criteria adopted by the second national soil census of China.

Fertility indicator (g/kg)	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
SOM	>40	30–40	20–30	10–20	6–10	<6
TP	>1	0.8–1.0	0.6–0.8	0.4–0.6	0.2–0.4	<0.2
TK	>25	20–25	15–20	10–15	5–10	<5
TN	>2	1.5–2.0	1.0–1.5	0.75–1.0	0.5–0.75	<0.5

stage. The third mode removes the direct dependence on the global best and relies on the current individual together with Lévy perturbation, which helps the population escape repeated convergence to the same local region. In this way, Eq. (4) is used to improve the search process after initialization, rather than leaving the algorithm confined to one fixed update behavior.

$$p = \tanh|\text{Sort}(i) - \text{Gof}|$$
 (5)

$$vb = 2 \times a \times r - a$$
 (6)

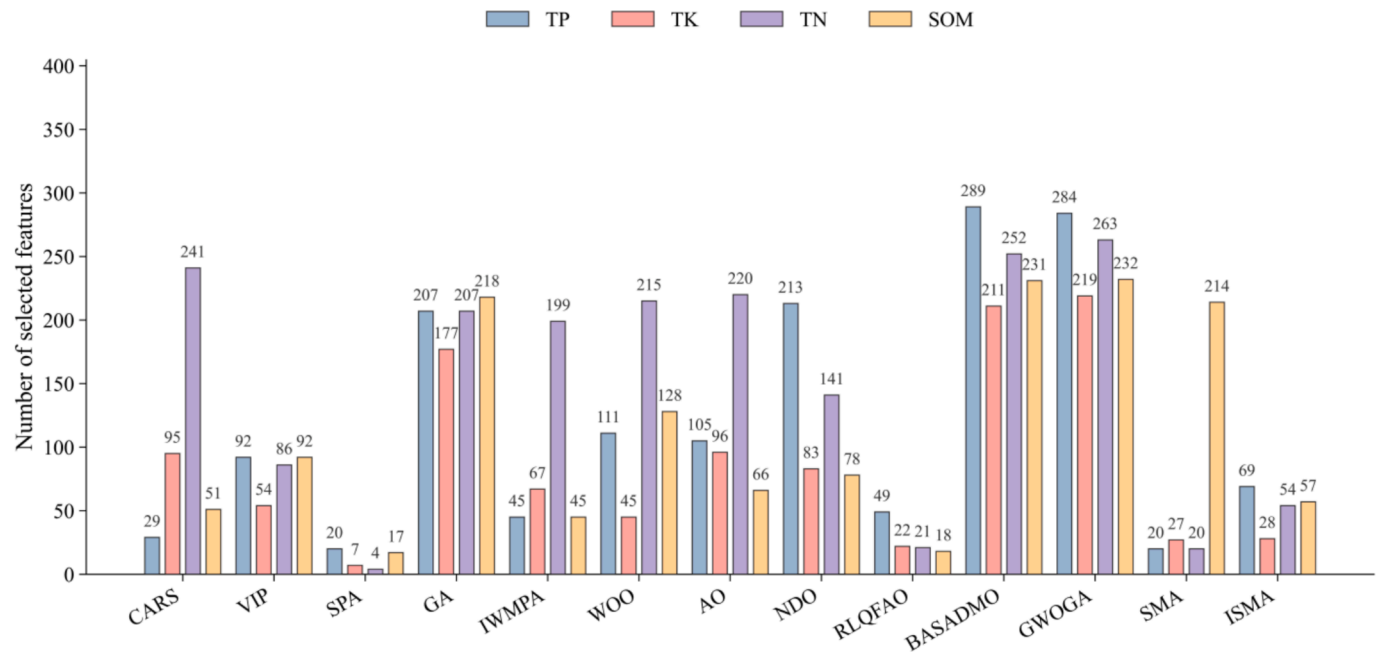


Fig. 9. Comparative evaluation of feature selection methods in estimation of soil fertility indicators.

Table 10
Estimation of soil fertility indicator content using ISMA and SMA based on ZY1-02E satellite image dataset.²

Method	Indicators	Model	Train			Test		
			R _c ²	RMSE _c (g/kg)	MAE _c (g/kg)	R _p ²	RMSE _p (g/kg)	MAE _p (g/kg)
SMA	TP	PLSR	0.191	0.110	0.087	0.148	0.110	0.092
		XGBoost	0.695	0.068	0.056	0.321	0.098	0.082
		SVM	0.531	0.084	0.070	0.284	0.101	0.095
	TK	RF	0.891	0.040	0.031	0.532	0.081	0.069
		PLSR	0.442	0.955	0.782	0.252	0.773	0.637
		XGBoost	0.471	0.930	0.750	0.253	0.773	0.608
		SVM	0.414	0.979	0.711	0.335	0.729	0.591
	TN	RF	0.504	0.900	0.701	0.292	0.752	0.610
		PLSR	0.473	0.265	0.197	0.439	0.161	0.132
		XGBoost	0.619	0.225	0.162	0.506	0.151	0.127
		SVM	0.505	0.256	0.167	0.494	0.153	0.130
		RF	0.845	0.131	0.093	0.476	0.158	0.131
ISMA	TP	PLSR	0.429	0.093	0.074	0.051	0.116	0.098
		XGBoost	0.834	0.050	0.040	0.667	0.069	0.057
		SVM	0.721	0.065	0.053	0.598	0.076	0.063
	TK	RF	0.886	0.042	0.033	0.675	0.068	0.056
		PLSR	0.250	1.108	0.923	0.220	0.791	0.627
		XGBoost	0.850	0.495	0.394	0.670	0.514	0.434
		SVM	0.704	0.696	0.549	0.569	0.587	0.504
	TN	RF	0.731	0.664	0.531	0.612	0.557	0.471
		PLSR	0.472	0.265	0.202	0.411	0.166	0.162
		XGBoost	0.883	0.125	0.088	0.678	0.122	0.102
		SVM	0.613	0.227	0.165	0.585	0.139	0.114
		RF	0.794	0.166	0.121	0.642	0.129	0.142

² In Table 10, the subscripts ‘c’ and ‘p’ denote metrics derived from the training and test sets, respectively. The optimal precision values are emphasized in bold.

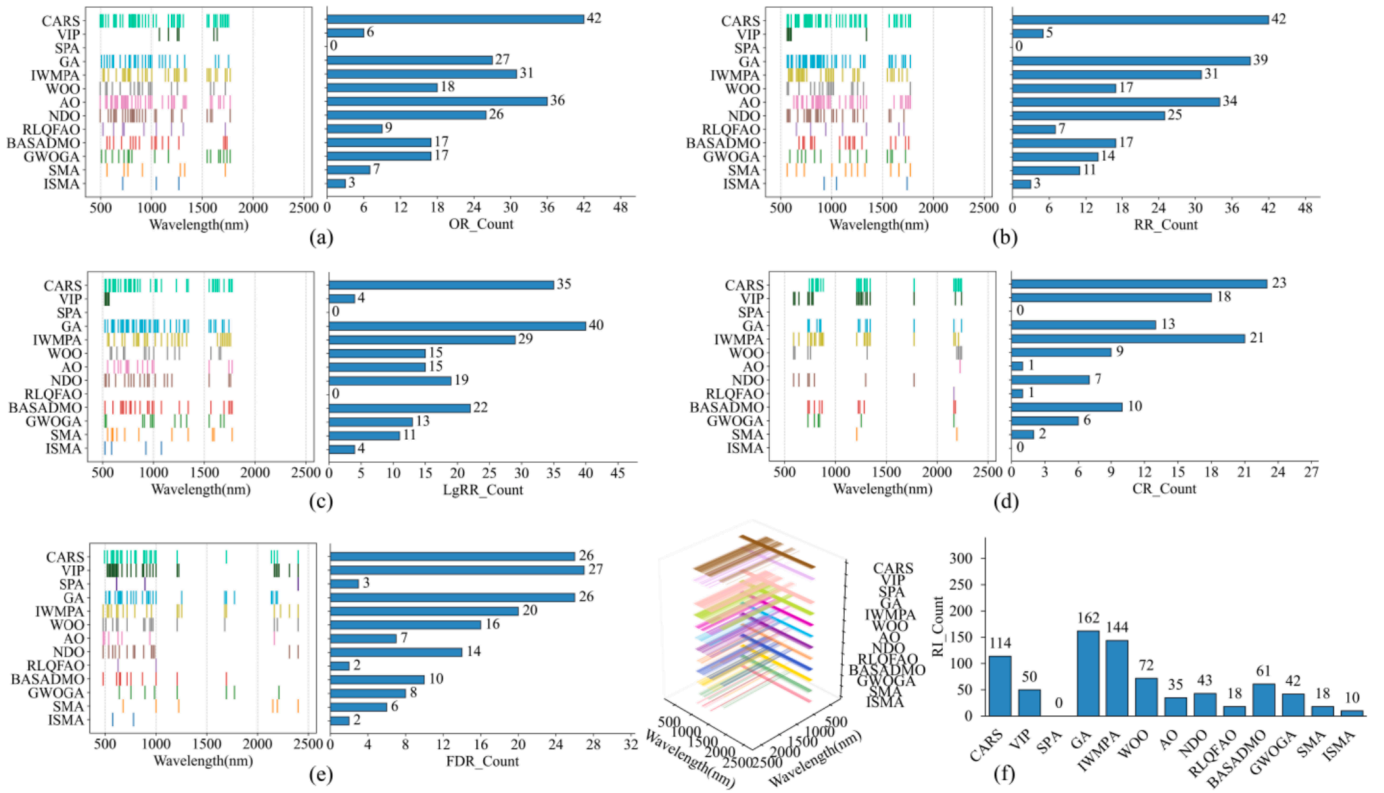


Fig. 10. Feature selection results for SOM using different feature selection methods based on HyMap airborne hyperspectral image dataset. (a) OR. (b) RR. (c) LgRR. (d) CR. (e) FDR. (f) RI.

Table 11

Common bands selected from different images under different spectral transformations.

Data Source	Wavelength
ZY1-02E	756.04 nm, 816.13 nm, 1750.95 nm
HyMap	754.57 nm, 817.77 nm, 1741.45 nm

$$vc = 2 \times b \times r - b \quad (7)$$

The expressions for a and b are as follows:

$$a = \arctanh\left(1 - \frac{t}{T}\right) \quad (8)$$

$$b = 1 - \frac{t}{T} \quad (9)$$

where $Sort(i)$ and Gof represent the ranked fitness and the global optimum's fitness; t is the iteration index.

We further constructed an evolutionary curve for the inertia weight, and sine functions can facilitate periodic dynamic adjustments within a predefined range (Shu et al., 2025). Furthermore, previous research found that a (0.4, 0.7) uniform distribution for late-stage inertia weights can improve precision (Hu et al., 2022). To adapt this to our specific research context, the threshold interval was systematically tuned and ultimately established as (0.4, 0.5). Within this interval, higher values facilitate global exploration by enabling the slime mould to traverse new regions with larger step sizes, whereas lower values prioritize local exploitation for fine-grained position adjustment. The formulation for the improved weight ω is defined as follows:

$$\omega = 0.4 + 0.5 \times rand + \sin\left(\pi + \frac{\pi t}{2Max_Iter}\right) \quad (10)$$

$$\theta = e^{\left(\frac{2t}{Max_Iter}\right)^3} \quad (11)$$

Even with improved initialization and multi-level updates, the search may still become concentrated in limited spectral regions, especially in later iterations. This problem is common in hyperspectral feature selection because many band combinations produce similar intermediate fitness values, which can cause stagnation. To reduce this risk, we add opposition-based learning (OBL) (Eq. (12)) and SSA-based perturbation (Eq. (15)) (Xue and Shen, 2020).

$$X_i^{t+1'} = UB + LB - rand \times X_i^{t+1'} \quad (12)$$

The role of OBL is to generate alternative candidate positions when the current population begins to cluster, so that the search can still examine unexplored regions of the spectral space. However, OBL mainly acts at a coarse scale and is not sufficient for later-stage fine adjustment. For this reason, we further introduce the SSA perturbation mechanism.

$$X_i^{t+1} = \begin{cases} X_i^{t+1'} & \text{if } rand < p_v \\ X_i^{t+1} & \text{otherwise} \end{cases} \quad (13)$$

$$p_v = \frac{2}{3} \left(1 - \left(\frac{t}{Max_Iter}\right)^{\frac{3}{2}}\right) \quad (14)$$

$$X_i^{t+1''} = \begin{cases} X_{best}^{t+1} + |X_i^{t'} - X_{best}^{t+1}| \times A^+ \times L & \text{if } rand < 0.6 \\ Norm \times \exp\left(\frac{X_{worst}^t - X_i^{t'}}{t^2}\right) & \text{otherwise} \end{cases} \quad (15)$$

where p is a probability factor; $Norm$ is a Gaussian random number; and X_{worst}^t and X_{best}^{t+1} denote the coordinates of the worst-performing and best-performing individuals at iteration t and the subsequent step $t + 1$; A represents a randomized binary matrix ($A \in \{-1, 1\}$) with the

Table 12

Estimation results for SOM based on HyMap airborne hyperspectral imagery.

Method	Model	Train			Test		
		R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	R_p^2	RMSE _p (g/kg)	MAE _p (g/kg)
SPA	PLSR	0.1479	5.9090	4.2944	0.1473	5.7211	4.4065
	XGBoost	0.7033	3.4870	2.7533	0.5175	4.3034	3.5674
	SVM	0.8353	2.5983	1.9484	0.5682	4.0712	3.4043
	RF	0.6446	3.8164	2.9589	0.4775	4.4783	3.2935
VIP	PLSR	0.4054	4.9360	4.0256	0.2561	5.3436	4.5632
	XGBoost	0.9290	1.7056	1.3980	0.5988	3.9241	3.2794
	SVM	0.5177	4.4455	2.7911	0.4450	4.6156	3.7763
	RF	0.8925	2.0987	1.7478	0.5438	4.1846	3.4770
GA	PLSR	0.4393	4.7934	3.9002	0.3090	5.1500	4.2554
	XGBoost	0.9210	1.7993	1.4683	0.5577	4.1206	3.4019
	SVM	0.5681	4.2070	2.7686	0.5126	4.3251	3.6742
	RF	0.9033	1.9907	1.6108	0.5674	4.0749	3.2199
CARS	PLSR	0.4777	4.6262	3.7786	0.2281	5.4433	4.4447
	XGBoost	0.8594	2.4003	1.9833	0.5713	4.0567	3.4305
	SVM	0.5388	4.3471	2.8820	0.4486	4.6005	3.9020
	RF	0.9045	1.9780	1.6093	0.5717	4.0546	3.1832
IWMPA	PLSR	0.4642	4.6858	3.8159	0.1047	5.8623	4.9145
	XGBoost	0.9188	1.8238	1.3833	0.5824	4.0038	3.2252
	SVM	0.5520	4.2847	2.7825	0.5517	4.1484	3.3948
	RF	0.8803	2.2151	1.7315	0.6259	3.7893	3.1295
WOO	PLSR	0.4185	4.8812	4.0114	0.0543	6.0248	5.1113
	XGBoost	0.9102	1.9180	1.4722	0.7298	3.2207	2.6788
	SVM	0.8577	2.4152	1.8789	0.7089	3.3425	2.8318
	RF	0.8735	2.2768	1.7685	0.7190	3.2844	2.7109
AO	PLSR	0.4091	4.9207	3.9754	0.3607	4.9535	4.2488
	XGBoost	0.8404	2.5575	2.0120	0.6682	3.5687	2.6089
	SVM	0.8008	2.8567	2.2747	0.6377	3.7290	2.9688
	RF	0.8844	2.1768	1.6952	0.5853	3.9895	3.2049
NDO	PLSR	0.4054	4.9362	4.0464	0.2204	5.4703	4.3934
	XGBoost	0.7594	3.1398	2.4625	0.6447	3.6927	3.2064
	SVM	0.7925	2.9159	2.2346	0.6383	3.7262	3.2068
	RF	0.8775	2.2406	1.7491	0.6468	3.6822	3.2903
RLQFAO	PLSR	0.4395	4.7926	3.8229	0.1343	5.7644	4.6694
	XGBoost	0.8232	2.6916	2.0740	0.7103	3.3345	2.5552
	SVM	0.7211	3.3803	2.6639	0.6892	3.4540	2.6492
	RF	0.8072	2.8107	2.1896	0.7085	3.3452	2.5908
BASADMO	PLSR	0.4541	4.7295	3.8303	0.3028	5.1729	4.2489
	XGBoost	0.8944	2.0792	1.7045	0.4322	4.6681	3.8935
	SVM	0.7747	3.0380	2.4966	0.4321	4.6686	3.9504
	RF	0.8383	2.5736	2.1350	0.4424	4.6260	3.9126
GWOGA	PLSR	0.4464	4.7627	3.8805	0.2833	5.2448	4.3801
	XGBoost	0.7871	2.9535	2.4732	0.4108	4.7555	3.8076
	SVM	0.6970	3.5233	2.8539	0.3997	4.7998	3.9005
	RF	0.7733	3.0478	2.5474	0.4128	4.7475	3.8185
SMA	PLSR	0.4282	4.6698	3.6935	0.3539	5.4670	4.1099
	XGBoost	0.8710	2.2181	1.7700	0.6109	4.2425	3.4422
	SVM	0.7749	2.9299	2.3224	0.5863	4.3743	3.3843
	RF	0.8903	2.0450	1.6572	0.5299	4.6631	3.4382
ISMA	PLSR	0.4091	4.9207	3.9754	0.3607	4.9535	4.2488
	XGBoost	0.8783	2.2332	1.7088	0.7566	3.0563	2.4846
	SVM	0.8195	2.7194	2.1929	0.6500	3.6650	2.8583
	RF	0.8363	2.5898	2.0223	0.6886	3.4573	2.6616

Table 13

Weighting schemes of the fitness function for different configurations.

Configuration	w ₁ (Accuracy)	w ₂ (Sparsity)	w ₃ (Importance)
E1	0.80	0.10	0.10
E2	0.90	0.05	0.05
E3	0.70	0.15	0.15
E4	0.60	0.20	0.20
E5	0.50	0.25	0.25
No Spa.	0.90	0.00	0.10
No Imp.	0.90	0.10	0.00
Acc.	1.00	0.00	0.00

pseudo-inverse $\mathbf{A}^+ = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$, and $\mathbf{L}$ signifies a unit matrix of ones.

Eq. (15) follows the original SSA update rule and is used here as a local perturbation operator. In our framework, OBL expands the search

when the population becomes too concentrated, whereas SSA refines the search or pushes individuals out of stagnant regions when further improvement becomes difficult. Therefore, the first two strategies improve the starting point and the general search trajectory, while OBL and SSA are added because those two steps alone still cannot fully prevent late-stage stagnation in a high-dimensional spectral search space.

3) Preliminary spectral feature screening strategy

To reduce the computational cost of the presented framework, PLS coefficients are employed for the preliminary screening of spectral features before the multi-level position update framework. To mitigate the influence of subjective intervention on feature selection, a stochastic triggering mechanism is introduced. Specifically, when the randomly generated value $P > 0.5$, the algorithm activates a weight polarization strategy. This process further amplifies the weights of significant features while suppressing those of redundant ones. By expanding the

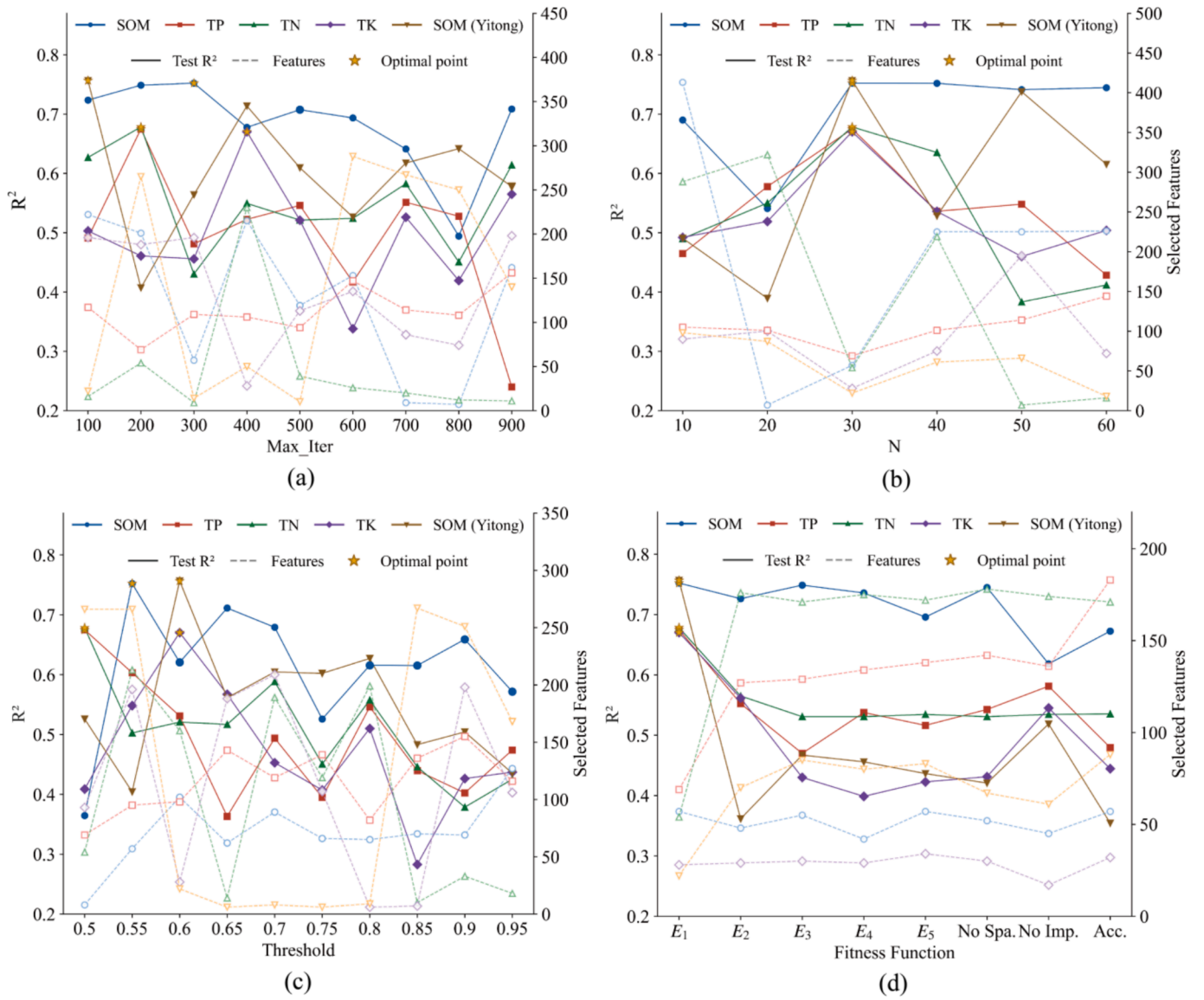


Fig. 11. Curves of parameter variations. (a) Max_Iter. (b) N. (c) Threshold. (d) Fitness Function.

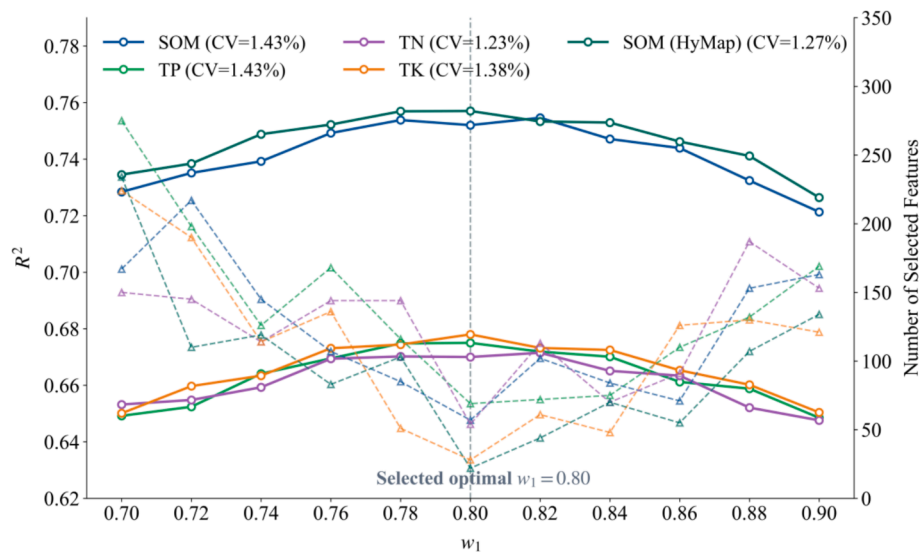


Fig. 12. Sensitivity analysis of the accuracy weight w_1 .

Table 14The results of retrieval performance driven by nonlinear feature importance (I_{XGB}).

Indicators	Model	Train			Test		
		R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	R_p^2	RMSE _p (g/kg)	MAE _p (g/kg)
SOM	PLSR	0.296	4.298	3.476	0.214	5.175	4.119
	XGBoost	0.830	1.657	1.228	0.692	2.444	1.941
	SVM	0.803	1.784	1.334	0.684	2.475	2.088
	RF	0.823	1.694	1.259	0.688	2.459	2.039
TP	PLSR	0.368	0.097	0.075	0.304	0.099	0.082
	XGBoost	0.910	0.036	0.029	0.434	0.089	0.075
	SVM	0.721	0.065	0.051	0.487	0.085	0.071
	RF	0.858	0.046	0.037	0.569	0.078	0.067
TK	PLSR	0.451	0.947	0.792	0.421	0.680	0.570
	XGBoost	0.833	0.512	0.434	0.598	0.571	0.489
	SVM	0.764	0.636	0.529	0.603	0.565	0.480
	RF	0.804	0.562	0.479	0.592	0.577	0.491
TN	PLSR	0.337	0.297	0.222	0.192	0.193	0.160
	XGBoost	0.671	0.209	0.158	0.500	0.152	0.116
	SVM	0.846	0.143	0.096	0.482	0.155	0.121
	RF	0.793	0.167	0.125	0.602	0.135	0.101
SOM (HyMap)	PLSR	0.377	5.050	4.161	0.232	5.426	4.312
	XGBoost	0.919	1.817	1.490	0.551	4.147	3.432
	SVM	0.859	2.400	1.977	0.550	4.155	3.508
	RF	0.785	2.961	2.433	0.516	4.307	3.659

Table 15

Design of the ablation study for ISMA with different strategy combinations.

Variant	Initial strategy (M1)	Pre-screening (M2)	Multi-layer update (M3)
ISMA	✓	✓	✓
M1	✓	–	–
M2	–	✓	–
M3	–	–	✓
M1 + M2	✓	✓	–
M1 + M3	✓	–	✓
M2 + M3	–	✓	✓

disparity in contribution degrees among features, this approach facilitates the retention of core spectral characteristics and the exclusion of irrelevant variables. A flowchart is provided in Fig. 4.

4) Calculation of the fitness function

The fitness function is considered by integrating three metrics: the dimensionality of selected features, their PLS-based importance weights, and the regression accuracy (R^2). Specifically, R^2 is computed via PLSR with 10-fold cross-validation, while the spectral features are identified through Eq. (16):

$$B_i = \begin{cases} 1 & \text{if } X_i > \text{Threshold} \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

Eq. (17) describes the degree of adaptation:

$$f(X_i) = \omega_1 T_{acc} + \omega_2 T_{spa} + \omega_3 T_{imp} \quad (17)$$

$$\begin{cases} T_{acc} = R^2 \\ T_{spa} = 1 - \frac{n}{N} \\ T_{imp} = \frac{\mu(|\rho_{PLS}|)}{\max(|\rho_{PLS}|) + \varepsilon} \end{cases} \quad (18)$$

where ω_1 , ω_2 , and ω_3 are the weight coefficients. In this study, we set $\omega_1 = 0.8$, $\omega_2 = 0.1$, and $\omega_3 = 0.1$. Specifically, the large weight assigned to T_{acc} creates the dominant selection pressure toward band subsets with strong predictive ability, which is necessary because the primary goal of the search is to maximize estimation performance. In contrast, T_{spa} and T_{imp} are assigned smaller weights so that they function as regularization terms rather than competing objectives. A relatively small sparsity weight is sufficient because feature count is already directly controlled

by the binary selection process and thresholding operation. If ω_2 is too large, the search can be driven toward overly compact subsets at the expense of predictive information. Similarly, the importance term is intended to encourage preference for physically informative wavelength regions, but an excessively large ω_3 may overemphasize locally large PLS coefficients and bias the search toward spectrally prominent yet not necessarily most predictive bands. Therefore, assigning small but nonzero weights to T_{spa} and T_{imp} allows them to gently guide the search toward more parsimonious and physically plausible solutions without derailing the primary optimization objective defined by T_{acc} . Since the relative contribution of each term may affect the search behavior, its impact is systematically assessed in the ablation analysis in Section 4.1.4.

Accuracy Term (T_{acc}): R^2 is the mean coefficient of determination from ten-fold cross-validation. Sparsity Term (T_{spa}): n is the number of selected bands and N is the total number of spectral variables. Significance Term (T_{imp}): $\mu(|\rho_{PLS}|)$ represents the mean of the absolute PLS coefficients for the selected subset, normalized by the global maximum absolute coefficient $\max(|\rho_{PLS}|)$ of the entire spectrum. This term leverages the global sensitivity information of the PLS model to guide the algorithm toward wavelength regions with high physicochemical relevance. $\varepsilon = 10^{-6}$ is a small positive constant to preclude division by zero.

$$f(X_i) = \begin{bmatrix} f(X_1) \\ f(X_2) \\ \vdots \\ f(X_N) \end{bmatrix} \quad (19)$$

2.3.3. Comparison of spectral feature selection methods

A comparative analysis was conducted on a diverse set of feature selection algorithms, specifically including Pearson correlation coefficient (PCC) (Rodgers and Nicewander, 1988), competitive adaptive reweighted sampling (CARS) (Li et al., 2009), variable importance in projection (VIP) (Teofilo et al., 2009), successive projections algorithm (SPA) (Araújo et al., 2001), genetic algorithm (GA) (Che et al., 2025), and improved weighting marine predators algorithm (IWMPA) (Tan et al., 2025), wave optics optimizer (WOO) (Peng et al., 2025), aquila optimizer (AO) (Abualigah et al., 2021), Newton's downhill optimizer (NDO) (Xiao et al., 2026), reinforced quantum aquila optimizer (RLQFAO) (Yang et al., 2026), binary adaptive weight dwarf mongoose optimization (BASADMO) (Lv et al., 2025) and grey wolf optimizer-

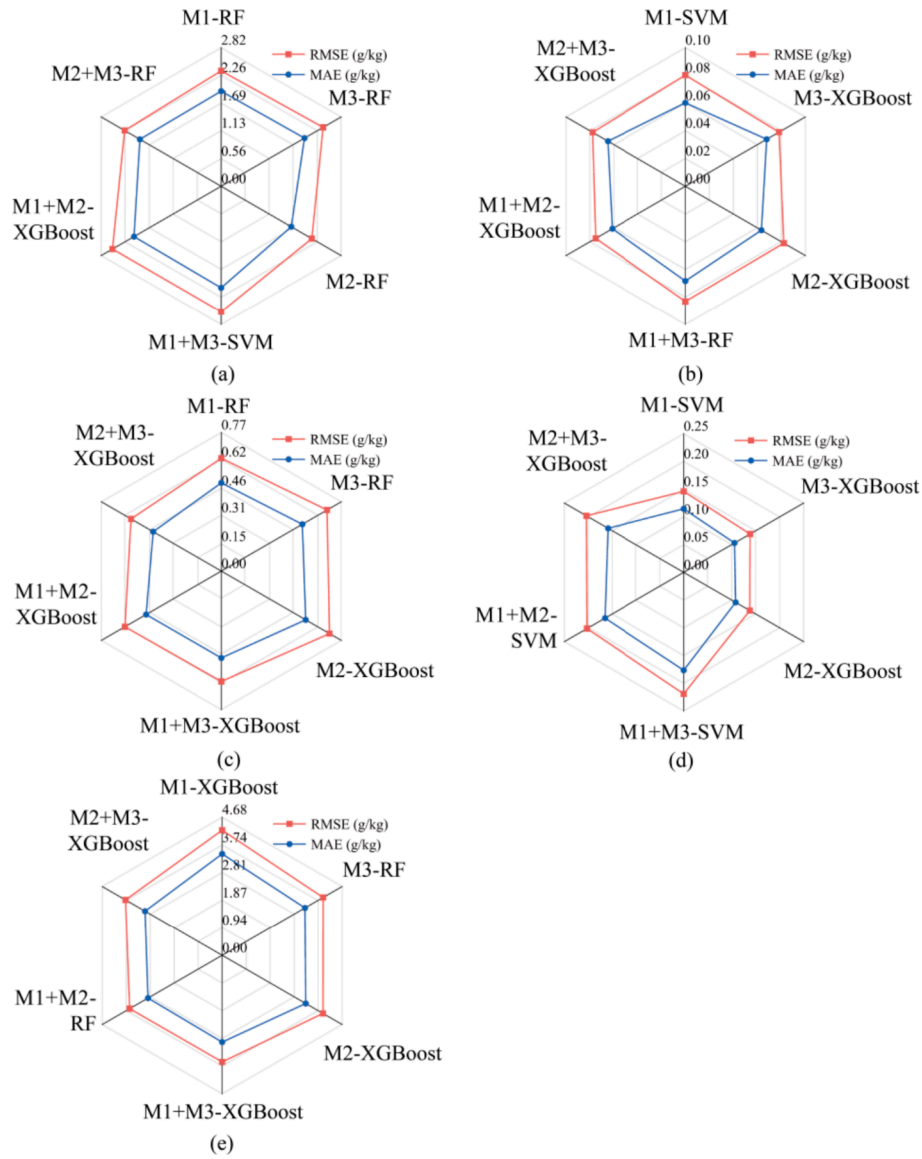


Fig. 13. Results of the ablation study for ISMA with different strategy combinations based on the optimal prediction model. (a) SOM. (b) TP. (c) TK. (d) TN. (e) SOM (HyMap).

Table 16
SOM estimation accuracies under alternative module execution sequences.

Method	Model	Train			Test		
		R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	R_p^2	RMSE _p (g/kg)	MAE _p (g/kg)
M132	PLSR	0.352	3.625	3.021	0.305	4.316	3.597
	XGBoost	0.721	2.327	1.939	0.651	2.872	2.393
	SVM	0.752	2.194	1.828	0.685	2.643	2.203
	RF	0.783	1.876	1.563	0.718	2.421	2.018
M213	PLSR	0.361	3.573	2.978	0.312	4.258	3.548
	XGBoost	0.711	2.368	1.973	0.646	2.913	2.428
	SVM	0.745	2.036	1.697	0.679	2.695	2.246
	RF	0.777	1.902	1.585	0.712	2.476	2.063
M321	PLSR	0.330	3.714	3.095	0.288	4.405	3.671
	XGBoost	0.703	2.401	2.001	0.638	3.015	2.513
	SVM	0.729	2.292	1.910	0.672	2.784	2.320
	RF	0.769	2.053	1.711	0.706	2.562	2.135

Table 17
TP estimation accuracies under alternative module execution sequences.

Method	Model	Train			Test		
		R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	R_p^2	RMSE _p (g/kg)	MAE _p (g/kg)
M132	PLSR	0.352	0.105	0.084	0.305	0.112	0.091
	XGBoost	0.724	0.073	0.060	0.551	0.092	0.075
	SVM	0.768	0.066	0.054	0.583	0.082	0.067
	RF	0.815	0.056	0.046	0.622	0.071	0.059
M213	PLSR	0.361	0.101	0.082	0.312	0.109	0.089
	XGBoost	0.719	0.074	0.060	0.548	0.093	0.076
	SVM	0.773	0.065	0.053	0.577	0.084	0.069
	RF	0.802	0.058	0.048	0.613	0.074	0.061
M321	PLSR	0.330	0.108	0.087	0.288	0.117	0.095
	XGBoost	0.709	0.076	0.062	0.539	0.094	0.077
	SVM	0.766	0.067	0.055	0.571	0.086	0.071
	RF	0.797	0.059	0.049	0.605	0.076	0.063

Table 18

TK estimation accuracies under alternative module execution sequences.

Method	Model	Train			Test		
		R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	R_p^2	RMSE _p (g/kg)	MAE _p (g/kg)
M132	PLSR	0.237	1.113	0.852	0.150	1.194	0.985
	XGBoost	0.540	0.810	0.607	0.357	0.701	0.583
	SVM	0.766	0.633	0.525	0.565	0.591	0.611
	RF	0.752	0.642	0.539	0.426	0.667	0.533
M213	PLSR	0.231	1.186	0.953	0.150	1.374	1.160
	XGBoost	0.636	0.730	0.571	0.459	0.849	0.676
	SVM	0.770	0.580	0.479	0.316	0.970	0.809
	RF	0.527	0.908	0.761	0.350	0.887	0.760
M321	PLSR	0.218	1.162	1.089	0.150	0.979	0.876
	XGBoost	0.632	0.790	0.662	0.419	0.682	0.651
	SVM	0.723	0.677	0.539	0.293	0.911	0.801
	RF	0.644	0.701	0.593	0.335	0.866	0.730

Table 19

TN estimation accuracies under alternative module execution sequences.

Method	Model	Train			Test		
		R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	R_p^2	RMSE _p (g/kg)	MAE _p (g/kg)
M132	PLSR	0.579	0.236	0.164	0.307	0.181	0.172
	XGBoost	0.725	0.173	0.098	0.602	0.167	0.157
	SVM	0.675	0.206	0.152	0.631	0.133	0.150
	RF	0.825	0.139	0.066	0.654	0.125	0.125
M213	PLSR	0.493	0.259	0.198	0.421	0.167	0.168
	XGBoost	0.507	0.256	0.168	0.606	0.157	0.123
	SVM	0.658	0.213	0.157	0.617	0.132	0.147
	RF	0.782	0.173	0.132	0.633	0.133	0.151
M321	PLSR	0.441	0.272	0.217	0.321	0.177	0.170
	XGBoost	0.612	0.230	0.163	0.584	0.141	0.156
	SVM	0.515	0.254	0.182	0.602	0.131	0.149
	RF	0.704	0.198	0.141	0.665	0.125	0.121

Table 20

SOM estimation accuracies based on HyMap airborne hyperspectral imagery under alternative module execution sequences.

Method	Model	Train			Test		
		R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	R_p^2	RMSE _p (g/kg)	MAE _p (g/kg)
M132	PLSR	0.287	5.659	4.460	0.164	5.882	4.989
	XGBoost	0.611	3.928	2.883	0.482	4.322	3.584
	SVM	0.536	4.461	3.578	0.364	4.549	3.664
	RF	0.542	4.418	3.328	0.368	4.673	3.920
M213	PLSR	0.267	5.554	4.064	0.192	5.670	4.292
	XGBoost	0.621	3.807	3.130	0.441	4.448	3.474
	SVM	0.441	4.889	4.114	0.350	5.337	4.369
	RF	0.660	3.419	2.783	0.476	4.581	3.806
M321	PLSR	0.348	5.409	4.638	0.215	5.772	4.544
	XGBoost	0.657	3.638	3.119	0.508	4.474	3.357
	SVM	0.458	4.508	3.778	0.331	5.167	4.099
	RF	0.568	3.905	3.231	0.374	5.150	4.149

genetic algorithm (GWOGA) (Shang et al., 2025). The different selection methods are listed in Table 6. Multiple experiments demonstrated that the optimal hyperparameters vary across the soil fertility indicators within ISMA, as shown in Table 7.

2.3.4. Modeling method

In soil property inversion modeling, statistical approaches based on multivariate relationships, such as PLSR (Li et al., 2024), are widely used due to their proficiency to establish correlations between soil attributes and spectral signatures. With the rapid advancement of machine learning, including support vector machine (SVM) (Liu et al., 2025),

random forest (RF) (Wang et al., 2024), and extreme gradient boosting (XGBoost) (Aksoy et al., 2024), all of which exhibit strong capabilities in nonlinear modeling.

2.3.5. Evaluation methods

To validate the effectiveness of ISMA for soil property inversion, we adopted the evaluation criteria outlined in Eqs. (20)–(22), R^2 , the root-mean-square error (RMSE), the mean absolute error (MAE) are introduced in this paper (Mena et al., 2025).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (20)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}} \quad (21)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}|}{n} \quad (22)$$

2.3.6. Statistical significance test

To determine whether the performance improvements of ISMA over baseline and state-of-the-art algorithms are statistically significant, the Wilcoxon Signed-Rank Test was conducted (San, 2025). As a non-parametric statistical test, it does not require the assumption of normality. In this study, a p-value of less than 0.05 was utilized as the threshold for statistical significance, thereby providing an objective validation of the effectiveness of the proposed improved mechanisms.

3. Results

3.1. Spectral transformations and correlations

Fig. 5 illustrates the spectral reflectance characteristics corresponding to the selected sampling locations for both the HyMap and ZY1-02E images. The results of the correlations of OR and RI with soil fertility indicators are listed in Table 8. The constituent bands of RI are the original reflectance that exhibit statistically significant ($p < 0.05$) correlations with the soil fertility indicators. RI was selectively incorporated based on its correlation enhancement relative to OR. For ZY1-02E data, it was used only for TP and TK, but for HyMap imagery, its use was extended to SOM, for which it also showed a notable benefit. The numbers of spectral features input to the ISMA from the ZY1-02E satellite data were 500, 370, 418, and 413 for TP, TK, TN, and SOM, respectively; for the HyMap airborne imagery, 529 spectral features were used for SOM estimation.

3.2. Spectral feature selection for SOM based on ZY1-02E satellite image dataset

Fig. 6 presents the spectral feature selection results for SOM based on ZY1-02E satellite image dataset. Compared to the five conventional feature selection algorithms and the original SMA, ISMA selects a relatively reduced number of spectral features (57 features). This demonstrates its efficacy in hyperspectral dimensionality reduction. Although SPA selects the minimum features, its inversion accuracy, as presented in Fig. 7, is suboptimal. This suggests that the strategy of using orthogonal projection to eliminate multicollinearity has limitations in capturing the essential spectral features for soil fertility indicators. The efficacy of the ISMA method is further substantiated by the physical and chemical significance of its selected features. Specifically, features around 2200 nm can be primarily attributed to vibrations of O–H functional groups (Wang et al., 2020).

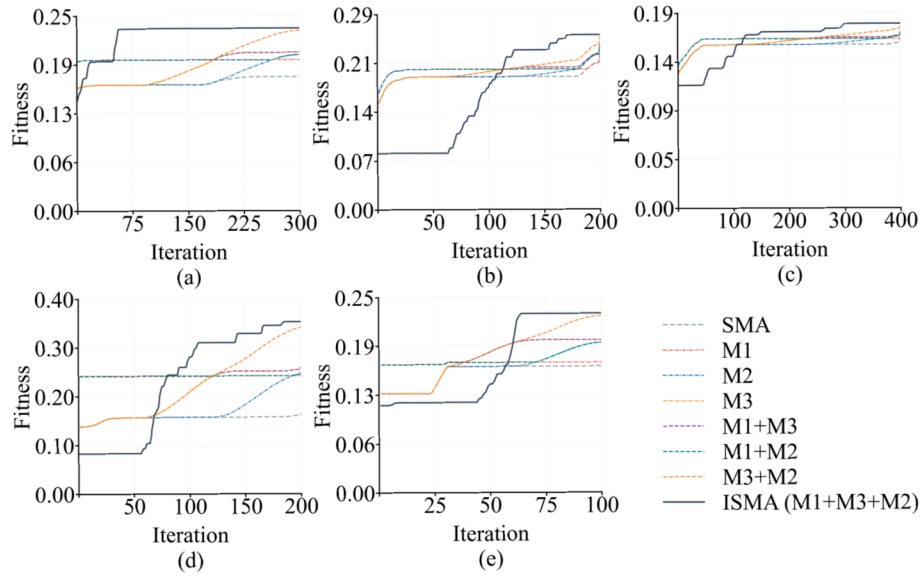


Fig. 14. Fitness convergence curves of all module combinations under the three-stage ablation study. (a) SOM. (b) TP. (c) TK. (d) TN. (e) SOM (HyMap).

Table 21
Statistical comparison of ISMA and SMA (10 independent runs).³

Dataset	ISMA (Mean $\pm$ SD)	SMA (Mean $\pm$ SD)	p-value
SOM	0.2393 $\pm$ 0.0089	0.2031 $\pm$ 0.0137	0.002**
TN	0.2992 $\pm$ 0.0200	0.2555 $\pm$ 0.0145	0.002**
TK	0.1757 $\pm$ 0.0139	0.1595 $\pm$ 0.0107	0.037*
TP	0.2499 $\pm$ 0.0132	0.2368 $\pm$ 0.0089	0.049*
SOM (HyMap)	0.2243 $\pm$ 0.0154	0.2049 $\pm$ 0.0129	0.014*

³ In Table 21, ** $p < 0.01$, * $p < 0.05$.

3.3. Accuracy assessment and spatial distribution of SOM based on ZY1-02E satellite image dataset

Based on the ZY1-02E dataset, we performed a random 3:1 partition to generate training and testing subsets. For SOM prediction using ZY1-02E satellite imagery (Fig. 7), ISMA achieved the highest performance when implemented within the XGBoost framework ($R^2 = 0.752$). These results demonstrated notable improvements over the comparable meta-heuristic algorithms such as GA and IWMPPA for soil property estimation. Additionally, when the 92-feature subset selected by VIP was utilized, the RF model's R^2 value declined to 0.721. This performance is likely attributable to the linear assumptions of VIP being ill-suited for

capturing the sensitive features within the non-linear spectral response of SOM. Among the evaluated algorithms, WOO, AO, NDO, and RLQFAO achieved competitive R^2 values for SOM retrieval (0.716, 0.703, 0.711, and 0.691, respectively). While outperforming SMA ($R^2 = 0.597$), they fell short of ISMA ($R^2 = 0.752$) due to their specific feature selection limitations (Fig. 6). Although WOO and AO primarily emphasized the vis-NIR region (500–900 nm), they underrepresented the 1800–2400 nm range, thereby overlooking critical C–H and O–H absorption features. Similarly, NDO clustered its 78 selected features mainly in the NIR region, lacking complementary SWIR wavelengths. Conversely, RLQFAO retained only 37 features, and this over-compression discarded informative features, leading to its relatively lower accuracy. These results indicated that effective retrieval requires a careful balance between dimensionality reduction and the retention of physically relevant spectral features.

The optimally trained model was applied to ZY1-02E imagery to map the spatial distribution of SOM content across western Jilin Province (Fig. 8). The majority of the region falls under Grade 4 and Grade 5 according to China's Second National Soil Survey (Table 9). Low SOM content is concentrated in lacustrine areas near Dabusu Lake (A), Hua'ao Pao (B), Zhangjia Pao (F), and Shuangliao City Baihe Nature Reserve (E). This is likely attributable to lower biological activity and reduced organic matter accumulation in these relatively enclosed aquatic and

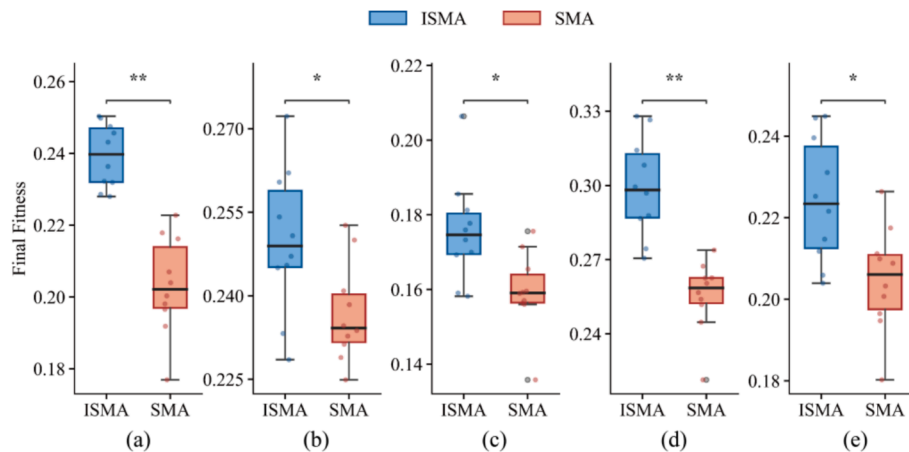


Fig. 15. Box plots of final fitness values from 10 independent runs. (a) SOM. (b) TP. (c) TK. (d) TN. (e) SOM (HyMap).

Table 22
Statistical comparison for the competing algorithms over 10 independent runs.

Dataset	RLQFAO		NDO		AO		WOO		BASADMO		GWOGA	
	Mean± SD	p-value	Mean± SD	p-value	Mean± SD	p-value	Mean± SD	p-value	Mean± SD	p-value	Mean± SD	p-value
SOM	0.2221± 0.0211	0.0322*	0.2124± 0.0310	0.0186*	0.2075± 0.0367	0.0137*	0.2174± 0.0260	0.0244*	0.1985± 0.0285	0.0125*	0.1820± 0.0332	0.0102*
TN	0.2823± 0.0296	0.0420*	0.2656± 0.0383	0.0186*	0.2575± 0.0320	0.0106*	0.2737± 0.0297	0.0322*	0.2450± 0.0312	0.0102*	0.2410± 0.0354	0.0101*
TK	0.1677± 0.0145	0.0420*	0.1625± 0.0221	0.0403*	0.1599± 0.0287	0.0389*	0.1651± 0.0186	0.0410*	0.1520± 0.0205	0.0210*	0.1440± 0.0241	0.0165*
TP	0.2305± 0.0221	0.0244*	0.2251± 0.0279	0.0137*	0.2216± 0.0220	0.0118*	0.2283± 0.0247	0.0186*	0.2110± 0.0238	0.0108*	0.1990± 0.0276	0.0102*
SOM (HyMap)	0.2143± 0.0181	0.0420*	0.2058± 0.0217	0.0244*	0.2010± 0.0264	0.0186*	0.2087± 0.0190	0.0322*	0.1920± 0.0224	0.0135*	0.1780± 0.0268	0.0105*

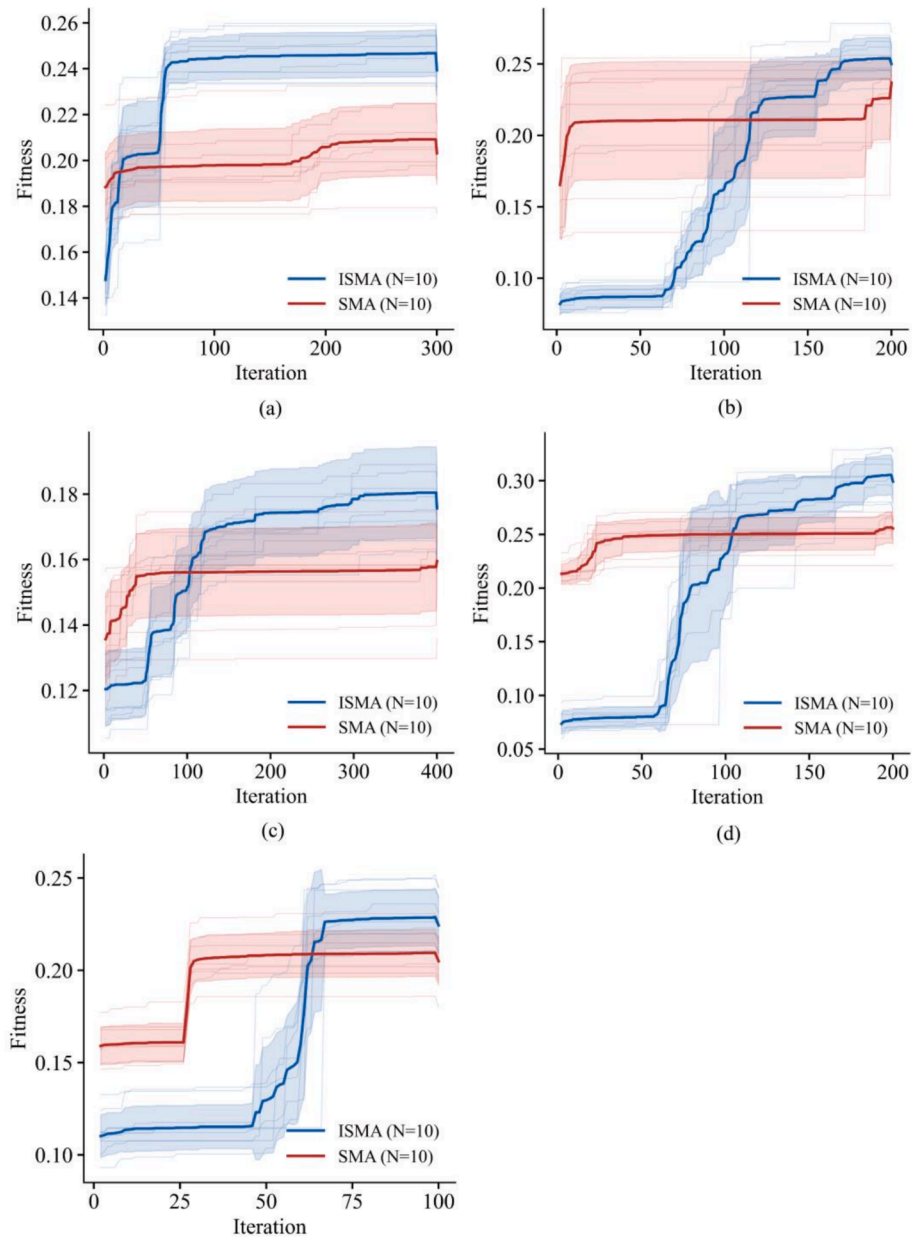


Fig. 16. Convergence curves of ISMA and SMA. Solid lines represent the mean fitness over 10 independent runs; shaded bands indicate ± 1 standard deviation. ISMA consistently converges to a higher final fitness with comparable or lower variance. (a) SOM. (b) TP. (c) TK. (d) TN. (e) SOM (HyMap).

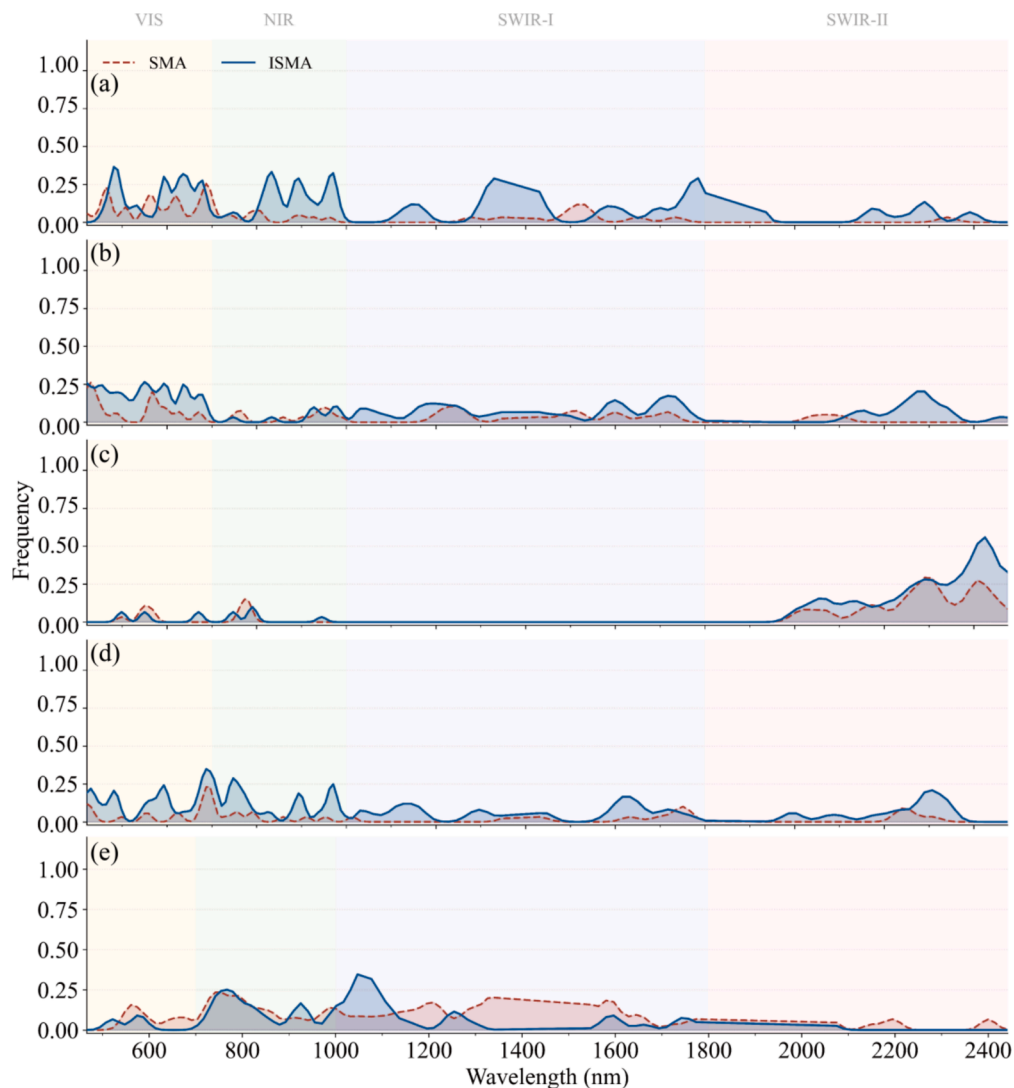


Fig. 17. Band selection frequency profiles of ISMA (blue, solid fill) and SMA (red dashed) across 10 independent runs. (a) SOM. (b) TP. (c) TK. (d) TN. (e) SOM (HyMap). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

catchment environments. Conversely, high SOM content is primarily observed in western Yong'an Township, northern Sanqingshan Town (D), and the south of the Chagan Lake scenic area (C). This high concentration is probably due to artificial factors, such as the implementation of scientific cultivation and utilization of local saline-alkali land.

4. Discussion

4.1. Application capability analysis of ISMA

4.1.1. Multi-indicator applicability analysis

In this study, ISMA was initially applied to estimate SOM. Following this primary validation, its application was extended to other soil fertility indicators to evaluate its performance using ZY1-02E satellite imagery, as shown in Fig. 9. An analysis of the feature selection resulting from the different spectral transformations indicated that, since the visible light region can reflect soil color and iron oxide absorption (Camargo et al., 2015; De Souza Bahia et al., 2017), the feature bands for TP in the 600–800 nm range selected by ISMA may exhibit weak characteristic signals through indirect associations with organic matter and iron oxides. ISMA selected feature bands for TK predominantly concentrated in the SWIR range. The bands were essentially consistent

with larger regression coefficients mentioned in the literature (Aksoy et al., 2024), including 450, 680, 960, 1410, 2370 nm, and their nearby bands. This consistency may be attributed to the incorporation of PLS-based weighting in ISMA. For TN estimation, the algorithm identified important bands within the NIR region, which were likely associated with overtones and combination bands related to soil constituents, as well as to the characteristic absorptions of N–H, O–H, and C–H groups concentrated within this spectral range (Thissen et al., 2004). In addition, Fig. 9 further shows substantial differences among feature selection methods in the number of selected features for different soil fertility indicators, highlighting distinct strategies for feature compression and information retention. Overall, conventional methods such as GA, WOO, and AO tended to retain more features, whereas ISMA produced a more compact feature subset for TP, TK, TN, and SOM. This suggests that ISMA can effectively reduce spectral redundancy while preserving informative features closely related to the target variables.

The estimation accuracies for the soil fertility indicator content using ISMA and SMA are listed in Table 10. The results demonstrate that ISMA, when integrated with XGBoost and RF, achieves superior predictive performance to SMA for TP, TK, and TN, with accuracies of 0.675, 0.670, and 0.678, respectively. The results highlight that the XGBoost model, by leveraging its gradient boosting mechanism, can more effectively capture the complex non-linear relationships within the spectral data,

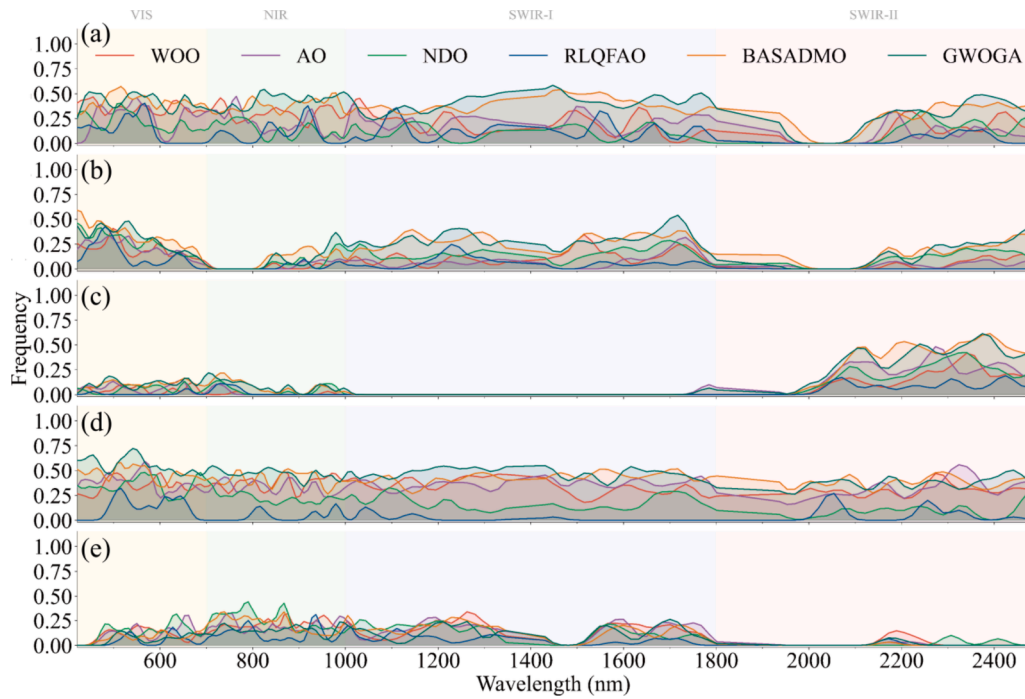


Fig. 18. Band selection frequency profiles of WOO, AO, NDO, RLQFAO, BASADMO and GWOGA across 10 independent runs. (a) SOM. (b) TP. (c) TK. (d) TN. (e) SOM (HyMap).

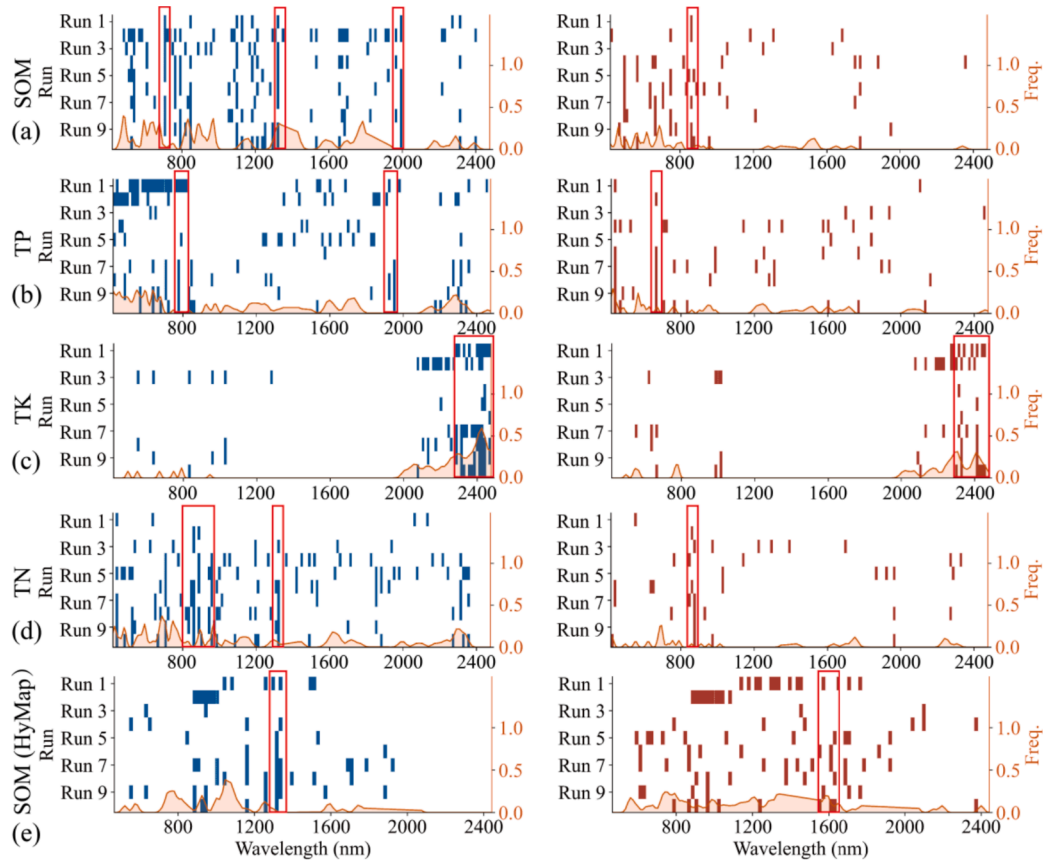


Fig. 19. Run-by-run band selection maps for ISMA (blue, left) and SMA (red, right) across 10 runs. Each dot indicates selection of the corresponding wavelength in that run. Orange frequency curve (right axis) overlays the per-band selection frequency. (a) SOM. (b) TP. (c) TK. (d) TN. (e) SOM (HyMap). (Red boxes indicate wavelength regions with high cross-run selection consistency.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

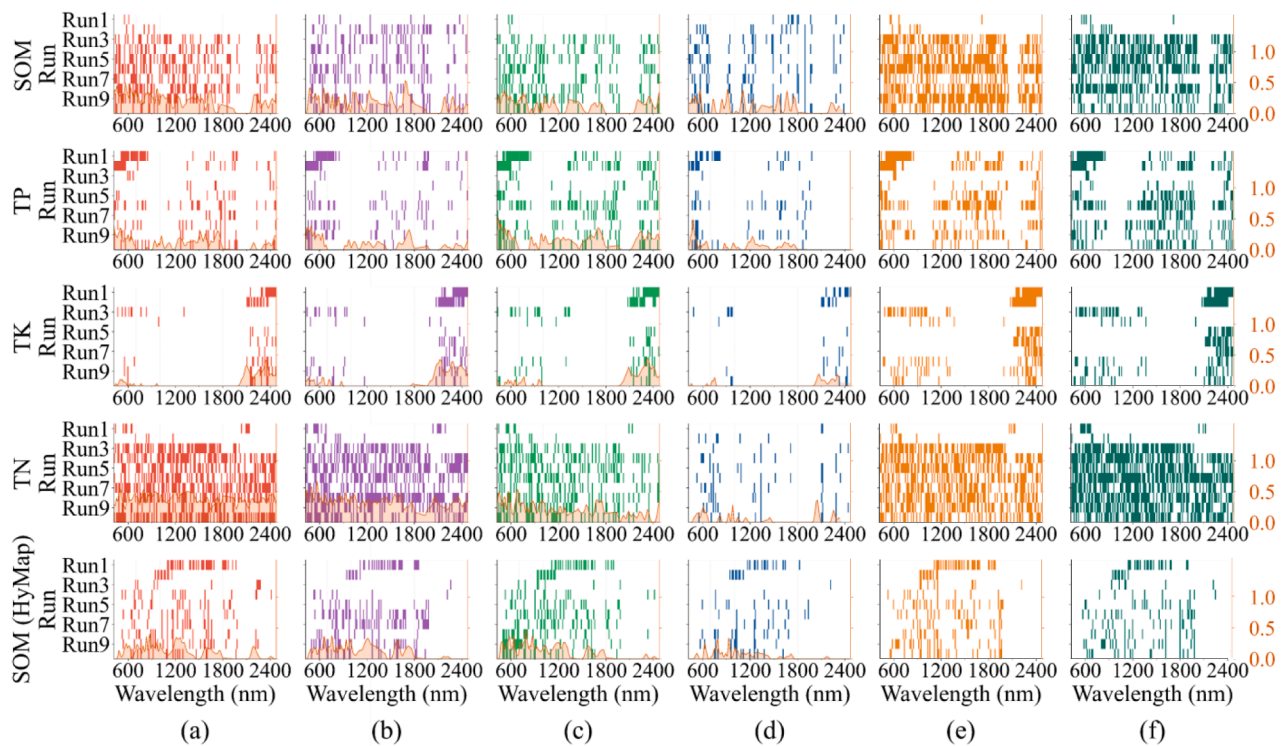


Fig. 20. Run-by-run band selection maps for SOM, TP, TK, TN and SOM (HyMap) across 10 runs. Each dot indicates selection of the corresponding wavelength in that run. Orange frequency curve (right axis) overlays the per-band selection frequency. (a) WOO. (b) AO. (c) NDO. (d) RLQFAO. (e) BASADMO. (f) GWOGA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

compared to the other models.

4.1.2. Cross-platform applicability analysis

To evaluate ISMA's applicability across remote sensing platforms—from satellite to airborne—an additional validation experiment was conducted. This experiment utilized an independent HyMap airborne hyperspectral dataset from 2017, partitioned at a 3:1 ratio for SOM estimation. Fig. 10 presents both the wavelength positions and the numbers of features selected by different methods for the HyMap imagery. In each subfigure, the left panel shows the spectral locations of the selected bands (x-axis: wavelength; y-axis: feature selection methods), while the right panel shows the corresponding number of selected features. Notably, SPA selected only three features, suggesting that its rigid sequential search struggled to find effective predictors within the HyMap dataset. This outcome highlights the advantage of adaptive optimization over fixed statistical rules, especially when seeking feature selection across different sensors. Furthermore, as shown in Table 11, the model identified several common spectral bands for SOM estimation across both HyMap and ZY1-02E hyperspectral imagery, notably at 754–756 nm, 816–817 nm, and 1741–1750 nm further validating the algorithm's effectiveness. The 816–817 nm band lies within the 800–1000 nm NIR interval, where organic matter can influence iron-oxide-related spectral responses and where soil structural properties contribute to light scattering (Heller Pearlshtien and Ben-Dor, 2020). The 1741–1750 nm range falls within a C–H overtone-related region in the SWIR-I zone (Stenberg et al., 2010). These common bands suggest that ISMA can identify physically meaningful spectral regions under different observation conditions. Although the difference in spatial resolution between the two sensors (30 m vs. 4.5 m) introduces variation in pixel mixing degree, our imagery was acquired over large freshly ploughed bare soil fields, ensuring the radiometric signals of both sensors are dominated by the soil surface and minimizing mixed-pixel interference on feature selection. The consistent features identified are not artifacts of overlapping spectral ranges or local soil types,

but correspond to meaningful absorption features of key soil constituents. Table 12 demonstrates that the SOM estimation model constructed with ISMA-selected feature bands (22 features) achieved substantial accuracy gains on the HyMap airborne hyperspectral imagery ($R^2 = 0.757$, RMSE = 3.056 g/kg, MAE = 2.485 g/kg).

4.1.3. Parameter settings analysis

The weighting schemes of the fitness function for different configurations are summarized in Table 13. Fig. 11 (a) shows that each dataset reaches peak R^2 at a specific iteration count, after which accuracy either fluctuates within a narrow range or declines. Meanwhile, the number of selected features stabilizes near its minimum at the optimal iteration, serving as a reliable convergence indicator. During the early phase, the algorithm explores the search space, improving accuracy and reducing the feature count as irrelevant features are eliminated. In the later stage, the algorithm mainly focuses on refining the current solutions, which causes small fluctuations in performance. Crucially, further iterations yield diminishing returns as accuracy does not consistently improve. Fig. 11 (b) demonstrates a consistent unimodal pattern across all indicators. Accuracy rises from $N = 10$ to $N = 30$ and subsequently plateaus or slightly declines for $N > 30$. When $N < 30$, the population lacks sufficient diversity to escape local optima, resulting in premature convergence and higher feature counts. Conversely, when $N > 30$, additional individuals improve neither prediction accuracy nor feature compactness. Fig. 11 (c) reveals that accuracy is non-monotonic with respect to the threshold, peaking between 0.50 and 0.60 for all indicators. The optimal values are 0.55 for SOM, 0.50 for TP and TN, and 0.60 for TK, suggesting that the optimal threshold depends on soil composition and sensor characteristics. At low thresholds in Eq. (16), too few features are retained, and spectrally important wavelengths are discarded, causing a sharp drop in accuracy. At high thresholds, the relaxed selection criterion leads to feature over-selection, allowing spectral noise and correlated features into the model, thereby inflating feature counts and degrading performance through multicollinearity. As

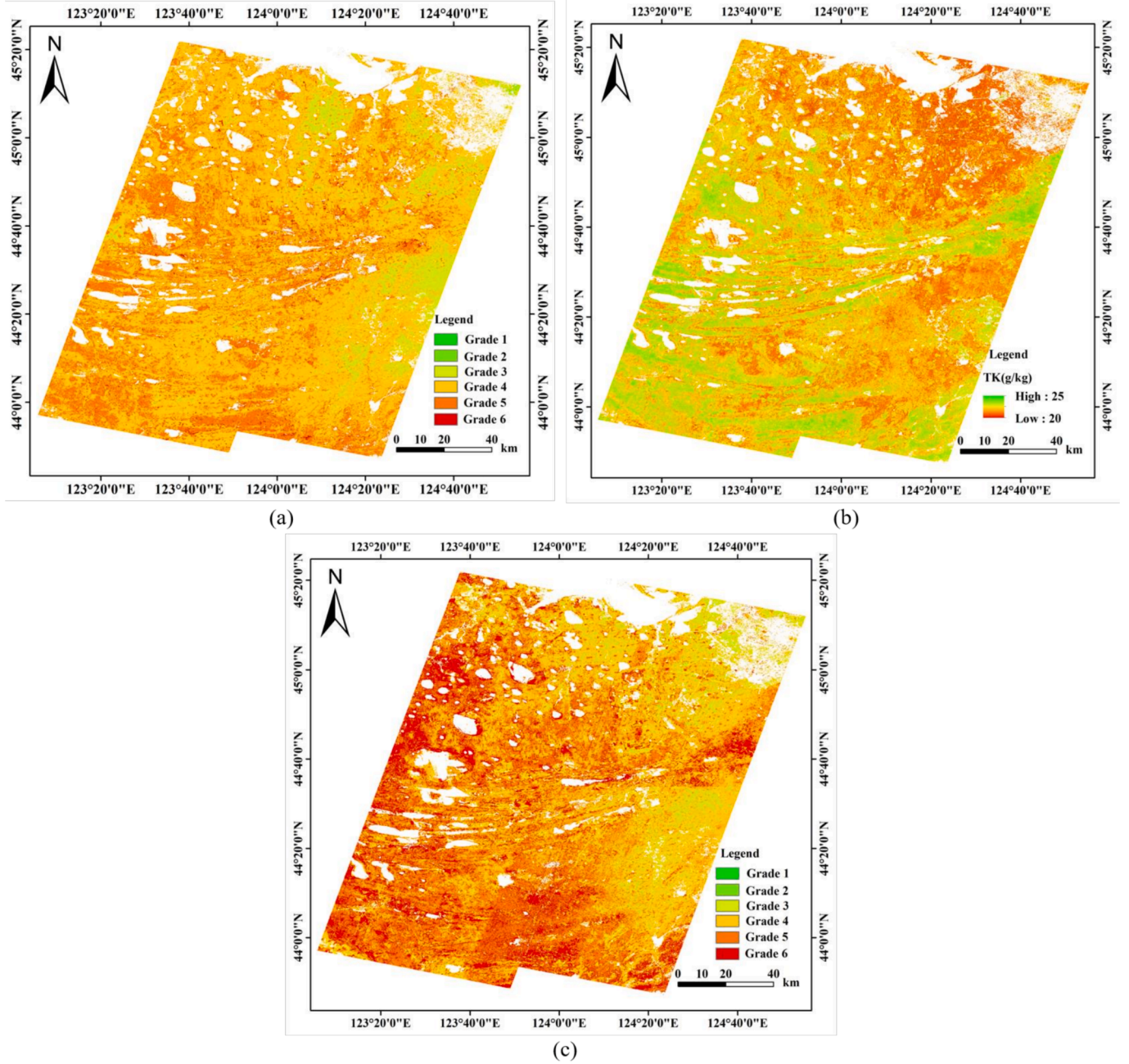


Fig. 21. Spatial distribution for TP, TK and TN in western Jilin Province based on ZY1-02E satellite image dataset. (a) TP. (b) TK. (c) TN.

shown in Fig. 11 (d), the choice of fitness weights is a key factor affecting both accuracy and feature compactness. The proposed configuration (E1) consistently achieves the highest R^2 across all indicators while maintaining a compact subset of 22 to 69 features, compared to 88 to 183 features for ablated variants. Removing sparsity regularization increases the feature count by 50% to 160% with negligible or negative accuracy gains. Finally, removing the PLS weight evaluation term degrades relative accuracy by 3% to 22%. This confirms that rewarding subsets with higher average absolute coefficients, relative to the global spectral baseline, successfully guides the search toward wavelength regions with stronger physicochemical relevance. To examine the sensitivity of model performance to the accuracy weight w_1 , a two-stage analysis was conducted. A coarse initial scan (Fig. 11 (d)) revealed a notable improvement in R^2 when $w_1 \geq 0.7$, motivating a fine-grained search over the interval $w_1 \in [0.70, 0.90]$ with a step size of 0.02. As shown in Fig. 12, the R^2 values of all five datasets remain highly stable

across this range, with a maximum coefficient of variation (CV) of less than 1.5%. Notably, $w_1 = 0.80$ lies within a stable plateau region ($w_1 \in [0.76, 0.84]$) rather than at an isolated optimum, confirming that the selected weight is not coincidental.

To evaluate whether consolidating the feature selection and regression steps under a unified nonlinear wrapper framework could yield a superior feature subset, an alternative variant (I_{XGB}) was implemented by replacing the ρ_{PLS} term in the original fitness function (Eq. (17)) with a nonlinear feature importance metric derived from XGBoost. This metric represents the relative contribution assigned to each spectral feature during the tree-building process of the XGBoost algorithm. As shown in Table 14, the nonlinear integration consistently yielded inferior retrieval accuracy compared to the proposed ISMA across all evaluated indicators. This performance deterioration is fundamentally driven by two factors. First, given the extreme dimensionality and limited sample size of soil hyperspectral datasets, placing a high-

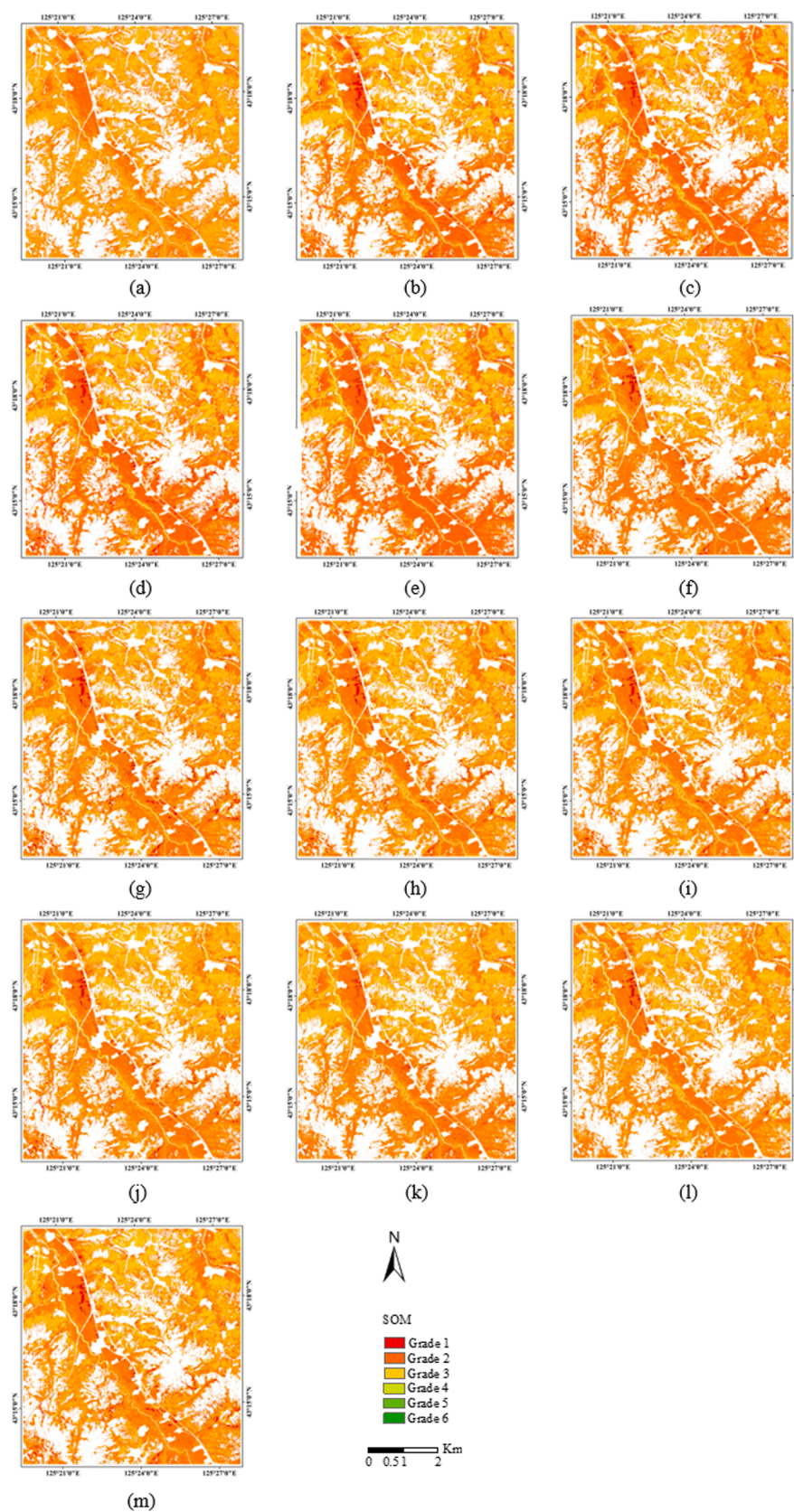


Fig. 22. Spatial distribution of SOM in Yitong Manchu Autonomous County based on HyMap airborne hyperspectral image dataset. (a) SPA-XGBoost. (b) VIP-XGBoost. (c) GA-RF. (d) CARS-RF. (e) IWMPA-RF. (f) WOO-XGBoost. (g) AO-XGBoost. (h) RLQFAO-XGBoost. (i) NDO-RF. (j) BASADMO-RF. (k) GWOGA-RF. (l) SMA-RF. (m) ISMA-XGBoost.

capacity model like XGBoost inside the inner optimization loop severely exacerbates overfitting. Second, by compressing these correlated inputs into orthogonal latent components, PLSR produces stable fitness scores across iterations, enabling the population to build reliable search trajectories. XGBoost, by contrast, assigns feature importance through a greedy tree-splitting process that is sensitive to small sample fluctuations, preventing the algorithm from consistently reinforcing promising spectral regions.

4.1.4. Ablation study of ISMA

To clarify the contribution of each proposed design, a factorial ablation study was conducted by decomposing ISMA into three functional modules, as shown in Table 15. The ablation study demonstrated that M1, M2, and M3 play distinct but complementary roles, as summarized in Fig. 13. For TP, M1 + M2 achieved the best test performance with XGBoost ($R^2 = 0.6044$, RMSE = 0.0752, MAE = 0.0613), suggesting a strong synergy between the initialization strategy and pre-screening. For TK, variants containing M2 + M3 were generally more competitive, and M1 + M2 and M1 + M3 yielded the strongest results. For TN, M2 and other M2-related variants performed particularly well, showing that pre-screening is highly effective when redundant or weakly informative features strongly interfere with model learning. For SOM based on the ZY1-02E dataset, M2 yielded the lowest prediction error, highlighting the complementary effect of pre-screening and iterative optimization. For SOM on the HyMap airborne hyperspectral imagery, M1 + M3 produced the most favorable results, suggesting that improved initialization and search dynamics jointly enhance model stability. Overall, M1 improved the initial population quality, M3 strengthened the iterative exploration-exploitation balance, and M2 reduced the dimensional burden and suppresses irrelevant information. Their interaction explained the performance advantage of the complete ISMA framework.

To validate whether the three stages possess a strict hierarchical dependency, rather than operating as independent add-ons, we conducted a sequence inversion experiment. As shown in Tables 16–20, any violation of the prescribed M1 → M2 → M3 order leads to measurable degradation in retrieval accuracy across all fertility indicators. The resulting convergence curves (Fig. 14) reveal structurally distinct profiles for each combination, demonstrating that no partial combination replicates the full system's behavior. In the early iterations, variants incorporating M1 begin from a visibly higher fitness level than all M1-absent configurations, reflecting the role of Chebyshev-based spectral initialization in biasing the starting population toward informative features through PLS-Lasso integration. In the mid-stage iterations, M2-containing variants exhibit a pronounced upward bend in slope that is absent in M2-absent curves, which flatten into a near-plateau. This divergence arises because the PLS-based stochastic weight mechanism amplifies discriminative features while suppressing redundant ones, thereby sustaining meaningful population search progress across the high-redundancy spectral regions where convergence would otherwise stall. In the late iterations, M1 + M2 curves level off into a characteristic plateau while ISMA continues to rise. That is a direct consequence of the absence of M3's multi-level position update, whose Lévy-perturbed and OBL-diversified search modes prevent the population from converging repeatedly to the same locally optimal feature subsets. The performance advantage of ISMA stems from the coordinated interaction among all three stages rather than the independent contribution of each.

4.2. Comparison of ISMA and SMA

4.2.1. Statistical comparison analysis

We conducted 10 independent runs for both ISMA and the baseline SMA. Each run used a different random seed to ensure genuine independence. For each algorithm and soil fertility indicator, we reported the mean and standard deviation (SD) of the final fitness value across all 10 runs (Table 21). Additionally, box plots illustrate the distribution of

final fitness values and include significance annotations (Fig. 15). RLQFAO, NDO, WOO, and AO were each independently run 10 times. Table 22 reports the statistical comparison from the Wilcoxon signed-rank tests for these competing algorithms. Furthermore, we provided convergence plots displaying the mean $\pm$ 1 SD band across 10 runs for both ISMA and SMA (Fig. 16), clearly visualizing consistency and stability. We applied the two-sided Wilcoxon signed-rank test ($\alpha = 0.05$) to the paired 10-run results for each dataset. The statistical test results are reported in Table 21. ISMA achieved statistically significant superiority over SMA on all indicators ($p < 0.05$).

4.2.2. Spectral feature selection analysis and algorithm stability

As shown in Fig. 17 and Fig. 19, ISMA consistently yields higher and more structured frequency peaks than SMA in physically meaningful spectral regions. Fig. 18 and Fig. 20 present the feature selection results of RLQFAO, NDO, WOO, and AO across 10 independent runs. For SOM, ISMA preferentially activates bands in the vis-NIR transition zone (550–900 nm), particularly near 670 nm and 720 nm. Secondary peaks appear in 1900–2200 nm, consistent with C–H and O–H overtone absorptions linked to humic substances. SMA produces diffuse, low-frequency activation across the full spectrum without these physically interpretable clusters. For TK, selection is concentrated almost entirely in the 1850–2400 nm range, with ISMA achieving frequency peaks of approximately 0.30 near 2000 nm and 2200 nm. This pattern aligns with the spectral response of clay minerals and feldspars that co-vary with potassium availability in soils (Galvao et al., 1997). For TN, ISMA shows moderate, distributed selection across 540–780 nm and 1400–1600 nm, with the frequency reaching 0.25 near 700 nm. For TP, selection frequencies remain relatively low (≤ 0.25) but consistently higher for ISMA than SMA across both 500–800 nm and 2000–2400 nm. The relatively low selection frequency magnitude is consistent with phosphorus having a weak direct spectral response, primarily detectable through its association with organic matter and iron oxides (Fink et al., 2016).

To visualize the within-algorithm reproducibility of band selection across independent stochastic runs, Fig. 19 presents run-by-run scatter maps for ISMA and SMA across all soil parameters. ISMA demonstrates superior band selection stability relative to SMA, as evidenced by (i) higher and physically interpretable frequency peaks; (ii) greater run-to-run coherence in band selection maps. ISMA exhibits visibly denser vertical alignment of dots compared to SMA, indicating that the same wavelength regions are reliably selected across different runs. This run-to-run coherence is especially pronounced in the vis-NIR region for SOM and TN, and in the SWIR for TK, consistent with the frequency peaks observed in Fig. 20. SMA selection maps are more horizontally scattered, with individual runs selecting different wavelength clusters. For SOM across both ZY1-02E and HyMap platforms, ISMA generates dense, reproducible selections in the 800–1300 nm and 1900–2100 nm regions, accurately capturing C–H overtones. Similarly, ISMA also captures the spectral features of TN within the 800–1000 nm and 1400–1600 nm regions. For TK, ISMA confines nearly all selections to the 1800–2400 nm range with high run-to-run overlap, whereas SMA activates sparse, scattered bands across the full spectrum with little cross-run consensus. For TP, ISMA maintains a discernible clustering preference near 600–750 nm and 2100–2300 nm that is absent or inconsistent in SMA.

4.2.3. Spatial distribution analysis of the soil fertility indicator content

Fig. 21 illustrates the spatial mapping for fertility indicators with ZY1-02E imagery. TP exhibits a west-to-east increasing gradient, with hotspots concentrated near Xinmiao and Chagan Lakes in Qian Gorlos. In contrast, TK levels are generally high (Grade 2), displaying a northeast-low to southwest-high pattern in the continuous mapping. The elevated TK in southwestern Qian'an is likely attributed to better drainage facilitated by the rising terrain, which minimizes leaching losses. Both TN and SOM share a similar west-low, east-high trend. Meanwhile, superior SOM levels in Qian Gorlos compared to Changling are driven by the prevalence of fertile chernozems and meadow soils.

As depicted in Fig. 22, the SOM content in Yitong Manchu Autonomous County is predominantly classified as Grade 2 or Grade 3. Regions exhibiting elevated SOM content are primarily aggregated along the Yitong River. This is likely because the alluvial soils on the riverbanks are in a phase of rapid accumulation, leading to higher SOM content. In contrast, mountainous areas tend to exhibit lower SOM. Factors such as higher elevations, abundant sunlight and well-drained conditions likely enhance microbial activity, thereby accelerating SOM decomposition.

It should be noted that the present spatial distribution analysis is based on two-dimensional hyperspectral imagery and therefore mainly reflects the horizontal variation in soil fertility indicators. Some of the mapped differences may also be associated with terrain-related conditions, such as drainage and surface heterogeneity, which cannot be fully characterized in a two-dimensional framework. In future work, hyperspectral LiDAR may provide a useful extension by integrating spectral reflectance with three-dimensional structural information, thereby supporting a more complete analysis of soil-property distribution under complex surface environments (Qian et al., 2021; Bai et al., 2026). A potential limitation of this study is the theoretical discrepancy between the heuristic selection metric and the nonlinear evaluation model. Although integrating the model's internal feature importance into the fitness function could theoretically achieve consistency between the search objectives and the model's learning mechanism, our preliminary tests indicated that such a strategy increased the risk of overfitting and reduced generalization stability. Future research will explore the integration of lightweight nonlinear constraints to further optimize the search-evaluation consistency. While the consistency of the algorithm across different spatial resolutions observed in this study is promising, a more robust validation would require multi-resolution imagery acquired over the same time period, so as to isolate the algorithm's performance from confounding spatial and temporal variability. Likewise, although the current evaluation covers four soil properties and two sensor platforms, further validation across more diverse soil types, sensors, and geographic regions would strengthen the generalizability of the findings. Accordingly, prospective research will design satellite-ground synchronization experiments to systematically investigate algorithmic consistency across multi-resolution imagery, and extend the evaluation to a broader range of environmental settings.

5. Conclusion

This study proposes ISMA for hyperspectral feature selection, validated on ZY1-02E satellite and HyMap airborne datasets. Instead of relying on isolated search enhancements, ISMA organizes the optimization process into three coordinated stages: spectrally-informed initialization and pre-screening for redundancy reduction, a multi-level position update for balancing exploration and exploitation, and a stagnation-avoidance strategy for improving late-stage search robustness. A customized fitness function was also designed to balance estimation accuracy with feature sparsity. Experimental results with XGBoost demonstrate ISMA's superior performance in retrieving TP, TK, TN, and SOM on ZY1-02E imagery, achieving accuracy of 0.675, 0.670, 0.678, and 0.752, respectively. Furthermore, its cross-platform robustness was confirmed on independent HyMap data ($R^2 = 0.757$ for SOM), providing a reliable framework for rapid soil quality assessment and sustainable land management.

CRedit authorship contribution statement

Lijun Tian: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chao Niu:** Validation, Supervision, Investigation, Formal analysis. **Xue Wang:** Supervision, Resources, Project administration, Conceptualization. **Yayu Yang:** Supervision, Resources, Conceptualization. **Kun Tan:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Xiaojun Xu:** Resources, Funding acquisition, Formal analysis. **Wen**

Zhang: Resources, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (No. 2024CSJZN1300), the Shanghai Municipal Education Commission Science and Technology Project (2024AI02002), the National Natural Science Foundation of China (No. 42171335), and the National Civil Aerospace Project of China (No. D040102).

Data availability

Data will be made available on request.

References

- Abasabadi, S., Nematzadeh, H., Motameni, H., Akbari, E., 2022. Hybrid feature selection based on SLI and genetic algorithm for microarray datasets. *J. Supercomput.* 78 (18), 19725–19753. <https://doi.org/10.1007/s11227-022-04650-w>.
- Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, A.A., Al-Qaness, M.A.A., Gandomi, A.H., 2021. Aquila Optimizer: a novel meta-heuristic optimization algorithm. *Comput. Ind. Eng.* 157. <https://doi.org/10.1016/j.cie.2021.107250>.
- Aksoy, S., Sertel, E., Roscher, R., Tanik, A., Hamzehpour, N., 2024. Assessment of soil salinity using explainable machine learning methods and Landsat 8 images. *Int. J. Appl. Earth Obs. Geoinf.* 130, 103879. <https://doi.org/10.1016/j.jag.2024.103879>.
- Amini, F., Hu, G., 2021. A two-layer feature selection method using Genetic algorithm and elastic net. *Expert Syst. Appl.* 166, 114072. <https://doi.org/10.1016/j.eswa.2020.114072>.
- Araújo, M.C.U., Saldanha, T.C.B., Galvão, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* 57 (2), 65–73. [https://doi.org/10.1016/S0169-7439\(01\)00119-8](https://doi.org/10.1016/S0169-7439(01)00119-8).
- Asadzadeh, S., De Souza Filho, C.R., 2016. A review on spectral processing methods for geological remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* 47, 69–90. <https://doi.org/10.1016/j.jag.2015.12.004>.
- Bai, J., Huang, H., He, B., Ma, L., Gao, S., Peng, S., Yang, X., Hao, Y., Dong, Y., Bi, K., 2026. The first large-footprint hyperspectral LiDAR model to attempt to reveal forest chlorophyll distribution heterogeneity from vertically layered spectral perspective utilizing 3D radiative transfer modeling. *Int. J. Appl. Earth Obs. Geoinf.* 146, 105060. <https://doi.org/10.1016/j.jag.2025.105060>.
- Bai, J., Niu, Z., Wang, L., 2024. A theoretical demonstration on the independence of distance and incidence angle effects for small-footprint hyperspectral LiDAR: basic physical concepts. *Remote Sens. Environ.* 315, 114452. <https://doi.org/10.1016/j.rse.2024.114452>.
- Camargo, L.A., Júnior, J.M., Barrón, V., Alleoni, L.R.F., Barbosa, R.S., Pereira, G.T., 2015. Mapping of clay, iron oxide and adsorbed phosphate in Oxisols using diffuse reflectance spectroscopy. *Geoderma* 251, 124–132. <https://doi.org/10.1016/j.geoderma.2015.03.027>.
- Che, J., Xia, W., Xu, Y., Hu, K., 2025. Multivariate wind speed forecasting with genetic algorithm-based feature selection and oppositional learning sparrow search. *Inf. Sci.* 695, 121736. <https://doi.org/10.1016/j.ins.2024.121736>.
- Maarouf, B.B., Rashid, T.A., Abdulla, J.M., Hassan, B.A., Alsadoon, A., Mohammadi, M., Khishe, M., Mirjalili, S., 2022. Current studies and applications of shuffled frog leaping algorithm: a review. *Arch. Comput. Meth. Eng.* 29, 3459–3474. <https://doi.org/10.1007/s11831-021-09707-2>.
- Dada, B.A., Nwulu, N.I., Olukanmi, S.O., 2025. Bayesian optimization with Optuna for enhanced soil nutrient prediction: a comparative study with genetic algorithm and particle swarm optimization. *Smart Agric. Technol.* 12, 101136. <https://doi.org/10.1016/j.jatech.2025.101136>.
- De Souza Bahia, A.S.R., Marques, J., La Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017. Prediction and mapping of soil attributes using diffuse reflectance spectroscopy and magnetic susceptibility. *Soil Sci. Soc. Am. J.* 81 (6), 1450–1462. <https://doi.org/10.2136/sssaj2017.06.0206>.
- Shu, B., Hu, G., Cheng, M., Zhang, C., Engineering, 2025. MSFPO: Multi-algorithm integrated particle swarm optimization with novel strategies for solving complex engineering design problems. *Comput. Methods Appl. Mech. Eng.* 437, 117791. <https://doi.org/10.1016/j.cma.2025.117791>.
- Esmaili, M., Abbasi-Moghadam, D., Sharifi, A., Tariq, A., Li, Q., 2023. Hyperspectral image band selection based on CNN embedded GA (CNNeGA). *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 16, 1927–1950. <https://doi.org/10.1109/jstars.2023.3242310>.

- Li, N., Zhang, Y., Shi, K., Zhang, Y., Sun, X., Wang, W., Qian, H., Yang, H., Niu, Y., 2023. Real-time and continuous tracking of total phosphorus using a ground-based hyperspectral proximal sensing system. *Remote Sens.* 15 (2), 507. <https://doi.org/10.3390/rs15020507>.
- Diaz-Gonzalez, F.A., Vuelvas, J., Correa, C.A., Vallejo, V.E., Patino, D., 2022. Machine learning and remote sensing techniques applied to estimate soil indicators-review. *Ecol. Ind.* 135, 108517. <https://doi.org/10.1016/j.ecolind.2021.108517>.
- Fink, J.R., Inda, A.V., Tiecher, T., Barrón, V., 2016. Iron oxides and organic matter on soil phosphorus availability. *Cienc. Agrotecnol.* 40, 369–379. <https://doi.org/10.1590/1413-70542016404023016>.
- Galvao, L.S., Vitorello, I., 1998. Variability of laboratory measured soil lines of soils from southeastern Brazil. *Remote Sens. Environ.* 63 (2), 166–181. [https://doi.org/10.1016/S0034-4257\(97\)00135-1](https://doi.org/10.1016/S0034-4257(97)00135-1).
- Galvao, L.S., Vitorello, I., Formaggio, A.R., 1997. Relationships of spectral reflectance and color among surface and subsurface horizons of tropical soil profiles. *Remote Sens. Environ.* 61, 24–33. [https://doi.org/10.1016/S0034-4257\(96\)00219-2](https://doi.org/10.1016/S0034-4257(96)00219-2).
- Got, A., Moussaoui, A., Zouache, D., 2021. Hybrid filter-wrapper feature selection using whale optimization algorithm: a multi-objective approach. *Expert Syst. Appl.* 183, 115312. <https://doi.org/10.1016/j.eswa.2021.115312>.
- Gui, N., Ge, D., Hu, Z., 2019. AFS: an attention-based mechanism for supervised feature selection. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Guo, J., Jin, S., Wang, K., 2024. Machine learning and genetic algorithm for mapping soil available phosphorus in coastal provinces in Southeast China. *Ecol. Ind.* 166, 112294. <https://doi.org/10.1016/j.ecolind.2024.112294>.
- Heller Pearlshien, D., Ben-Dor, E., 2020. Effect of organic matter content on the spectral signature of iron oxides across the VIS–NIR spectral region in artificial mixtures: an example from a red soil from Israel. *Remote Sens.* 12 (12), 1960. <https://doi.org/10.3390/rs12121960>.
- Hu, P., Pan, J., Chu, S., Sun, C., 2022. Multi-surrogate assisted binary particle swarm optimization algorithm and its application for feature selection. *Appl. Soft Comput.* 121, 108736. <https://doi.org/10.1016/j.asoc.2022.108736>.
- Ji, R., Tan, K., Wang, X., Tang, S., Sun, J., Niu, C., Pan, C., 2025. PatchOut: a novel patch-free approach based on a transformer-CNN hybrid framework for fine-grained land-cover classification on large-scale airborne hyperspectral images. *Int. J. Appl. Earth Obs. Geoinf.* 138, 104457. <https://doi.org/10.1016/j.jag.2025.104457>.
- Jiang, J., Zhang, X., Yuan, Z., 2024. Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. *Expert Syst. Appl.* 249, 123633. <https://doi.org/10.1016/j.eswa.2024.123633>.
- Kok, M., Sarjant, S., Verweij, S., Vaessen, S.F., Ros, G.H., 2024. On-site soil analysis: a novel approach combining NIR spectroscopy, remote sensing and deep learning. *Geoderma* 446, 116903. <https://doi.org/10.1016/j.geoderma.2024.116903>.
- Ladoni, M., Bahrami, H.A., Alavipanah, S.K., Norouzi, A.A., 2010. Estimating soil organic carbon from soil reflectance: a review. *Precis. Agric.* 11 (1), 82–99. <https://doi.org/10.1007/s11119-009-9123-3>.
- Li, H., Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 648 (1), 77–84. <https://doi.org/10.1016/j.aca.2009.06.046>.
- Li, S., Chen, H., Wang, M., Heidari, A.A., Mirjalili, S., 2020. Slime mould algorithm: a new method for stochastic optimization. *Future Gener. Comp. Sy.* 111, 300–323. <https://doi.org/10.1016/j.future.2020.03.055>.
- Li, X., Pan, W., Li, D., Gao, W., Zeng, R., Zheng, G., Cai, K., Zeng, Y., Jiang, C., 2024. Can fusion of vis-NIR and MIR spectra at three levels improve the prediction accuracy of soil nutrients. *Geoderma* 441, 116754. <https://doi.org/10.1016/j.geoderma.2023.116754>.
- Liu, J., Li, D., Shan, W., Liu, S., 2024. A feature selection method based on multiple feature subsets extraction and result fusion for improving classification performance. *Appl. Soft Comput.* 150, 111018. <https://doi.org/10.1016/j.asoc.2023.111018>.
- Liu, M., Yang, K., Yan, Y., Song, Z., Tian, F., Li, F., Yu, Z., Zhang, R., Yang, Q., Lu, Y., 2025. Multi-spectral evaluation of total nitrogen, phosphorus and potassium content in soil using Vis-NIR spectroscopy based on a modified support vector machine with whale optimization algorithm. *Soil Tillage Res.* 252, 106567. <https://doi.org/10.1016/j.still.2025.106567>.
- Lv, H., Chen, C., Zhang, H., Shi, C., Wang, R., 2025. A Two-Stage Feature selection Method for Hyperspectral image Classification: Hybrid Filter Preprocessing and improved Dwarf Mongoose Optimization. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 18, 29271–29294. <https://doi.org/10.1109/JSTARS.2025.3625974>.
- Maroufpoor, S., Maroufpoor, E., Bozorg-Haddad, O., Shiri, J., Yaseen, Z.M., 2019. Soil moisture simulation using hybrid artificial intelligent model: Hybridization of adaptive neuro fuzzy inference system with grey wolf optimizer algorithm. *J. Hydrol.* 575, 544–556. <https://doi.org/10.1016/j.jhydrol.2019.05.045>.
- Mena, F., Pathak, D., Najjar, H., Sanchez, C., Helber, P., Bischke, B., Habelitz, P., Miranda, M., Siddamsetty, J., Nuske, M., Charfuelan, M., Arenas, D., Vollmer, M., Dengel, A., 2025. Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction. *Remote Sens. Environ.* 318, 114547. <https://doi.org/10.1016/j.rse.2024.114547>.
- Miao, F., Wu, Y., Yan, G., Si, X., 2025. Dynamic multi-swarm whale optimization algorithm based on elite tuning for high-dimensional feature selection classification problems. *Appl. Soft Comput.* 169, 112634. <https://doi.org/10.1016/j.asoc.2024.112634>.
- Peng, Y., Gu, S., Liang, Y., Ouyang, K., Li, Y., Wang, K., Wu, G., Fan, C., 2025. Wave Optics Optimizer: a novel meta-heuristic algorithm for engineering optimization. *Commun. Nonlinear Sci. Numer. Simul.* 152, 109337. <https://doi.org/10.1016/j.cnsns.2025.109337>.
- Qian, L., Wu, D., Liu, D., Song, S., Shi, S., Gong, W., Wang, L., 2021. Parameter simulation and design of an airborne hyperspectral imaging LiDAR system. *Remote Sens.* 13 (24), 5123. <https://doi.org/10.3390/rs13245123>.
- Qu, X., Jiang, S., Gu, X., Zhou, J., Liu, X., Zong, F., Li, M., Ji, Y., 2025. Mapping soil organic matter in cultivated land using Landsat 8 image and GA-AdaBoost algorithm. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 18, 25427–25438. <https://doi.org/10.1109/jstars.2025.3614884>.
- Rodgers, J.L., Nicewander, W.A., 1988. Thirteen ways to look at the correlation coefficient. *Am. Stat.* 42 (1), 59–66. <https://doi.org/10.1080/00031305.1988.10475524>.
- Saberioon, M., Gholizadeh, A., Ghaznavi, A., Chabrilat, S., Khosravi, V., 2024. Enhancing soil organic carbon prediction of LUCAS soil database using deep learning and deep feature selection. *Comput. Electron. Agric.* 227, 109494. <https://doi.org/10.1016/j.compag.2024.109494>.
- Şan, M., 2025. Combined innovative trend analysis methods for seasonal trend testing. *J. Hydrol.* 649, 132418. <https://doi.org/10.1016/j.jhydrol.2024.132418>.
- Shang, Y., Zheng, M., Li, J., Zheng, X., 2025. An effective feature selection approach based on hybrid Grey Wolf Optimizer and Genetic Algorithm for hyperspectral image. *Sci. Rep.* 15 (1), 1968. <https://doi.org/10.1038/s41598-024-84934-8>.
- Shi, X., Song, J., Wang, H., Lv, X., Zhu, Y., Zhang, W., Bu, W., Zeng, L., 2023. Improving soil organic matter estimation accuracy by combining optimal spectral preprocessing and feature selection methods based on pXRF and vis-NIR data fusion. *Geoderma* 430, 116301. <https://doi.org/10.1016/j.geoderma.2022.116301>.
- Shin, S.K., Lee, S.J., Park, J.H., 2025. Prediction of soil properties using Vis-NIR spectroscopy combined with machine learning: a review. *Sensors* 25 (16), 5045. <https://doi.org/10.3390/s25165045>.
- Stenberg, B., Rossel, R.A.V., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* 107, 163–215. [https://doi.org/10.1016/S0065-2113\(10\)07005-7](https://doi.org/10.1016/S0065-2113(10)07005-7).
- Su, H., 2022. Dimensionality reduction for hyperspectral remote sensing: advances, challenges, and prospects. *Nat. Remote Sens. Bull.* 26 (8), 1504–1529. <https://doi.org/10.11834/jrs.20210354>.
- Tan, J., Ding, J., Wang, Z., Han, L., Wang, X., Li, Y., Zhang, Z., Meng, S., Cai, W., Hong, Y., 2024. Estimating soil salinity in mulched cotton fields using UAV-based hyperspectral remote sensing and a seagull optimization algorithm-enhanced random forest model. *Comput. Electron. Agric.* 221, 109017. <https://doi.org/10.1016/j.compag.2024.109017>.
- Tan, K., Zhu, L., Wang, X., 2025. A hyperspectral feature selection method for soil organic matter estimation based on an improved weighted marine predators algorithm. *IEEE Trans. Geosci. Remote Sens.* 63, 5500411. <https://doi.org/10.1109/tgrs.2024.3422475>.
- Teofiloff, R.F., Martins, J.P.A., Ferreira, M.M., 2009. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemom.* 23 (1), 32–48. <https://doi.org/10.1002/cem.1192>.
- Thissen, U., Peppers, M., Üstün, B., Melssen, W.J., Buydens, L.M.C., 2004. Comparing support vector machines to PLS for spectral regression applications. *Chemom. Intell. Lab. Syst.* 73 (2), 169–179. <https://doi.org/10.1016/j.chemolab.2004.01.002>.
- Wang, N., Chen, S., Huang, J., Frappart, F., Taghizadeh, R., Zhang, X., Wigneron, J.-P., Xue, J., Xiao, Y., Peng, J., 2024. Global soil salinity estimation at 10 m using multi-source remote sensing. *J. Remote Sens.* 4, 0130. <https://doi.org/10.34133/remotesensing.0130>.
- Wang, S., Cheng, X., Zheng, D., Song, H., Han, P., Yuen, P., 2020. Prediction of the soil organic matter (SOM) content from moist soil using synchronous two-dimensional correlation spectroscopy (2D-COS) analysis. *Sensors* 20 (17), 4822. <https://doi.org/10.3390/s20174822>.
- Wang, W., Wu, W., Zhang, W., Hu, D., Xu, W., Bai, J., Wang, Y., 2026. Spatiotemporal dynamics of tree species-level aboveground carbon storage at the canal scale under green engineering with a random forest model. *Remote Sens.* 18 (3), 475. <https://doi.org/10.3390/rs18030475>.
- Wu, F., Tan, K., Wang, X., Ding, J., Liu, Z., 2023. A novel semi-empirical soil multi-factor radiative transfer model for soil organic matter estimation based on hyperspectral imagery. *Geoderma* 437, 116605. <https://doi.org/10.1016/j.geoderma.2023.116605>.
- Xia, H., Chen, L., Xu, H., 2025. Multi-strategy dung beetle optimizer for global optimization and feature selection. *Int. J. Mach. Learn. Cybern.* 16 (1), 189–231. <https://doi.org/10.1007/s13042-024-02197-1>.
- Xiao, W., Lian, J., Ouyang, K., Gu, S., Ke, Z., Wei, D., Sha, X., Wang, J., Fu, S., Qiu, M., 2026. Newton downhill optimizer with application to engineering optimization and breast cancer feature selection. *Biomed. Signal Process. Control* 117, 109184. <https://doi.org/10.1016/j.bspc.2025.109184>.
- Xue, J., Shen, B., 2020. A novel swarm intelligence optimization approach: sparrow search algorithm. *Syst. Sci. Control Eng.* 8 (1), 22–34. <https://doi.org/10.1080/21642583.2019.1708830>.
- Yang, H., Yu, M., Zhang, J., Xiong, Y., Wang, D., Xu, J., 2026. A reinforced quantum aquila optimizer for multi-threat 3D UAVs path planning in complex environments. *App. Math. Model.* 155, 116736. <https://doi.org/10.1016/j.apm.2025.116736>.
- Zhang, J., Ding, J., Zhang, Z., Wang, J., Zeng, X., Ge, X., 2025. Study on the inversion and spatiotemporal variation mechanism of soil salinization at multiple depths in typical oases in arid areas: a case study of Wei-Ku Oasis. *Agric Water Manag* 315, 109542. <https://doi.org/10.1016/j.agwat.2025.109542>.