Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

GF-5 hyperspectral inversion of soil parameters using a VAE style-based spectral fusion model☆

Depin Ou [a], Jie Li [b], Zhifeng Wu [c,*], Kun Tan [d,e,*], Weibo Ma [f,g], Xue Wang [d,e], Yueqin Zhu [a]

[a] *National Institute of Natural Hazards, Ministry of Emergency Management, Beijing 100085, China*
[b] *South China Institute of Environmental Sciences, Ministry of Ecology and Environment of the People's Republic of China, Guangzhou 510655, China*
[c] *School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China*
[d] *Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China*
[e] *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*
[f] *Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment, Nanjing 210042, China*
[g] *Engineering Research Center of Ministry of Education for Mine Ecological Restoration, China University of Mining and Technology, Xuzhou Jiangsu, 221116, China*

ARTICLE INFO

Keywords:
GF-5 hyperspectral image
Spectra data fusion
Soil organic matter
Soil heavy metals
VAE

ABSTRACT

Inverting soil parameters through hyperspectral techniques is currently one of the highly popular research topics and the major challenges in quantitative remote sensing. To date, indoor spectral data-based inversion models cannot be directly applied to satellite-based hyperspectral data, due to the weak model migration capability caused by the large differences between the two spectral data. Therefore, the present study aims to improve the inversion soil parameter accuracies using satellite-based GF-5 hyperspectral remote sensing data by merging multiple hyperspectral data. First, indoor Analytical Spectral Devices (ASD) hyperspectral and pre-processed GF-5 data of soil samples were used to develop a variational auto-encoder (VAE)-based spectral fusion model capable of transforming GF-5 spectra into indoor spectra. Second, traditional machine learning regression algorithms, namely Partial Least Squares Regression (PLSR) and Support Vector Regression (SVR), were used to build an inversion model using the mixed spectra data to determine the spatial distributions of soil organic matter (SOM), arsenic (As) and copper (Cu) contents across a large study area. The results demonstrated the effectiveness of the VAE-based spectral fusion model in removing substantial noise information while preserving the spectral features from the GF-5 data. The optimal inversion accuracies of the SOM, As, and Cu contents showed coefficients of determination (R^2) of 0.87, 0.88, and 0.85, which are 38%, 55%, and 28% higher than those obtained using the original GF-5 data-derived model, respectively. In addition, the spatial distributions of the SOM, As, and Cu contents demonstrated that the GF-5 satellite data are more intuitive and effective for large-scale soil composition analysis.

1. Introduction

As the Earth's epidermis, the soil is a surface material located at the interface between the atmosphere and the lithosphere, playing an essential role in global climate change and biogeochemical cycles (Song et al., 2022). The assessment of soil composition has theoretical relevance for studying carbon peak, carbon neutrality, global warming, and extreme global weather evolution. Traditional in-situ sampling methods are costly, labor-intensive, and time-consuming, delaying soil monitoring activities (Ben-Dor, 2002). In contrast, hyperspectral remote sensing technology can rapidly detect minute changes in soils and extract soil components (Du et al., 2020). Indeed, several hyperspectral methods have been developed and used to invert soil components worldwide, attracting great attention from researchers worldwide (Ben-Dor, 2002; Ben-Dor et al., 2019).

The main objective of hyperspectral soil component inversion is to obtain sensitive bands or feature bands with accurate and universal identification of soil compositions. However, only a small number of soil

☆ This article is part of a special issue entitled: 'QRSM' published in Computers and Electronics in Agriculture.

* Corresponding authors.
E-mail addresses: zfwu@gzhu.edu.cn (Z. Wu), tankuncu@gmail.com (K. Tan).

https://doi.org/10.1016/j.compag.2025.110214
Received 21 October 2024; Received in revised form 25 February 2025; Accepted 27 February 2025
Available online 4 March 2025

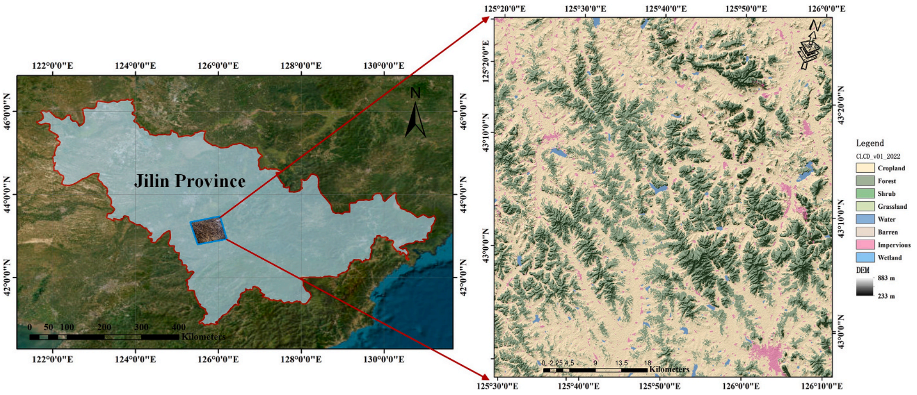

Fig. 1. Study area and its land cover product.

components exhibit distinct and consistent sensitivity bands. Indeed, water has sensitive band features around 1400 and 1900 nm (Viscarra Rossel and Behrens, 2010), while clay minerals exhibit sensitive band features around 2200 nm (Lagacherie et al., 2008). In contrast, there are no particular sensitivity features regarding soil organic matter (SOM). However, Ou et al. (2022) highlighted a good correlation of the SOM scattering characteristics at around 2200 nm using the Kubelka-Munk inversion method. While other researchers have indicated a good correlation of the SOM scattering characteristics at the visible near-infrared (VNIR) band range (Al-Abbas et al., 1972; Chabrillat et al., 2019). On the other hand, soil heavy metal arsenic (As) is more sensitive in the short-wavelength infrared (Ou et al., 2021). In addition, Wu et al. (2007) demonstrated the presence of distinct spectral absorption features of soil chromium (Cr) and copper (Cu), when concentrations greater than 4000 mg/kg at 0.61 and 0.83 μm, respectively. However, soils with heavy metal concentrations at such high concentrations are uncommon in nature. Therefore, sensitive features of soil heavy metals and SOM are still to be investigated.

Data-driven models are currently the most widely tools for implementing hyperspectral inversion of soil components. Indeed, these models can provide good inversion at the local scale and enable rapid soil component mapping over research areas (Ben-Dor, 2002). Most indoor spectrum data and hyperspectral remote sensing data modeling methods implement a similar process, involving data pre-processing, feature selection/extraction, and statistical Regression modeling-mapping (Gholizadeh et al., 2015; Shi et al., 2014; Wang et al., 2018). In general, preprocessing involves filtering, transformation, and other techniques; Feature selection/extraction mainly involves the use of some models, such as the Competitive Adaptive Reweighted Sampling (CARS) algorithm and Pearson correlation analysis (Cohen et al., 2009); Statistical regression modeling is performed by several methods, such as Partial Least Squares Regression (PLSR) and Support Vector Regression (SVR) (Smola and Schölkopf, 2004). Statistical models can provide satisfactory inversion results when applied to indoor and airborne spectral data. Tan et al. (2021) employed the CARS and stacking integration techniques to develop heavy metal inversion models of As, Cr, Pb, and Zn for airborne HyMap hyperspectral data of the northeastern black land, showing correlation coefficients greater than 0.6. In recent years, Deep Learning has emerged as a substantial breakthrough in the machine learning field. Because of its great advantage in feature extraction, it has been used in several studies on hyperspectral inversions. Padarian et al. (2019) used a multitask CNN model to construct prediction models for soil organic carbon, clay, pH, and total nitrogen (N) based on 20,000 LUCAS-derived surface soil spectra from 23 European Union countries, demonstrating the effectiveness of the multitask CNN model in reducing the prediction error of soil organic carbon by 87 and 62 % compared to the PLSR and Cubist regression methods, respectively. Whereas Tsakiridis et al. (2020) implemented a local multichannel 1D convolutional neural network to predict the organic matter components based on the LUCAS dataset, with a performance of 0.86, indicating a significantly improved accuracy over the standard regression methods.

Although indoor hyperspectral and hyperspectral remote sensing data can be used to construct a reliable statistical regression model. Indoor spectral models cannot be applied directly to remote sensing data and vice versa (Wan et al., 2022). Laboratory spectroscopy can provide comprehensive insights for constructing high-quality inversion models through the elimination of a large variety of environmental disturbances (e.g., atmosphere, soil particle size, and soil moisture) (Camargo et al., 2015; Chabrillat et al., 2019; Piekarczyk et al., 2016). However, numerous environmental factors can influence airborne and satellite-based hyperspectral remote sensing data, resulting in low precision or even model failure. Moreover, unlike laboratory spectra, it is often challenging to introduce sensitive features in remote sensing spectral data due to the low soil heavy metal contents. Therefore, in order to enhance the soil spectral features of hyperspectral remote sensing data, it is crucial to eliminate environmental factors. Multi-source data fusion has been attracting great interest. Indeed, deep generative models, such as variational auto-encoder (VAE) and generative adversarial network (GAN), have been successfully applied to hyperspectral image classification in several related studies, providing novel methods for data fusion (Wang et al., 2019; Yu et al., 2020).

In multi-source spectral data fusion, deep generative models need primarily to address two issues. First, unlike classification tasks, where various categories of spectra have distinct differentiability, quantitative analysis of soil composition requires accurate differentiable features; Second, assessing the generative spectra-derived features accuracy. In this context, the main objective of the present study is to integrate the spectral features of soil components from indoor spectra and remote sensing hyperspectral data in order to achieve high-precision inversions of SOM and soil heavy metal contents. The novelty of the present study lies primarily in the feature fusion and spectral generation of laboratory spectra and satellite-based remote sensing data using an improved VAE-based technique. Our proposed method can generate stable and reliable full-spectrum data while preserving the spectral soil features. The constructed statistical regression inversion model is highly accurate, providing a reliable and stable baseline product for large-scale soil pollution and fertility assessments.

2. Study area and datasets

2.1. Study area

The study area is located in the southeastern part of Yitong Manchu Autonomous County in Western Jilin Province, China, which ranges from 125.31°E-126.21°E and 42.93°N-43.33°N, covering a total area of about 3800 km^2 (Fig. 1). The minimum and maximum elevations in the study area are 233 and 833 m, respectively, with mostly hilly and

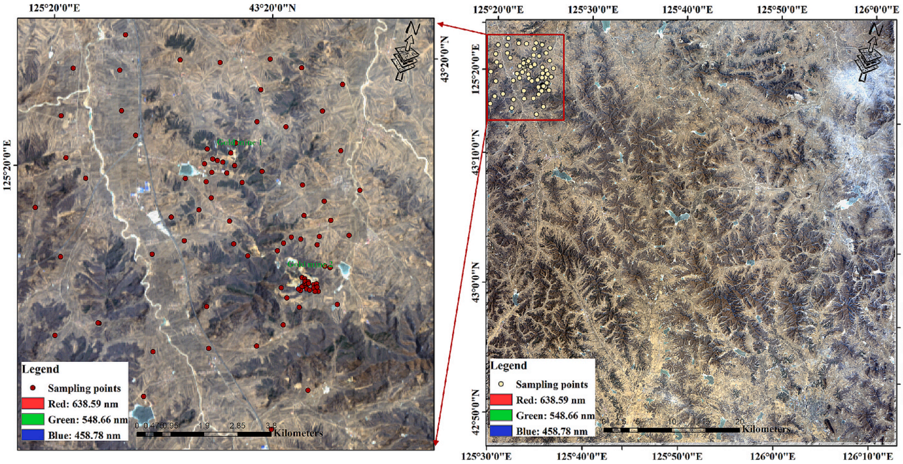

Fig. 2. Geographic locations of field sampling points overlaid on the GF-5 hyperspectral image.

Table 1
Basic information on soil composition in the study area.

	Descriptive statistics						Pearson Correlation		
	Max	Min	Mean	C.V	Kurtosis	Skewness	SOM	As	Cu
SOM (g/kg)	49.84	14.76	30.55	0.21	0.72	0.11	1.00	–	–
As (mg/kg)	419.96	6.35	44.12	1.55	13.55	3.40	0.03	1.00	–
Cu (mg/kg)	94.45	10.37	20.02	0.63	16.09	3.73	0.12	−0.07	1.00

relatively flat terrain. The study area is located in a black soil area in the northeastern part, where the main soil type is a moderately fertile dark brown loam. The land cover product of China (CLCD) is obtained from Yang and Huang (Yang and Huang, 2023). From the CLCD data, it can be seen that cropland dominates the area, with forested land being the second largest. Crops are relatively homogeneous, consisting mainly of corn and rice in the lowlands. There are few industries in the study area, but there are numerous small gold mines, some of which are still operational, constituting the main sources of pollution in the study area.

2.2. Datasets

2.2.1. Field sampling data and laboratory spectra

Soil samples were collected from 91 sampling points, evenly distributed across a 100 km^2 cultivated area in the upper northwestern part of the study area, over the April 18-April 22, 2017 period. The geographic locations of the field sampling points are shown in Fig. 2. All soil samples were collected from the topsoil layer (0–5 cm). Real-time kinematic (RTK) positioning was used concurrently to record the precise location data of the sampling sites. The collected soil samples first underwent a preliminary removal of impurities, then air-dried, ground, and sieved through a 100-mesh sieve. In addition, 91 soil spectra were measured using an ASD FieldSpec3 (400–2500 nm) and a 6.5-watt-tungsten-halogen lamp.

In this study, the potassium dichromate volumetric method was used to determine the SOM contents, while the ICP-MS method was used to determine the soil As and Cu contents. Table 1 provides an overview of the soil composition observed in the study area. The SOM content was

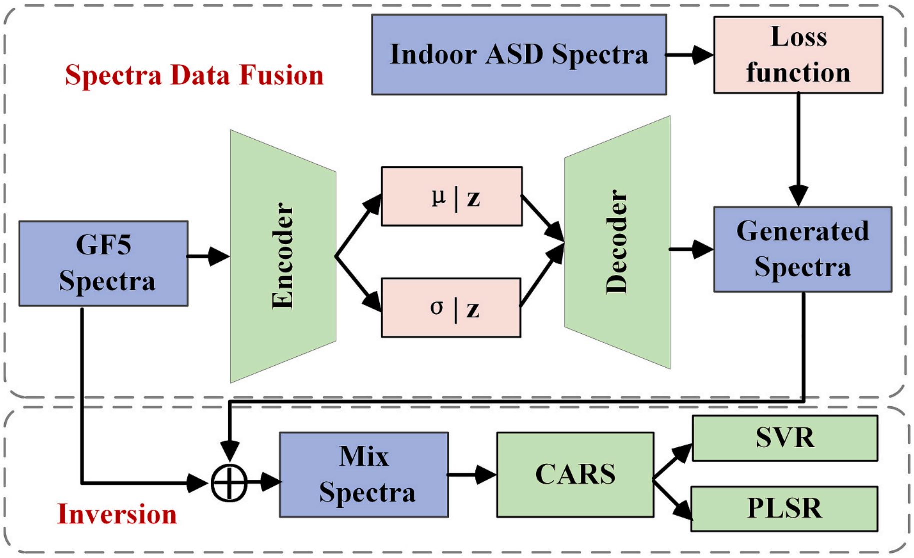

Fig. 3. Flowchart of the methodology used in this study.

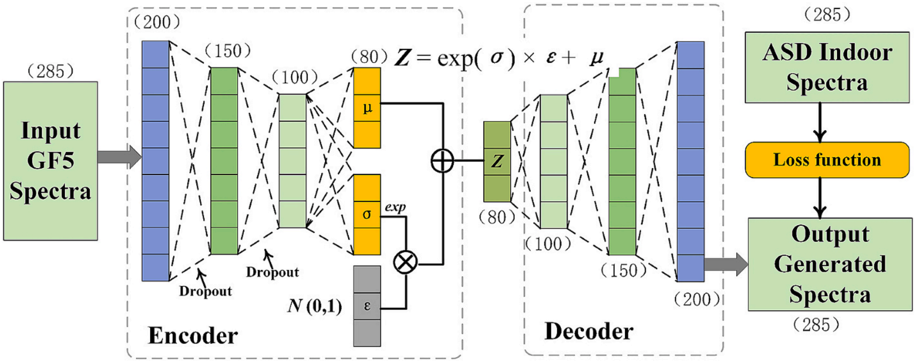

Fig. 4. VAE style-based spectral fusion model.

30 g/kg, with a low coefficient of variation, suggesting fewer effects of human activities on the SOM contents. In contrast, the coefficients of variation of soil As and Cu were high. Moreover, the high kurtosis and skewness values indicated that soil samples were more likely to be influenced by human activities. The low correlation between the three analyzed soil parameters suggests that there were no concomitant phenomena.

2.2.2. GF-5 hyperspectral data

The Gaofen-5 (GF-5) satellite, launched on May 9, 2018, is part of the China High-Resolution Earth Observation System. The GF-5 satellite is equipped with VNIR and SWIR hyperspectral sensors, of which the VNIR (400–1000 nm) sensor has a spectral resolution of 5 nm and contains 150 bands, while the SWIR (1000–2500 nm) sensor has a spectral resolution of 10 nm and contains 180 bands. The swath width and spatial resolution are 60 km and 30 m, respectively.

In this study, a high-quality GF-5 hyperspectral image was acquired on February 1, 2019. The acquired GF-5 hyperspectral images were preprocessed using radiometric calibration, geometric correction, and atmospheric correction. The calibration coefficients reported by Tan et al. (2020) were used as radiometric calibrations, while atmospheric correction was performed using FLAASH software (Berk et al., 1999). The preprocessed GF-5 hyperspectral image data had a size of 2008 × 2083 × 330 pixels, with a spatial resolution of 30 m and a digitization footprint of 0.00699 Mb/ha (Fig. 3). Water vapor absorption leads to low reflectance and significant noise near 1400 nm and 1800 nm (Guanter et al., 2006). Therefore, the spectral regions between 1342–1451 nm and 1771–1977 nm were removed, while 285 bands were finally retained.

As shown in Fig. 2, the ice in the study area had not melted. In addition, there was a lack of leaves since the images were acquired in February. Thin clouds on the right side of the GF-5 image may affect the inversion mapping result and must, therefore, be removed. Although the area from which the soil samples were collected is relatively small (100 km^2), the overall image exhibits high radiometric consistency (Tan et al., 2020). Ge et al.(2022) and Wu et al. (2021) have demonstrated the effectiveness of targeted regional sampling for producing large scale maps of soil composition with enhanced precision. Meng et al.'s (2024) SOC results showed a high degree of consistency in the organic carbon content of this study area. Therefore, the soil samples are representative, and the inversion results can reflect the overall distribution trend.

2.2.3. Digital Elevation model

The Digital Elevation Model (DEM), with a resolution of 12.5 m, was obtained from the Alaska Satellite Facility Distributed Active Archive Data Center (ASF DAAC) (Logan et al., 2014).

3. Methods

Fig. 3 shows the flowchart of the methodology used in this study. First, GF-5 and indoor ASD spectra were merged using an implemented VAE-style deep network. Although the generated spectra contain the spectral information of ASD, some GF5 spectral features were also lost. Therefore, the GF5 spectra were combined with the generated spectra. Finally, the CARS method was used to select features of the mixed spectra, while the SVR and PLSR methods were used to construct the soil composition inversion model.

3.1. Spectral data fusion

Despite the two-year difference between soil sampling and hyperspectral image acquisition, low variability and high correlation of soil heavy metal content were reported (Wu et al., 2021). The results of (Castaldi et al., 2018; Lagacherie et al., 2012; Qin et al., 2021) also show that the time difference between the image and the sampling points can also obtain better inversion results. In particular, Castaldi et al. (2018) used 2015 APEX data with 2009 LUCAS soil sample data for soil organic matter inversion, which also suggests that time span has less impact on soil composition inversion results. Meng et al.'s (2024) SOC result shows a moderate decrease (0.40 g/kg) from 2001 to 2021, which indicates that soil composition in the study area changed very little over the two-year period. Direct applications of indoor spectra-derived soil composition inversion models to imaging data lead often to very poor accuracy, which is due to the significant difference between indoor and imaging spectra. Indoor spectra are less sensitive to environmental perturbations, thus the model accuracies can reach 0.9 or higher. In contrast, imaging spectral data make it challenging to construct similarly accurate models. Therefore, a VAE style-based spectral fusion model was constructed to mitigate the time interval-related spectral differences and to fully utilize the indoor spectral and imaging spectral information to improve the inversion accuracy.

3.1.1. VAE style-based spectral fusion model

Variational auto-encoder (VAE) is an extension of the AE data generation model that includes distribution in the latent space (Phillips and Abdulla, 2022). The training of VAE involves the introduction of probability distributions into the latent space in order to prevent overfitting and ensure appropriate latent space properties for the data generation process (Doersch, 2016). The principle of the proposed VAE style-based spectral fusion model is shown in Fig. 4. The model is similar to the conventional VAE model, except that the final loss function optimization is calculated using the indoor and generated spectra. Due to the high dimensional input spectra, the model employs three implicit layers in the encoding and decoding processes, resulting in a closer approximation of the generated spectra to the indoor spectra. To minimize overfitting, we used a dropout of 0.4 to the encoding layer, with ReLu as the activation function between the hidden layers.

3.1.2. Loss function

The encoding process of the model ensures that the data distribution meets a normal distribution. The reparameterization trick function can be computed mainly using the following equation:

$$\boldsymbol{Z} = exp(\sigma) \times \varepsilon + \mu \tag{1}$$

Where σ denotes the sample variance; ε is obtained from a random normal distribution; μ denotes the sample mean.

The Kullbeck-Leibler divergence (KLD) was used to transform the latent distribution to a standard Gaussian (Cha et al., 2019). As shown in Eq. (2), it is sufficient to minimize the KLD in the loss function.

$$\mathrm{KLD} = \sum \left(exp(\sigma) - (1 + \sigma) + \mu^2 \right) \tag{2}$$

Since our samples consisted of soils, the homogeneity of soil spectral is high. In addition, the trained spectra may be different when using methods that directly result in minimized mean square error (MSE). Therefore, the spectral angle distance (SAD) is considered part of the loss function, making the generated spectra more convergent to the indoor spectra while retaining more spectral features (Sohn and Rebello, 2002). The spectral angle distance can be determined using Eq. (3).

$$\boldsymbol{Min}\mathrm{SAD}(X^*, X_i) = \cos^{-1}\left(\frac{(X^*)^{\mathrm{T}} X_i}{\left((X^*)^{\mathrm{T}} X^*\right)^{1/2} \left(X_i^T X_i\right)^{1/2}} \right) \tag{3}$$

Where X^* denote the generated spectra from the sample X_i denotes the ASD indoor spectra. The SAD value is small when two spectra are more similar.

The KLD can control the distribution of the generated spectral data, while SAD can maintain the generated spectra's waveforms more similar to indoor spectra. However, in quantitative remote sensing, the spectra between different components may also differ significantly from the overall spectra. Therefore, distance measures are required to make the distribution of the generated spectra similar to the indoor spectra, maintaining the soil component feature patterns on the spectra. The Maximum Mean Discrepancy (MMD) was used to determine the distance between two distinct and related distributions according to the following equation (Borgwardt et al., 2006):

$$\boldsymbol{Min}\mathrm{MMD}(\mathrm{X}, \mathrm{Y}) = \left\| \frac{1}{n} \sum_{i=1}^{n} \varnothing(x_i) - \frac{1}{m} \sum_{j=1}^{m} \varnothing(y_i) \right\|_H^2 \tag{4}$$

Where H denotes the distance measured by the mapping (∅) of the regenerated Hilbert space. ∅ can be expressed as a kernel function. In this study, we used the Gaussian kernel function. Hence, the final loss function can be expressed as follows:

$$\boldsymbol{Min}\mathrm{loss} = \mathrm{MMD} + \alpha^* \mathrm{SAD} + \mathrm{KLD} \tag{5}$$

Where α is used to regulate the spectral dominant generating features. Higher and lower α values indicate a more similar spectral spectrum and distribution feature, respectively. The final adjustment of the α value can balance the spectral waveform and spectral distribution. In this study, α was set to 0.1.

3.1.3. Spectral fine-tuning spectra

Since there is a dropout in the neural network, changing random seeds during algorithm evaluation may exhibit an effect on the algorithm accuracy results (Picard, 2021). Moreover, minor changes in soil spectra can also influence the correlation between soil compositions and spectra. Therefore, defining the optimal number of random seeds is necessary to improve the correlation between soil compositions and spectra. In this study, the optimal number of random seeds for each component was defined within a range from −9999 to 9999 to improve the correlations between the observed SOM, As, and Cu contents of the training samples and the generated spectra. Savitsky-Golay filters were first used to determine the three optimal generated spectra, then the final generated spectra were obtained after averaging. However, the generated spectra may lose some of the original features. Hence, they were combined with the original GF-5 spectra to perform the fusion of multiple spectral features.

In the VAE style-based spectral fusion model experiment, indoor spectra with the same dimensions as GF5 spectra were created using linear interpolation. The optimizer used employs the Adam algorithm, with optimal step size and epoch of $1e^{-4}$ and 100, respectively (Kingma and Ba, 2014). The dataset was divided into training and testing sub-datasets according to sample component gradient, with a ratio of 2:1. However, the deep network may easily exhibit overfitting due to the small size of the training sub-dataset. Tobler's first law of geography (Tobler, 1970) shows that closely related substances are similar. Therefore, the four-neighborhood spectra of each training sample were selected as pseudo-labeled samples for training purposes. Due to the severe unmixing issue in the 30-m-resolution GF-5 spectral data, the spectral angular distance (SAD < 0.08) was used in the selection process to determine whether additional data are required in the training step.

3.2. Inversion method

There is a large amount of redundant information between contiguous hyperspectral bands, as they are highly correlated. Therefore, the implementation of feature selection and feature extraction techniques is required prior to inversion to improve its accuracy (Asadzadeh and de Souza Filho, 2016). The Competitive Adaptive Reweighted Sampling (CARS) algorithm is a spectral feature selection algorithm that is based on the Darwinian evolutionary theory of "survival of the fittest". Indeed, the CARS method uses Monte Carlo iteration and competition to select several subsets of wavelengths according to their importance, then select the optimal wavelengths and eliminate those with significant errors using the decay index method and Adaptive Reweighted Sampling (ARS) algorithm through multiple cross-validations. The CARS algorithm was first applied to the fused spectra to select features, then SVR and PLSR were used for model training and testing purposes.

3.3. Hydrologic and topographic factors analysis

The Soil and Water Assessment Tool (SWAT) model was used to generate the stream network and analyze the accumulation and transport behaviors of soil parameters contents in the study area (Arnold and Fohrer, 2005).

The topographic factors Topographic Wetness Index (TWI), SLPOE, and LS-factor were generated by the open-source software SAGA and used for the subsequent correlation discussion and analysis of the soil composition(Conrad, 2006).

3.4. Model evaluation methods

In this study, four statistical metrics were used for modeling evaluation.

(1) Coefficient of determination, R^2: denotes the degree of variation in the variables explained by the model. The closer the value is to 1, the better the prediction.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{6}$$

(2) Root-mean-square error, RMSE: measures the degree of deviation between the predicted value and the true value. The smaller the value, the higher the accuracy, and the better the prediction.

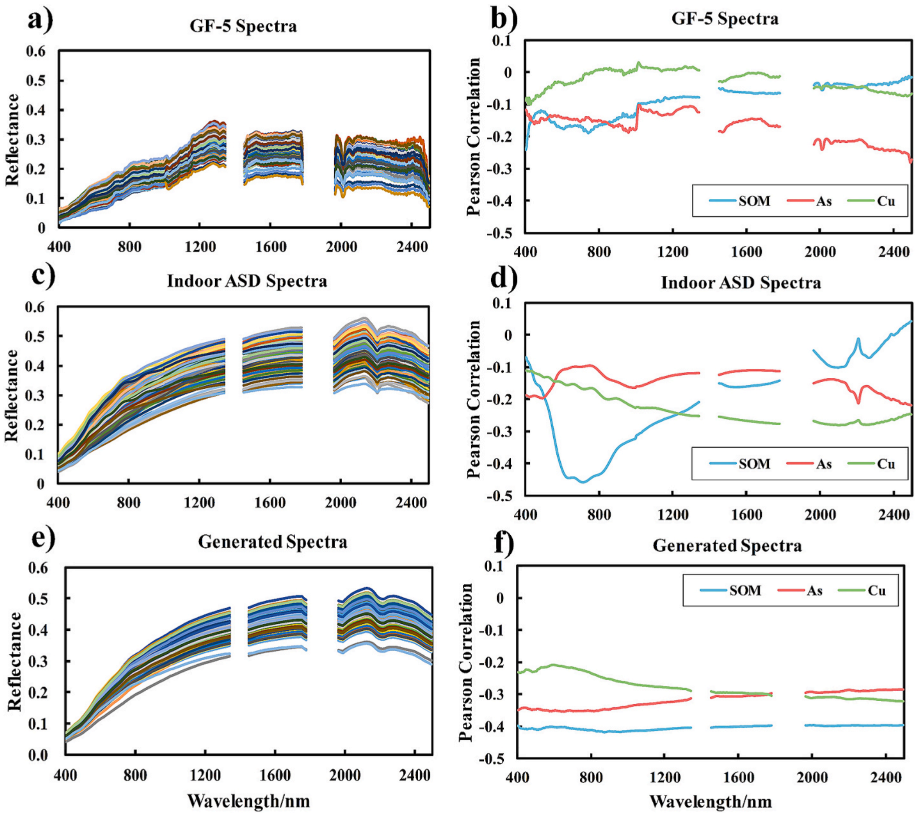

Fig. 5. Different spectra and their Pearson correlation coefficients with the soil parameters. Soil spectra extracted from the GF-5 (a), indoor ASD (c), and proposed data fusion model (e), and their Pearson correlation coefficients with the observed soil parameters (b, d, and f).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{n}} \tag{7}$$

(3) Mean absolute error, MAE: Average value of the difference between the predicted value and the real value. The smaller the value, the better the prediction

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y}_i - y_i| \tag{8}$$

(4) Ratio of performance to inter-quartile distance, RPIQ: A larger RPIQ value indicates improved model performance.

$$RPIQ = \frac{\mathrm{Q3} - \mathrm{Q1}}{RMSE} \tag{9}$$

Where Q1 is the value below which we can find 25 % of the samples; Q3 is the value below which we find 75 % of the samples.

4. Results and analysis

4.1. Spectral generation analysis

The results of GF-5, indoor, and generated spectra are shown in Fig. 5. Overall, the spectral reflectance of GF-5 was relatively low. In addition, the spectral curves were not sufficiently smooth and contained comparatively more noise. Since the 400–1000 nm and 1000–2500 nm sensors were distinct, the continuity at the 1000 nm splice was relatively poor. The atmosphere might significantly impact the spectra, particularly around 1400 nm and 1900 nm, as the satellite is located at 705 km altitude. The indoor spectral reflectance was greater, while the spectrum and spectral differentiation were more uniform and enhanced, respectively. In contrast, the spectra generated by the proposed method were closely similar to the indoor spectra in terms of reflectance and noise, showing a high degree of similarity between the spectral curves and the indoor spectra. Indeed, although GF-5 showed a lack of spectral features over 2200 nm, the generated spectra were relatively similar to the indoor spectra. The generated spectra maintained, to a certain extent, the waveform features of the indoor spectra. However, it remains challenging to determine whether some important GF-5 spectrum features were not lost.

The Pearson correlation coefficients of the generated spectra with three soil components (SOM, Cu, and As contents) are shown in Fig. 5. The observed SOM contents showed the strongest negative correlations with the GF-5 spectra of −0.19 and −0.28 at 741 nm and above 2000 nm, respectively. Whereas the soil Cu contents showed a weak correlation with the GF-5 spectra in the entire spectral range. Regarding the indoor hyperspectral, the SOM contents exhibited strong negative correlations in the Visible-NIR range, reaching −0.46 at 711 nm, while the soil As contents showed a moderate negative correlation of −0.22 after 2000 nm. In addition, the correlation coefficient between the soil Cu contents and the indoor spectra decreased with increasing wavelength, reaching the value of −0.28 at 2100 nm. The comparison of the generated spectra with the GF-5 hyperspectral data highlighted a substantial improvement in the correlation of the three soil parameters. The SOM and As contents showed the strongest negative correlations of −0.42 and −0.35 at 873 and 642 nm, respectively, while the soil Cu contents exhibited a substantial improvement of the correlation coefficient, reaching −0.32 after 2200 nm. In general, the generated spectra can approach the indoor spectra' correlation, even by combining the advantages of the GF-5 and indoor spectra. However, the overall

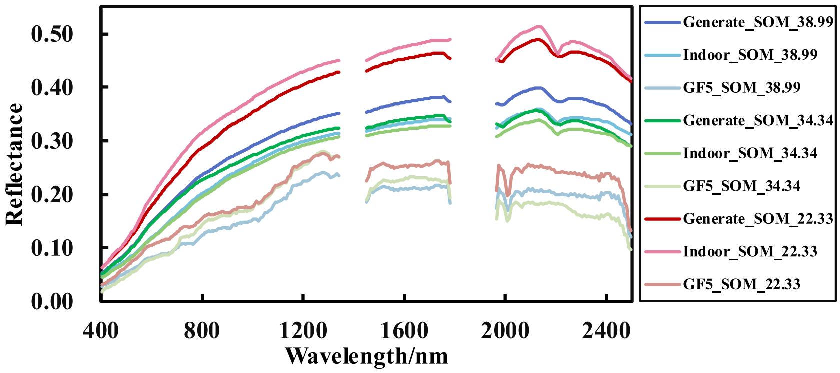

Fig. 6. Comparison between the GF5, indoor, and generated spectra.

Table 2
Evaluation of the optimal soil composition inversion models.

Soil	Regression models	No. of Features	Training sub-dataset				Testing sub-dataset			
			R_c^2	RMSEc	MAEc	RPIQc	R_p^2	RMSEp	MAEp	RPIQp
SOM	CARS + SVR	51	0.94	1.60	0.76	4.71	**0.87**	2.21	1.77	3.41
As	CARS + SVR	64	0.96	14.34	4.14	2.63	**0.88**	21.98	15.32	1.35
Cu	CARS + PLSR	36	0.95	2.99	2.29	1.95	**0.85**	4.47	3.51	1.22

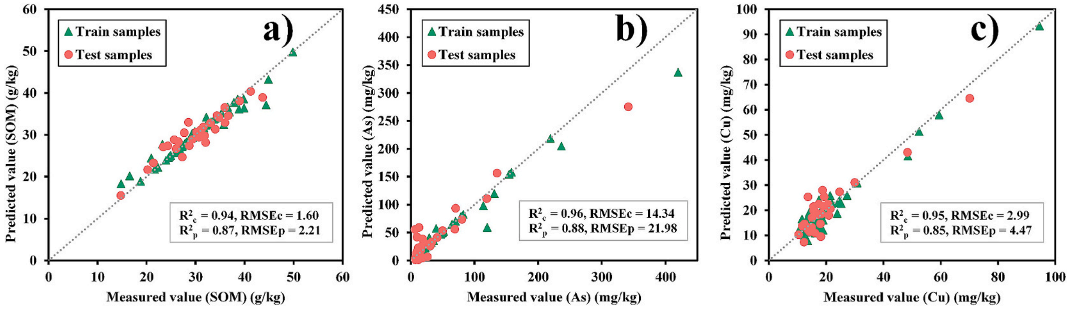

Fig. 7. Scatter plots of the predicted and measured contents of SOM (a), As (b), and Cu (c).

correlation of all spectra was more similar, suggesting the loss of some spectral features. Therefore, the final inversion process requires the addition of the original GF-5 spectra.

Fig. 6 shows the spectra of three selected samples with different SOM contents for comparison. GF5_SOM_22.33 represents the GF5 spectral for SOM with 22.33 g/kg, Indoor_SOM_22.33 represents the indoor ASD spectral for SOM with 22.33 g/kg, Generate_SOM_22.33 represents the generated spectral by our proposed method for SOM with 22.33 g/kg, separately. The obtained results showed a lower and pronounced reflectance of the GF-5 spectra in the visible spectral range. The spectra generated by our method are all close to the target spectrum (indoor spectrum), with a higher degree of differentiation. In addition, the two spectral curves of GF5_SOM_22.33 and GF5_SOM_34.34 obviously overlapped in the 800–1400 nm wavelength. In contrast, the spectra generated by our proposed model were more distinct, indicating that these spectra were generated based on the entire spectral features, thereby demonstrating the better performance of the generated model.

4.2. Soil composition inversion model

The optimal inversion models for the SOM, AS, and Cu contents in the soils obtained by the CARS-based SVR and PLSR methods for mixed spectral features (original GF5 and proposed model-generated spectra) are indicated in Table 2. Fig. 7 shows scatter plots of the predicted and measured values. The SOM training and testing sub-datasets showed R^2 values of 0.94 and 0.87, respectively, with low RMSE values and high RPIQ values greater than 3.4. These findings demonstrated the effectiveness of the model in predicting the soil parameters with very small predicted errors. The scatter plots showed a linear distribution of the training and testing data close to the 1:1 line (Fig. 7a), indicating that the overfitting phenomenon was not obvious. The soil As training and testing sub-datasets showed R^2 values of 0.96 and 0.88, with RPIQ values of 2.63 and 1.35, respectively, demonstrating good prediction accuracy of the model. However, some high predicted errors were observed due to the uneven spatial distributions of the soil sampling points and soil As contents, showing high RMSE values. Fig. 7b showed less pronounced overfitting between the soil As training and testing sub-datasets is less pronounced, particularly under high soil As contents. The soil Cu training and testing sub-dataset revealed R^2 values of 0.95 and 0.85, respectively, with relatively low RMSE values and RPIQ values greater than 1.2, demonstrating the high performance of the model in predicting the soil Cu contents. Although Fig. 7c showed uneven sample data distributions, the model had a high accuracy in predicting high soil Cu contents.

4.3. Analysis of the spatial distributions of soil parameters

The spatial distributions of the SOM, As, and Cu contents are shown in Figs. 8, 9, and 10. By overlaying the DEM-derived hydric network vectors, we can more visibly analyze the spatial distributions of the

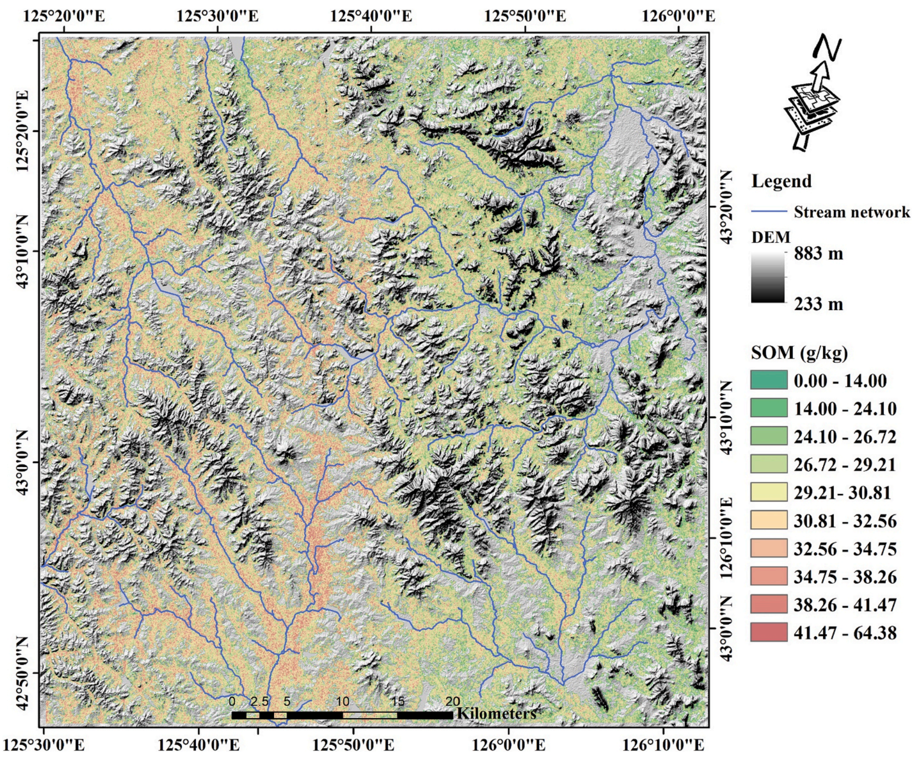

Fig. 8. Spatial distribution of the SOM contents in the study area.

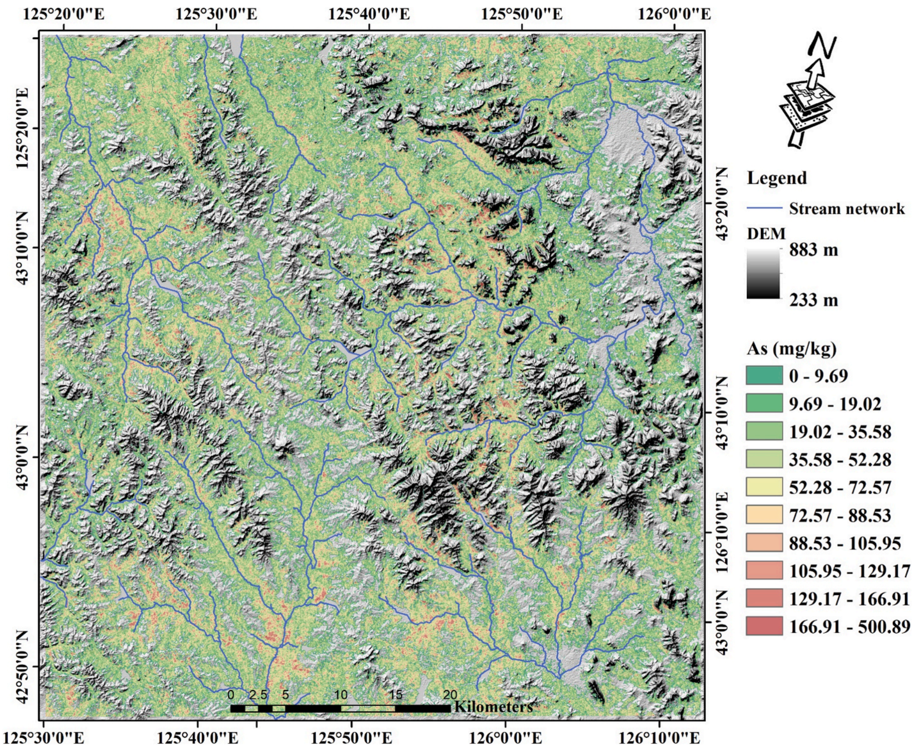

Fig. 9. Spatial distribution of the soil As contents in the study area.

analyzed soil parameters. Even though our sampling points represent only 1/38th of the entire map of the study area, the map clearly shows the distributions of the soil parameter contents in areas where there were no sampling points. The left part of the original image was less affected by clouds compared to the right part, particularly in the upper right part, which was partially masked, but some remnants were still discernible.

The SOM content distribution in Fig. 8 showed high SOM accumulation in areas where the stream networks flow. This finding can be explained by the fact that the stream network traverses a low-lying area, resulting in SOM accumulation and migration through soil erosion. There were greater SOM amounts in the low-lying and flat areas than in

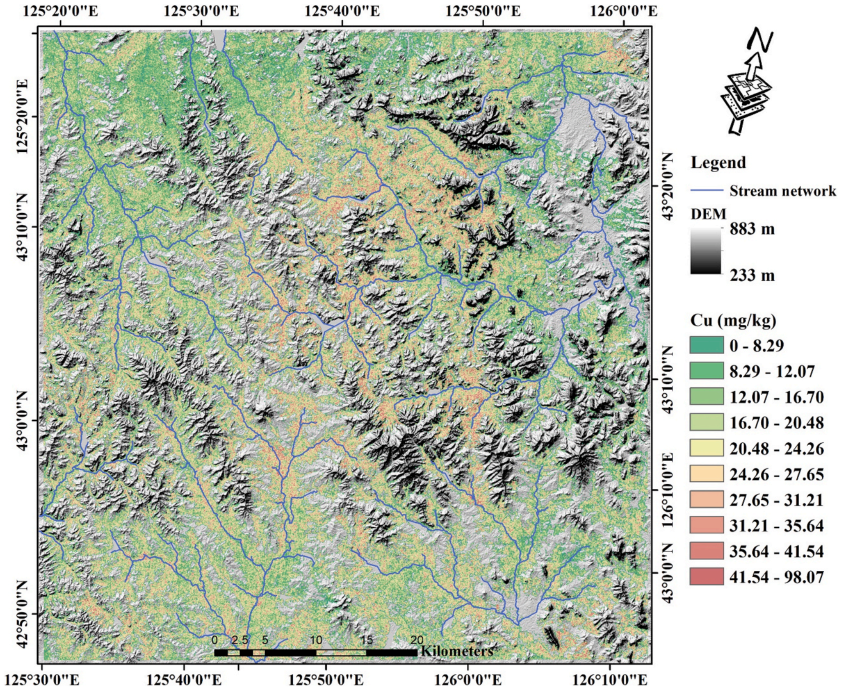

Fig. 10. Spatial distribution of the soil Cu contents in the study area.

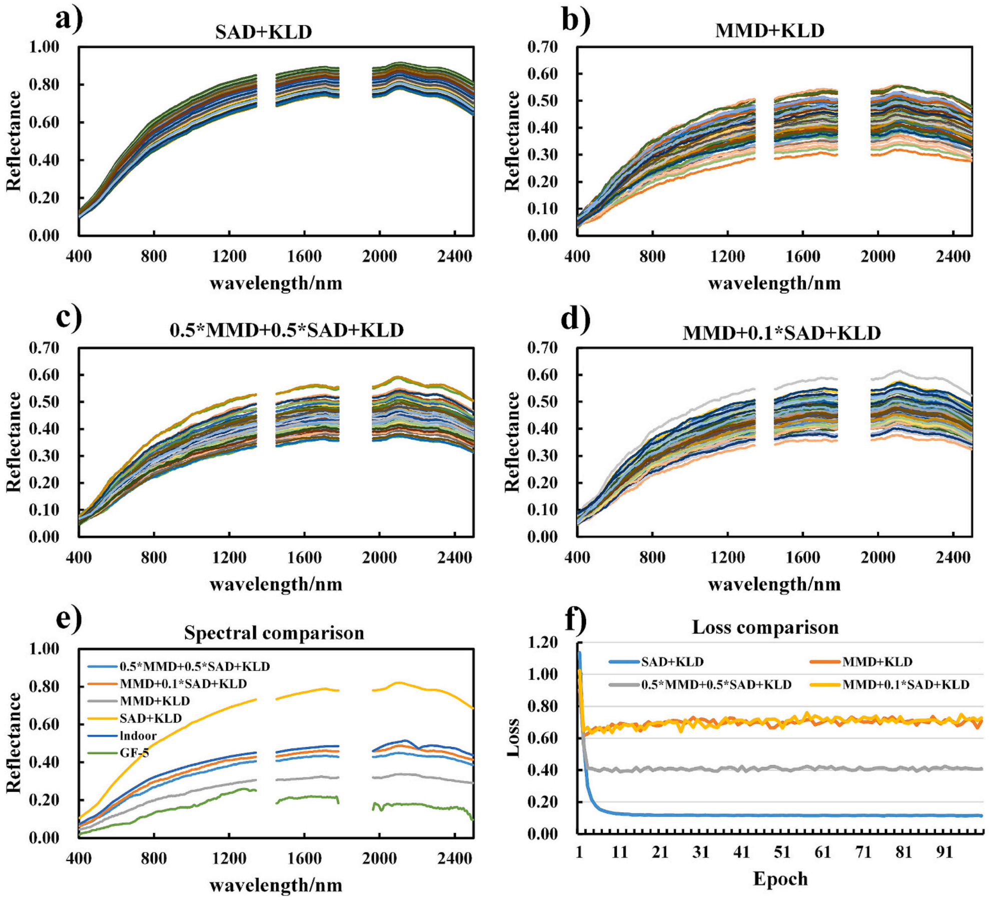

Fig. 11. Comparison of the generated spectra between different loss functions, namely SAD + KLD (a), MMD + KLD (b), 0.5*MMD + 0.5*SAD + KLD (c), and MMD + 0.1*SAD + KLD (d), as well as with the indoor spectra (e); GF5 spectra loss curve (f).

Table 3
Comparison of the proposed method with different feature models.

Soil	Features	Models	No. of Features	Training sub-dataset				Testing sub-dataset			
				R_c^2	RMSEc	MAEc	RPIQc	R_p^2	RMSEp	MAEp	RPIQp
SOM	Indoor	CARS + PLSR	51	0.99	0.63	0.51	11.89	**0.82**	2.60	1.95	2.90
	GF5		38	0.95	1.44	1.12	5.22	0.43	4.64	3.89	1.63
	VAE		51	0.98	1.01	0.85	7.43	**0.79**	2.84	1.97	2.66
	Mix		51	0.99	0.63	0.50	12.04	**0.88**	2.13	1.71	3.54
	Indoor	CARS + SVR	51	0.98	0.93	0.43	8.09	0.50	4.34	3.57	1.74
	GF5		38	0.96	1.27	0.60	5.93	0.49	4.38	3.39	1.72
	VAE		51	0.98	0.92	0.41	8.20	0.45	4.53	3.41	1.67
	Mix		51	0.94	1.60	0.76	4.71	**0.87**	2.21	1.77	3.41
As	Indoor	CARS + PLSR	34	0.98	9.55	7.32	3.95	**0.86**	24.09	18.45	1.23
	GF5		34	0.93	18.53	15.08	2.03	0.51	45.11	32.98	0.66
	VAE		38	0.97	13.06	9.59	2.89	0.65	38.41	29.22	0.77
	Mix		64	1.00	2.24	1.84	16.86	**0.80**	29.24	23.83	1.01
	Indoor	CARS + SVR	34	0.94	17.40	5.74	2.17	0.41	49.63	30.18	0.60
	GF5		34	0.70	37.99	11.61	0.99	0.33	52.82	32.57	0.56
	VAE		38	0.74	35.91	14.64	1.05	0.38	50.87	30.11	0.58
	Mix		64	0.96	14.34	4.14	2.63	**0.88**	21.98	15.32	1.35
Cu	Indoor	CARS + PLSR	62	0.99	1.09	0.81	5.32	**0.84**	4.72	3.67	1.15
	GF5		42	0.92	3.68	2.84	1.58	0.57	7.68	5.93	0.71
	VAE		56	0.90	4.09	3.48	1.42	0.54	7.89	6.00	0.69
	Mix		36	0.95	2.99	2.29	1.95	**0.85**	4.47	3.51	1.22
	Indoor	CARS + SVR	62	0.98	1.83	1.23	3.17	**0.79**	5.29	4.03	1.03
	GF5		42	0.85	5.07	1.99	1.15	0.60	7.34	5.43	0.74
	VAE		56	0.98	1.99	0.45	2.92	0.69	6.48	4.91	0.84
	Mix		36	0.94	3.29	1.46	1.77	**0.76**	5.75	4.16	0.95

other parts of the study area where small river channels are present. According to the field survey, erosion around the river in the study area was severe. In addition, most of the study area consisted of soil with high sand contents, explaining the low SOM contents, which is consistent with the SOM inversion results, suggesting high inversion accuracy. The results showed also an overlap of the As contents with the stream networks in the study area (Fig. 9). The main source of As in the study area is mining. The spatial distribution of the soil As contents was similar to that of SOM due to the complexation of As by dissolved organic matter (DOM), which also has a sorption effect (Haitzer et al., 1998; Zhang et al., 2021). The roots of rice cultivated in some regions may exhibit some effect on soil As contents through absorption. Overall, the study area showed a moderate pollution level due mainly to mining activities. On the other hand, the soil Cu content distribution was less consistent with the stream network distribution in the study area (Fig. 10). The highest soil Cu contents were found in the central part of the study area, which might be due to the influences of atmospheric factors, requiring more comprehensive field investigations. The inverted soil parameter contents in the eastern part of the study area were substantially higher than those in the other parts due to the greater effects of atmospheric factors in the eastern part of the study area.

5. Discussion

5.1. Effect of the loss function

Different loss functions significantly affect the generated spectra. In this study, different loss functions (SAD + KLD, MMD + KLD, 0.5*MMD + 0.5*SAD + KLD, and MMD + 0.1*SAD + KLD) were used to compare and analyze the generated spectra using our proposed model. The KLD was ignored in the comparative analysis, as it had little effect on the generated spectra in this study. Fig. 11 shows the generated spectra using different loss functions. As can be seen, the spectral values were generally relatively high, reaching over 0.9, but their waveforms were relatively good. Fig. 11b shows the spectra generated by the MMD + KLD function, indicating lower spectral values than those of the other generated spectra, thus the spectral features were lost. Fig. 11c and 11d show the results of the combined MMD and SAD loss functions with different parameter values. Although these combined loss functions were comparable and retained more features than those obtained using only the MMD loss function, there were inconsistencies in the range of the retained features.

Fig. 11e shows the results of the indoor and GF-5 spectra selected randomly. It can be observed that the reflectance of the MMD + 0.1*SAD + KLD and 0.5*MMD + 0.15*SAD + KLD loss functions were close to those of the indoor spectra. It should be noted that the MMD + 0.1*SAD + KLD loss function exhibited the highest degree of similarity with the indoor spectra. The MMD + KLD, on the other hand, exhibited more lost features, while the SAD + KLD retained more features, with obvious deviation from the indoor spectra. Therefore, the obtained demonstrated that the MMD + 0.1*SAD + KLD was the most optimal loss function in this study. Certainly, this combination can be flawed and does not guarantee consistent spectra with the indoor spectra, as the GF-5 spectra vary considerably. Fig. 11f demonstrates that the loss was stable at an epoch number of 10. For the stability of the generated spectra, 100 training iterations were performed to build the final model.

5.2. Comparative Experimental analysis of the inversion accuracy

In this study, a series of comparative analyses were performed to validate the features of the generated and mixed spectra. In total, four sources of spectra were selected for feature comparisons, namely indoor spectral (Indoor), GF-5 original (GF5), proposed model (VAE), and mixed spectra (Mix). The features of each spectrum type were first selected using the CARS algorithm, then PLSR and SVR were used to develop inverse models. Table 3 provides the comparison results of the proposed method with different features. Direct training with GF-5 data resulted in extremely inaccurate model results for the three soil parameters (SOM, Cu, and As). In contrast, the accuracy of direct inversion using indoor spectra was very high, as indoor spectra are not greatly affected by environmental factors-related disturbances compared to GF-5. The spectra generated by our proposed model showed slightly higher and lower inversion accuracies than those obtained using the GF-5 and indoor spectra, respectively. Indeed, the slightly lower accuracies of our proposed model are due to the fact that the generated spectra may remove some of the original features. On the other hand, the inversion accuracy of the mixed spectra was higher than that of the indoor spectra. This finding is due to the consideration of both indoor spectral features

Table 4
Comparison by different inversion models.

Methods	Soil	Training sub-dataset				Testing sub-dataset			
		R_c^2	RMSEc	MAEc	RPIQc	R_p^2	RMSEp	MAEp	RPIQp
FD-RF	**SOM**	0.71	3.54	2.72	2.25	0.69	3.19	2.23	2.07
(Zhou et al., 2021)	**As**	0.44	51.36	29.03	0.73	0.45	49.46	26.24	0.58
	Cu	0.39	8.42	5.02	0.68	0.40	12.15	5.29	0.45
DNN-CARS	**SOM**	0.57	4.17	3.30	1.84	0.55	4.08	3.07	1.51
(Wei et al., 2021)	**As**	0.73	35.44	24.86	0.98	0.56	33.60	23.15	0.85
	Cu	0.72	7.57	4.80	0.77	0.56	9.56	5.00	0.61

Table 5
Correlation of soil composition with topographic factors.

Soil Composition	Image Range	Topographic Factor		
		TWI	Slope	LS-factor
SOM	Whole Image	0.1041	**−0.2029**	−0.1509
	Left Half Image	0.1359	**−0.2388**	−0.1683
	Right Half Image	0.0594	−0.1490	−0.1125
As	Whole Image	0.0319	−0.0762	−0.0622
	Left Half Image	0.0638	−0.1309	−0.1146
	Right Half Image	−0.0034	−0.0267	−0.0198
Cu	Whole Image	0.0442	−0.1123	−0.0791
	Left Half Image	0.0890	**−0.1694**	−0.1190
	Right Half Image	−0.0080	−0.0574	−0.0444

and spectra generated by our proposed model, demonstrating the effectiveness of spectra data fusion in inverting SOM, Cu, and As contents in the soils. In fact, the inversion accuracies of the SOM, As, and Cu contents obtained using the mixed spectral feature model were 38, 55, and 28 % higher than those of the original spectral feature model, respectively.

Table 4 shows the inversion of soil composition using the FD-RF method proposed by Zhou et al. (2021) and the DNN-CARS method used by Wei et al. (2021) respectively. From the table, it can be seen that the FD-RF method has the highest accuracy for the inversion of SOM, and the R^2 values of test set can reach 0.69, while our proposed method is able to reach 0.87, which is 18 % higher. For As and Cu, relatively better results can be obtained by utilizing DNN-CARS, and the R^2 of the test set can reach 0.56 for both of them, while it is around 0.45 using FD-RF. However, our proposed method can achieve 0.88 for As and 0.85 for Cu, which are 32 % and 29 % higher, respectively. This means the high accuracy advantage of our method.

5.3. Topographic correlation analysis

Table 5 shows the correlation of soil composition with topographic factors. From the hyperspectral image, the right half of the image may be affected by thin clouds, so we used the whole image, the left half of the image and the right half of the image to obtain the relationship between topographic factors and soil composition content respectively. From the table, it can be seen that SOM has the highest correlation with slope, which can reach −0.24, while As and Cu show lower negative correlation, indicating that the surface fluctuation has a certain effect on the distribution of soil components. At the same time, it corroborates that the hydrodynamics of the study area plays a role in the transportation and accumulation of soil components. The correlation coefficient of the left half of the image is higher than that of the right half of the image, which indicates that the thin clouds have a certain effect on the soil composition mapping, so more consideration needs to be given to removing the effect of the thin clouds in the follow-up in order to ensure the accuracy of the mapping.

6. Conclusions

This paper proposes a VAE style-based spectral data fusion model to address the low inversion accuracy caused by the large time interval of the GF-5 image and soil sample data acquisition. Compared to the indoor and original GF-5 spectra, the correlations of the proposed model-generated spectra with three soil parameters (SOM, As, and Cu) were significantly enhanced. The results demonstrated that the proposed model generates indoor-like spectra with variable interclass spacing. According to the comparison results, the mixed spectral feature model showed higher SOM, As, and Cu inversion accuracies than those of the original spectra feature model by 38, 55, and 28 %, respectively. In addition, the inversion accuracies of the SOM, As, and Cu contents in the testing step showed R^2 values of 0.87, 0.88, and 0.85, respectively. The spatial distribution maps revealed that SOM and As contents were similarly concentrated in the lowlands, while the soil Cu contents were not detectable. Although the proposed model can better fuse indoor and GF-5 image spectra, it has some shortcomings. Furthermore, physical mechanisms should be considered in spectral generation in the future to prevent feature losses.

CRediT authorship contribution statement

Depin Ou: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Jie Li:** Writing – original draft, Visualization, Resources, Methodology, Investigation. **Zhifeng Wu:** Writing – original draft, Visualization, Supervision, Project administration, Methodology, Investigation, Data curation. **Kun Tan:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition. **Weibo Ma:** Visualization, Software, Funding acquisition, Formal analysis. **Xue Wang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Formal analysis, Data curation. **Yueqin Zhu:** Validation, Methodology, Investigation, 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.

Acknowledgment

This work is jointly funded by the National Natural Science Foundation of China (grant nos. 42201385, 41871337, 42307604) and the China Postdoctoral Science Foundation (grant no. 2022 M720866).

Data availability

Data will be made available on request.

References

Al-Abbas, A.H., Swain, P.H., Baumgardner, M.F., 1972. Relating organic matter and clay content to the multispectral radiance of soils. Soil Sci. 114, 477–485.

Arnold, J.G., Fohrer, N., 2005. SWAT2000: current capabilities and research opportunities in applied watershed modelling. Hydrological Processes: an International Journal 19, 563–572.

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.

Ben-Dor, E., 2002. Quantitative remote sensing of soil properties. Adv. Agron. 75, 173–243.

Ben-Dor, E., Chabrillat, S., Demattê, J., Thenkabail, P., Lyon, J., Huete, A., 2019. Characterization of soil properties using reflectance spectroscopy, hyperspectral remote sensing of vegetation. CRC Press, Boca Raton.

Berk, A., Anderson, G.P., Bernstein, L.S., Acharya, P.K., Dothe, H., Matthew, M.W., Adler-Golden, S.M., Chetwynd Jr, J.H., Richtsmeier, S.C., Pukall, B., 1999. MODTRAN4 radiative transfer modeling for atmospheric correction, Optical spectroscopic techniques and instrumentation for atmospheric and space research III. International Society for Optics and Photonics 348–353.

Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.-P., Schölkopf, B., Smola, A.J., 2006. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22, e49–e57.

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 251–252, 124–132.

Castaldi, F., Chabrillat, S., Jones, A., Vreys, K., Bomans, B., Van Wesemael, B., 2018. Soil organic carbon estimation in croplands by hyperspectral remote APEX data using the LUCAS topsoil database. Remote Sens. (basel) 10, 153.

Cha, J., Kim, K.S., Lee, S., 2019. On the transformation of latent space in autoencoders. arXiv preprint arXiv:1901.08479.

Chabrillat, S., Ben-Dor, E., Cierniewski, J., Gomez, C., Schmid, T., van Wesemael, B., 2019. Imaging spectroscopy for soil mapping and monitoring. Surv. Geophys. 40, 361–399.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. Noise Reduction in Speech Processing 1–4.

Conrad, O., 2006. SAGA—program structure and current state of implementation. SAGA–Analysis and Modelling Applications, edited by: Böhner, J., McCloy, KR, and Strobl, J., Göttinger Geographische Abhandlungen, Göttingen, 39-52.

Doersch, C., 2016. Tutorial on variational autoencoders. arXiv preprint arXiv: 1606.05908.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H., Liu, W., 2020. Advances of four machine learning methods for spatial data handling: a review. Journal of Geovisualization and Spatial Analysis 4, 13.

Ge, X., Ding, J., Teng, D., Xie, B., Zhang, X., Wang, J., Han, L., Bao, Q., Wang, J., 2022. Exploring the capability of Gaofen-5 hyperspectral data for assessing soil salinity risks. Int. J. Appl. Earth Obs. Geoinf. 112, 102969.

Gholizadeh, A., BorůVka, L., Saberioon, M.M., Kozák, J., Vašát, R., NěMeček, K., 2015. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. Soil Water Res. 10, 218–227.

Guanter, L., Richter, R., Moreno, J., 2006. Spectral calibration of hyperspectral imagery using atmospheric absorption features. Appl. Opt. 45, 2360–2370.

Haitzer, M., Höss, S., Traunspurger, W., Steinberg, C., 1998. Effects of dissolved organic matter (DOM) on the bioconcentration of organic chemicals in aquatic organisms—a review—. Chemosphere 37, 1335–1362.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lagacherie, P., Bailly, J.-S., Monestiez, P., Gomez, C., 2012. Using scattered hyperspectral imagery data to map the soil properties of a region. Eur. J. Soil Sci. 63, 110–119.

Lagacherie, P., Baret, F., Feret, J.-B., Netto, J.M., Robbez-Masson, J.M., 2008. Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. Remote Sens. Environ. 112, 825–835.

Logan, T.A., Nicoll, J., Laurencelle, J., Hogenson, K., Gens, R., Buechler, B., Barton, B., Shreve, W., Stern, T., Drew, L., 2014. Radiometrically terrain corrected ALOS PALSAR Data available from the Alaska Satellite Facility, AGU Fall Meeting Abstracts, pp. IN33B-3762.

Meng, X., Bao, Y., Luo, C., Zhang, X., Liu, H., 2024. SOC content of global Mollisols at a 30 m spatial resolution from 1984 to 2021 generated by the novel ML-CNN prediction model. Remote Sens. Environ. 300, 113911.

Ou, D., Tan, K., Lai, J., Jia, X., Wang, X., Chen, Y., Li, J., 2021. Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction. Geoderma 385, 114875.

Ou, D., Tan, K., Wang, X., Wu, Z., Li, J., Ding, J., 2022. Modified soil scattering coefficients for organic matter inversion based on Kubelka-Munk theory. Geoderma 418, 115845.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. Geoderma Reg. 16.

Phillips, T., Abdulla, W., 2022. Variational autoencoders for generating hyperspectral imaging honey adulteration data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 214–221.

Picard, D., 2021. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. arXiv preprint arXiv: 2109.08203.

Piekarczyk, J., Kaźmierowski, C., Królewicz, S., Cierniewski, J., 2016. Effects of soil surface roughness on soil reflectance measured in laboratory and outdoor conditions. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9, 827–834.

Qin, Y., Zhang, X., Zhao, Z., Li, Z., Yang, C., Huang, Q., 2021. Coupling relationship analysis of gold content using gaofen-5 (GF-5) satellite hyperspectral remote sensing data: a potential method in chahuazhai gold mining area, Qiubei County, SW China. Remote Sens. (basel) 14, 109.

Shi, T., Chen, Y., Liu, Y., Wu, G., 2014. Visible and near-infrared reflectance spectroscopy-An alternative for monitoring soil contamination by heavy metals. J. Hazard. Mater. 265, 166–176.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.

Sohn, Y., Rebello, N.S., 2002. Supervised and unsupervised spectral angle classifiers. Photogramm. Eng. Remote Sens. 68, 1271–1282.

Song, X., Wang, P., Van Zwieten, L., Bolan, N., Wang, H., Li, X., Cheng, K., Yang, Y., Wang, M., Liu, T., 2022. Towards a better understanding of the role of Fe cycling in soil for carbon stabilization and degradation. Carbon Research 1, 1–16.

Tan, K., Ma, W., Chen, L., Wang, H., Du, Q., Du, P., Yan, B., Liu, R., Li, H., 2021. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. J. Hazard. Mater. 401, 123288.

Tan, K., Wang, X., Niu, C., Wang, F., Du, P., Sun, D.-X., Yuan, J., Zhang, J., 2020. Vicarious calibration for the AHSI instrument of Gaofen-5 with reference to the CRCS Dunhuang test site. IEEE Trans. Geosci. Remote Sens. 59, 3409–3419.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr 46, 234–240.

Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma 367, 114208.

Viscarra Rossel, R., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54.

Wan, L., Zhou, W., He, Y., Wanger, T.C., Cen, H., 2022. Combining transfer learning and hyperspectral reflectance analysis to assess leaf nitrogen concentration across different plant species datasets. Remote Sens. Environ. 269, 112826.

Wang, F., Gao, J., Zha, Y., 2018. Hyperspectral sensing of heavy metals in soil and vegetation: feasibility and challenges. ISPRS J. Photogramm. Remote Sens. 136, 73–84.

Wang, X., Tan, K., Du, Q., Chen, Y., Du, P., 2019. Caps-tripleGAN: GAN-assisted CapsNet for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57, 7232–7245.

Wei, L., Zhang, Y., Lu, Q., Yuan, Z., Li, H., Huang, Q., 2021. Estimating the spatial distribution of soil total arsenic in the suspected contaminated area using UAV-Borne hyperspectral imagery and deep learning. Ecol. Ind. 133, 108384.

Wu, F., Wang, X., Liu, Z., Ding, J., Tan, K., Chen, Y., 2021. Assessment of heavy metal pollution in agricultural soil around a gold mining area in Yitong County, China, based on satellite hyperspectral imagery. J. Appl. Remote Sens. 15, 042613.

Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., Ma, H., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71, 918–926.

Yang, J., Huang, X., 2023. The 30 m annual land cover datasets and its dynamics in china from 1985 to 2022. 2023. Earth Syst. Sci. Data 3907–3925.

Yu, W., Zhang, M., Shen, Y., 2020. Spatial revising variational autoencoder-based feature extraction method for hyperspectral images. IEEE Trans. Geosci. Remote Sens. 59, 1410–1423.

Zhang, F., Li, X., Duan, L., Zhang, H., Gu, W., Yang, X., Li, J., He, S., Yu, J., Ren, M., 2021. Effect of different DOM components on arsenate complexation in natural water. Environ. Pollut. 270, 116221.

Zhou, W., Yang, H., Xie, L., Li, H., Huang, L., Zhao, Y., Yue, T., 2021. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. Catena 202, 105222.