Open Access

SCHÓLARENA

Journal of Forensic and Crime Studies

RESEARCH ARTICLE

ISSN: 2638-3578

A Preliminary Study of Geographic Traceability of Soil Physical Evidence: Machine Learning Recognition of Elemental Fingerprints and Morphological Features

Zhen Jia, Hongyuan He*, Genyuan Cui and Yujing Li

School of investigation, People's Public Security University of China, Beijing, 100038, China

*__Corresponding Author__: Hongyuan He, School of investigation, People's Public Security University of China, Beijing, 100038, China, Tel: +86-13311296819, E-mail: 13311296819@189.cn

Citation: Zhen Jia, Hongyuan He, Genyuan Cui, Yujing Li (2024) A Preliminary Study of Geographic Traceability of Soil Physical Evidence: Machine Learning Recognition of Elemental Fingerprints and Morphological Features, J Forensic Crime Stu 6: 101

Abstract

Soil is a type of physical evidence that is often found at crime scenes, but it is not efficiently utilized in the process of solving a crime. In this research, soil samples collected from 20 soil sites scattered in five cities of China were investigated to compare their application values and conditions. The traceability of soil physical evidence was analysed from two perspectives, namely soil composition and morphology.The elemental fingerprints of the samples were determined using inductively coupled plasma mass spectrometry (ICP-MS), and the samples were source classified by combining principal component analysis, partial least squares discriminant analysis, and support vector machine models. Images of the soil in the visible band were collected using a hyperspectral imaging system, and the samples were classified according to their source by combining deep learning models, including BPNN and CNN. Good classification results were obtained for both soil traceability methods based on both techniques. Among them, the ICP-MS method combined with the PLS model was able to achieve 100% classification accuracy for trace soil samples, but the experimental cost was high and the pre-processing process was complicated. Meanwhile, the hyperspectral method combined with the CNN model was able to achieve 99.19% fast and non-destructive identification, however, there was a high level of demand for soil testing. Notably, the two traceability methods can be applied to different occasions, and the selection of appropriate analytical detection methods in practical application will be conducive to the effective use of on-site soil physical evidence of the juries, providing assistance in the detection of cases and the breakthrough of difficult cases.

Keywords: Soil; ICP-MS; Hyper Spectral; Elemental Fingerprint; Machine Learning

Introduction

Soil is the environmental medium that people were most often exposed to in their lives .Soil evidence can often be found at crime scenes, for example, on the tyres of suspects' vehicles, on the soles of suspects' shoes, and on the tools used in committing the crime [1]. Once soil is attached to an object, its ability to adhere depends on the later activity of the subject, the nature of the soil itself, and the surface condition of the attached subject. However, the elemental fingerprints in soil, especially metal element fingerprints, remain relatively constant over time and space without interference from external factors [2]. Due to the wide variety of soil, their complex composition, and their poor uniformity, the test material is subject to environmental impacts [3,4]. Thus, the identification of soil evidence is quite difficult, and test results are generally not used independently as evidence but can improve the evidentiary value of the attached object, provide clues to breaches in security, and narrow down the scope of the investigation.

Current research on soil characteristics in the environment has focused on three areas: moisture content [5-7], organic matter content [8-10], and elemental fingerprinting [11-12]. For soil samples collected in the field, moisture varies considerably over time; organic matter also varies significantly due to small sample sizes [13]. Notably, elemental fingerprints had established a certain research foundation in the field of forensic soil analysis [14-15]. In one case [16], the investigators used scanning electron microscopy-energy spectrometry (SEM-EDS) and X-ray fluorescence spectroscopy (XRF) to detect soil on the tools of a suspected grave robber and soil in the robbed grave, then identified the same soil sample elements from both sources, thereby providing strong evidence to identify the suspect.

Inductively coupled plasma mass spectrometry (ICP-MS) has the advantages of simple sample preparation and injection techniques, fast mass scanning, short duty cycle, and low interference with the ion information provided. Its detection sensitivity and accuracy are higher than those of SEM-EDS and XRF [17-18]. It has extremely low detection limits for most elements and is recognised as the most ideal method for inorganic elemental analysis.

High standard precision lossless spectral imaging is a new type of high-speed lossless spectral imaging technology, which effectively combines mechanical vision spatial imaging and spectral analysis technology to obtain higher resolution, wider band spectral images, and richer spectral data information [19-20]. In the field of soil testing, hyperspectral imaging technology is widely used,such as in estimating soil water content [39], inversion of soil organic matter content [40] and forensic chemical analysis of soil [41-42], while hyperspectral remote sensing techniques [43-45], are often used for hyperspectral collection of soils over large areas. There have been no reports of traceability analyses of soil physical evidence combining both morphological and elemental perspectives.

Machine learning is an analytical method that has often been used in recent years in the field of intelligent identification of physical and chemical evidence; it can quickly and accurately identify large amounts of data and is widely used in the analysis of soil hyperspectral image data [46]. A total of 96 ground soil samples were collected through hyperspectral acquisition by Tahmasbian [47], who developed a partial least squares regression model to correlate the total carbon (TC), total nitrogen (TN) ,δ 13 C, and δ 15 N values obtained from isotope ratio mass spectrometry with soil reflectance spectra. Wu [48] used a visible-NIR hyperspectrometer to collect 140 in situ hyperspectral images of soil profiles and compared the predictive ability of partial least squares regression and partial least squares-support vector machine (LS-SVM) models for soil salinity content. With the advancements in computer science, deep learning represented by convolutional neural networks (CNNs) has gradually developed, extracting features layer by layer through convolution and pooling. In addition, it possesses the characteristics of weight sharing and local connectivity, which can train large-scale data more effectively than traditional machine learning models [49]. Meanwhile, Riese [50] used five CNN models that are available on GitHub to target the soil hyperspectral data in the LUCAS dataset and showed that the shallow CNN outperformed the ResNet network and CoordConv network in addressing this problem. The shallow CNN was found to be better than the residual neural network with complex structure and parameters in one-dimensional data processing.

In this study, soil samples were collected from 20 sampling points in five cities of China, and multi-elemental analysis and spectral information were collected using ICP-MS and hyperspectral imaging, respectively, to establish the best machine learning identification models for different data types. Finally, the two methods were evaluated for traceability of the soil evidence, providing new ideas for the use of soil evidence in investigative practice. The research process was shown in Figure 1.

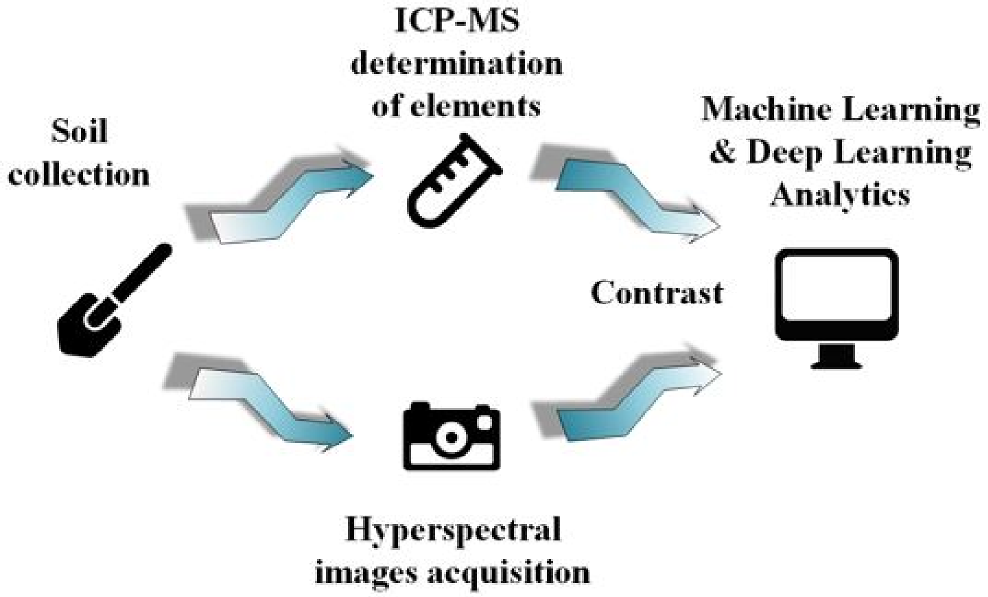
ICP-MS determination of elements
Soil collection
Machine Learning & Deep Learning Analytics
Contrast
Hyperspectral images acquisition

Figure1: Research process for Soil physical evidence

Materials and Methods

Apparatus and Reagents

Apparatus

The main instruments used in this paper were 7800 Inductively Coupled Plasma Mass Spectrometer (ICPMS) produced by Agilent Technologies, EPMA-1720 Electron Probe Mass Spectrometer (EPMS) produced by Shimadzu, DKQ-1800 Intelligent Temperature Controlled Electric Heater (DKQ-1800) produced by Shanghai Yi Yaoyao Instrument Science and Technology Development Co. Ltd.; Hyperspectral imaging system SEC-E1100 (Wayho Technology Co.); and Synergy ultrapure water system from Merck Millipore. The specific parameters of the instrument are shown in Table 1.2.

Table1: ICP-MS Instrument Operating Parameters

Projects	Parameters	Projects	Parameters
Radio frequency power	1500 W	Cooling gas flow	15.0 L·min^{-1}
Carrier gas flow	1.0 L·min^{-1}	Auxiliary gas flow	1.0·L^{-1}
Sampling depth	0.085 metres	Analysis model	Collision reaction cell
Repetition rate	3 times	Acquisition Mode	Peak-hopping acquisition

Table2: Hyperspectral imaging system Operating Parameters

Product model	**SEC-E1200**
Spectral range	450-950nm
LCTF scanning accuracy	1nm
LCTF half-height width (FWHM)	10nm@550nm
LCTF response time	10-200ms
LCTF field angle	±5°
Image sensor type	Backlit scientific CMOS
Image sensor size	1.2 inches
Image sensor pixel size	6.5μm×6.5μm
Image resolution	2448×2048

Reagents

UPS grade concentrated nitric acid (68 %) (Suzhou Jingrui Chemical Co., Ltd); analytical pure hydrofluoric acid (40 %) (Merck Millipore, Germany); high purity helium He (≥99.999 %) (Henan Yuanzheng Special Gases Co., Ltd); high purity argon Ar (≥99.999 %) (Henan Yuanzheng Special Gases Co., Ltd); ICP-MS mixed internal standard stock solution (Agilent Technologies Co.) ICP-MS multi-element mixed standards (Agilent Technologies Ltd.)

Experimental Methods

Soil Sampling and Pre-Treatment

A total of 100 soil observation samples were obtained using a five-point sampling method. The sampling method is shown in Figure 2. Each sample was repeated three times during testing.

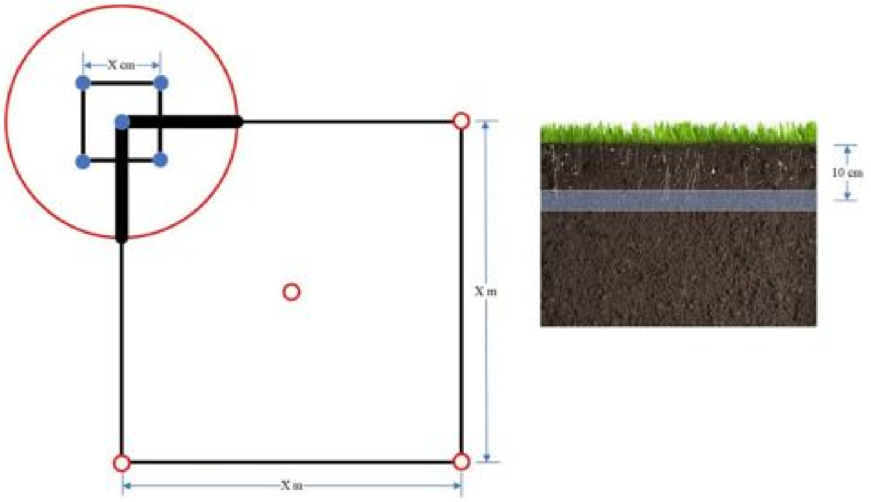

Figure 2: Sampling method (Five-point sampling of the four corners and center of the square)

Five soil samples were taken from each sampling site using the five-point sampling method; each sample was obtained by mixing five soil samples from a smaller area (the five-point sampling method was still followed in the smaller area).Where the sampling spacing X depends on the type of sampling points and the total area.

For example, in parks, the X value can be 0.1~0.2 meters, while in flower beds, the X value is set to 0.2~0.5 meters. The sampling depth was 0.1 meters, and each portion was about 5kg.

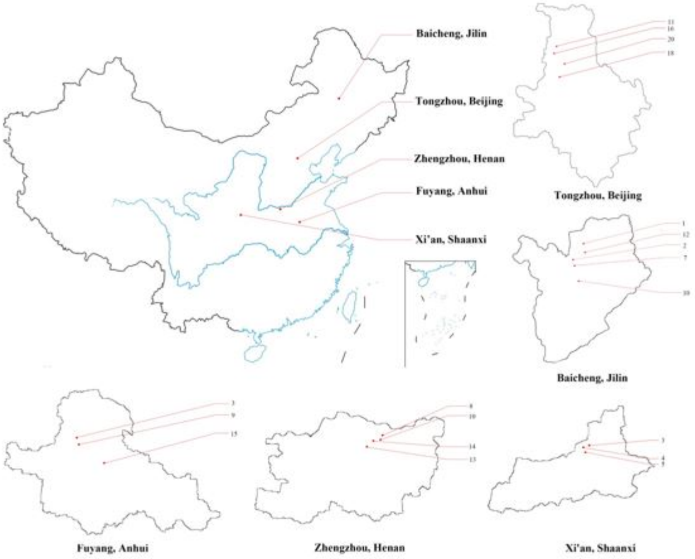

Figure 3: Sampling cities and sampling points

The sampling cities and sampling points are shown in Figure 3, and the source information and numbers of the soil samples are shown in Table 3. The soil was dried overnight in a blast drying oven at 80°C and passed through a 50-mesh sieve to remove plant and animal residues and large stony soil. A image of the soil samples from the 20 sampling points was taken under a 200x scanning electron microscope as shown in Figure 4. Observe for the presence of plant and animal remains and other impurities.

Table3: Soil sample number and source

Sample No	City	Sampling sites	Location details	Sampling quality(/kg)
1	Baicheng	Olympic Sports Park	Artificial garden	8.4
2	Baicheng	Municipal Public Security Bureau	Artificial garden	7.2
3	Fuyang	Fortune Plaza, Taihe County	Artificial garden	5.3
4	Xi'an	Xi'an Secondary School	Artificial garden	10.6
5	Xi'an	Bell Tower Flower Bed	Artificial garden	11.8
6	Xi'an	Kuo Du Police Station	Artificial garden	6.9
7	Baicheng	First School Garden	Artificial garden	5.3
8	Zhengzhou	Long Lake Park	Artificial garden	7.7
9	Fuyang	Taihe Family Community	Natural soil	7.6
10	Zhengzhou	Purple Mountain Road flower bed	Natural soil	8.4
11	Tongzhou	Golden Jubilee Garden	Natural soil	9.3
12	Baicheng	Bookish Blue County	Natural soil	6.2
13	Zhengzhou	Forest Park	Natural soil	5.5
14	Zhengzhou	People's Park	Artificial garden	5.9
15	Fuyang	Tai Wo Hospital	Natural soil	8.2
16	Tongzhou	Luhe Hospital	Natural soil	8

17	Baicheng	South Lake Park	Artificial garden	12.4
18	Tongzhou	Canal Secondary School	Artificial garden	7.5
19	Baicheng	Hanlin Academy	Natural soil	5.9
20	Tongzhou	Xihaizi Park	Natural soil	7.1

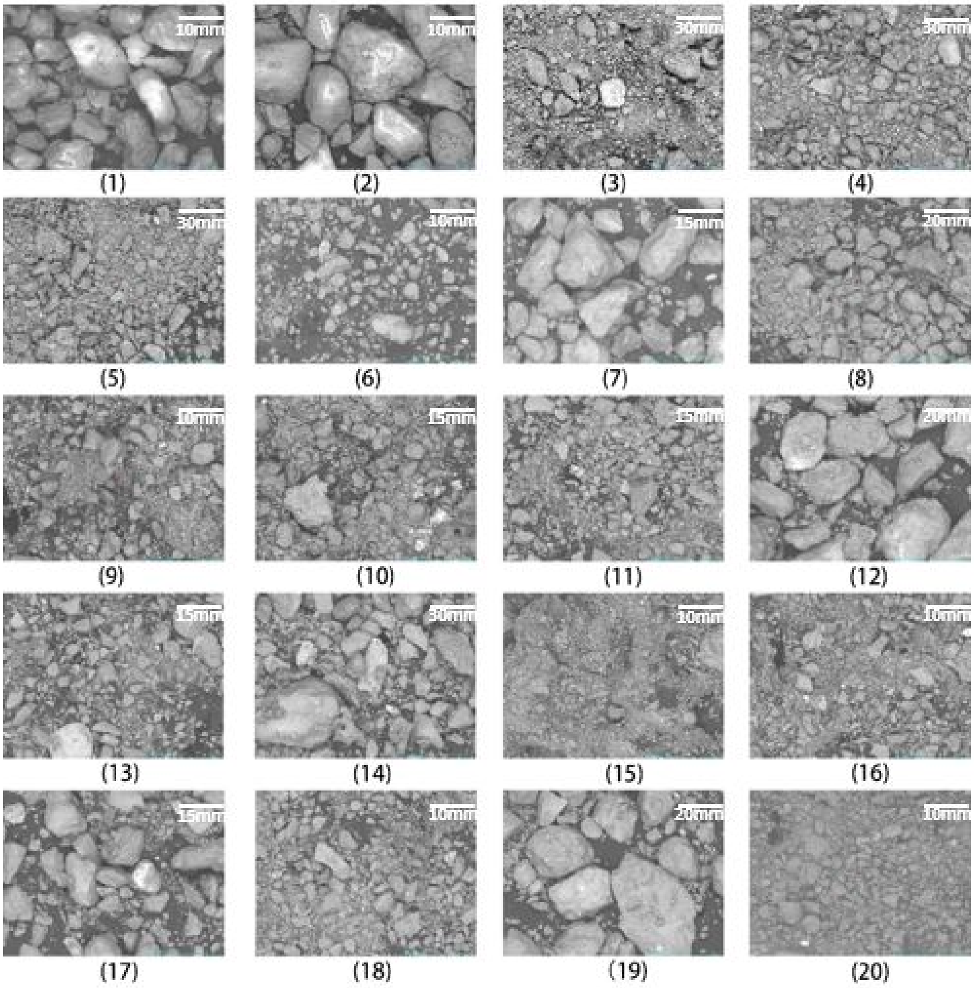

Figure 4: Soil quality of 20 sampling sites × 200 SEM

ICP-MS Multi-Element Determination and Data Analysis Methods

Weigh 0.25 g of sample accurately, add 5 mL of 65 % nitric acid and 2 mL of 40 % hydrofluoric acid and carry out microwave digestion. The microwave digestion procedure refered to the instrument operating instructions is shown in Table 4.

Table4: Microwave digestion procedure

Steps	**Temperature（℃）**	**pressure（atm）**	**Holding time（min）**
1	120	15	5
2	150	20	5
3	180	30	10
4	190	35	20

The digested samples were placed on an electric heating plate at 160 °C for 2 h to drive out the acid until the liquid volume was 1~2 mL. The samples were then diluted with ultrapure water and fixed to 50 mL.

Eighteen elements and isotopes of Na, Al, ^{43}Ca, ^{44}Ca, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Ag, Cd, Ba and ^{206}Pb, ^{207}Pb, ^{208}Pb were selected as the elements and isotopes to be measured, and a 5% HNO_3 solution was diluted to 0.5, 1, 5, 10, 20, 50, 200 and 500 μg/L as the standard solution.

The internal standard elements were selected as 45Sc, 72Ge, 103Rh and 209Bi. 45Sc was selected for elements with mass numbers 23-52; 72Ge for elements with mass numbers 55-75; 103Rh for elements with mass numbers 107-137; and 209Bi for elements with mass numbers 206-208.

Principal component analysis was used to downscale and visualize the data, converting linearly correlated data into a set of linearly uncorrelated or less linearly related data by orthogonal transformation. For elemental fingerprinting of soil samples, there is often a linear correlation between the content of multiple elements. Lionel [51] found a prominent correlation between Cr, Ni and clay when they analyzed the spatial distribution and sources of six trace elements (As, Cd, Cu, Cr, Ni, Pb) in soils from south-western France. Zhang [52] surveyed the spatial distribution and sources of six trace elements in the Chinese Loess Plateau. The study showed that soil elements in different sub-basins contain important correlation information. For example, Ca is highly homologous with Fe in two sub-basins, and Cu, K, Mg, Mn, Na and Zn are of similar origin in individual sub-basin soils. Principal component analysis can therefore be used to effectively resolve the correlations between soil elements and visualize the differences and similarities between the different samples.

Partial least squares discriminant analysis and support vector machine models (SVMs) were used to classify and identify the elemental fingerprints of soil samples. Support vector machine (SVM) is an optimized discriminant model that attempts to generate optimal hyperplanes or decision boundaries in a high-dimensional space to best distinguish between different classes. These hyperplanes are often constructed by analyzing the data points that are most likely to be misclassified (i.e. those in the vicinity of the pre-optimization hyperplane). These support vectors are iteratively weighted in the training phase to obtain distances (i.e. margins) that maximize the distance between classes separated by hyperplanes [53]. Based on the sample size, Linear, quadratic and cubic polynomials were chosen as kernel functions for the training of the model in this experiment. Bayesian optimization and cross-validation were used to optimize the penalty coefficient (C) and the regularization parameter (ε) of the model The partial least squares discriminant combines the advantages of principal component analysis and multivariate linear discriminant analysis and can effectively perform regression or classification analysis on independent variables with multiple correlations. Principal component analysis is used to obtain several main factors that contain information about the original data and are not correlated with each other, and multiple regression analysis is done based on the factors obtained by dimensionality reduction [54]. The number of principal components was traversed to determine the principal components to be retained for the purpose of obtaining the best classification effectiveness and the minimum number of factors. Principal component analysis, support vector machines, and partial least squares models were implemented by MATLAB 2019b software

Hyperspectral Image Acquisition and Processing Methods

Weigh about 2.0g of the sieved soil and disperse it in the center of the hyperspectral image acquisition unit with a uniform white liner in the background. The hyperspectral unit uses four 50W halogen lamps to illuminate the soil sample uniformly at an angle of 45°. The focal length of the image acquisition unit is adjusted via the control console until the sample soil is clearly visible at 700nm. A matching white board is used for correction prior to acquisition. A total of 101 images of the soil at different wavelengths were acquired by selecting the visible light range from 400nm to 900nm and acquiring every 5nm. The reflectance values of the observed points in each soil image were finally calculated from the image information of the whiteboard to obtain the soil reflectance spectral images.

The acquired hyperspectral images are also pixel fused to reduce noise interference and analysis load before data analysis is carried out. The 10*10 spectral summation strategy is used to average the 10*10 pixel size image segments after removing background information from the images using an algorithm. The pixel fusion strategy refers to previous studies by Wang [55] and others.

Hyperspectral images of soil samples were classified and identified using partial least squares discriminant analysis, back propagation neural network and convolutional neural network. The data volume of hyperspectral image observation samples is large and common machine learning models tend to fail for large-scale training samples [56]. Therefore, an attempt was made to use currently popular deep learning models, including feedforward neural networks (BPNN) and convolutional neural networks (CNN), for the recognition of hyperspectral image data.

Software

Principal component analysis, support vector machine model and partial least squares model were implemented by MATLAB 2019b software; BPNN and CNN deep learning models were implemented based on python 3.7.12 and Tensorflow 2.8.0.

Results and Discussion

Elemental Fingerprint Acquisition

Standard Curve Fitting

A regression curve was established with the calculated concentration as the horizontal coordinate and the response intensity ratio (corresponding intensity of the element to be measured / corresponding intensity of element He) as the vertical coordinate, and the overall effect was improved. The linear equations and correlation coefficients are shown in Table 5.

Table5: Fitting results of standard curve

Element	Linear equation	Correlation coefficient(r)
Na	y=0.0014*x-0.0050	0.9983
Al	y=1.3136E-004*x+2.4250E-004	0.9997
^{43}Ca	y=2.5198E-005*x+2.0085E-005	0.9998
^{44}Ca	y=2.4502E-004*x+6.1830E-004	1

Multi-Element Determination

A total of 100 soil samples from 20 sampling sites were pre-treated and injected for the determination of 18 elements and isotopes in the samples, with each sample measured three times with a relative standard deviation of less than 15 %. The results showed that the elemental fingerprints varied considerably between the 20 sampling sites. The elemental determinations of five parallel samples from the 20 sampling points were averaged and normalized to obtain the relative elemental concentrations of all soil samples. The relative elemental concentrations eliminate errors in the processing of the samples and provide a visual representation of the relative abundance of different elements between samples. A heat map was plotted using the relative elemental concentrations, as shown in Figure 5, where the higher the relative elemental concentration, the darker the color. As can be seen from the graph, there is a large variation in the elemental fingerprints of the 20 sampling points.

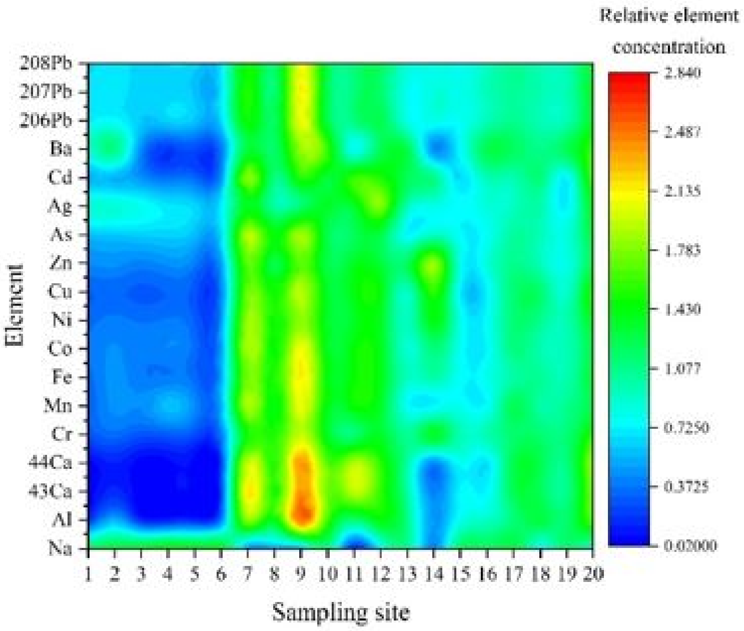

Figure 5: Relative element concentration heat diagram

Elemental Fingerprinting Analysis

Visualization Analysis

To visualize the differences in the elemental fingerprints of the soil samples between different sampling points, the information on the content of the 18 elements and isotopes was downscaled using principal component analysis, and the results showed that the first three principal components were able to retain 93.26% of all the information; the first five principal components were able to retain 97.56% of all the information. The loadings of the first three principal components are shown in Table 6, and the cumulative contributions of the first five principal components are shown in Table 7.

Table6: Loads of the first three principal components

Element	Principle component 1	Principle component 2	Principle component 3
Na	-0.1963	-0.4863	0.0184
Al	0.2375	-0.2389	0.0128
^{43}Ca	0.2403	-0.0698	0.0929
^{44}Ca	0.2397	-0.0898	0.0781
Cr	0.2296	0.1326	-0.011
Mn	0.239	-0.0833	-0.1156
Fe	0.2484	0.0302	-0.0502
Co	0.2461	0.0716	-0.1053
Ni	0.2398	0.2156	-0.009
Cu	0.2396	0.2228	-0.0089
Zn	0.2151	0.3951	-0.0409
As	0.2404	-0.0294	-0.1481
Ag	0.1677	0.0124	0.798

Cd	0.2307	0.1769	0.2692
Ba	0.1739	-0.4904	0.3055
^{206}Pb	0.2392	-0.1704	-0.171
^{207}Pb	0.2363	-0.1987	-0.2009
^{208}Pb	0.2391	-0.1867	-0.17

Table7: Cumulative contribution rate of the first five principal components

	PC score	**Contribution rate /%**	**Cumulative contribution rate /%**
PC 1	15.9308	83.85	83.85
PC 2	1.1433	6.02	89.86
PC 3	0.7145	3.76	93.62
PC 4	0.4657	2.45	96.07
PC 5	0.2827	1.49	97.56

PC=principal component

Similar loadings exist between the two isotopes of calcium and the three isotopes of lead, for the first principal component the element sodium has opposite loadings to the other elements, indicating that the distribution of the element sodium differs significantly from the other elements. The first three principal components were selected to plot the three-dimensional distribution, as shown in Figure 6. There are differences in the elemental fingerprints between the 20 sampling points and most of the samples are well clustered, but there is confusion between samples from certain sampling points. For example, No. 6, No. 12 and No. 19 are marked in red circles. Continued use of machine learning model analysis may improve the objectivity and accuracy of the classification.

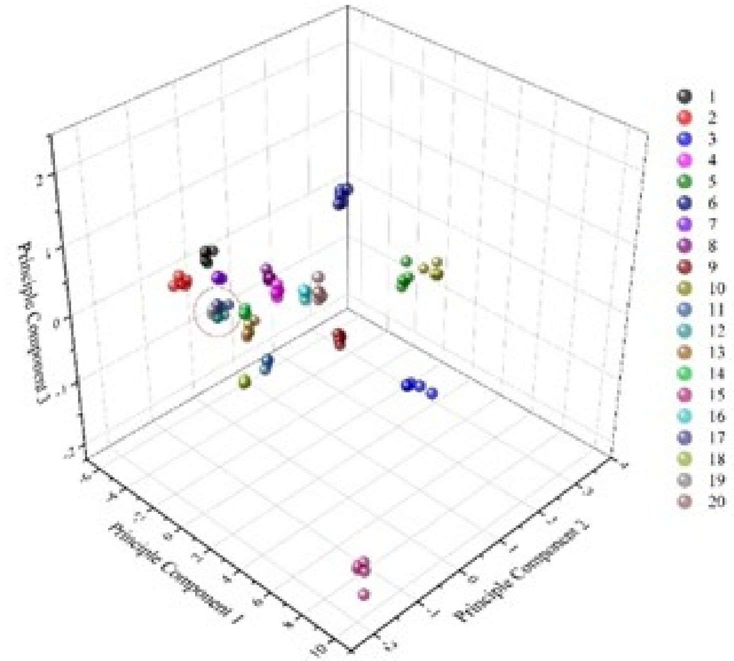

Figure 6: Element fingerprint visualization between sampling points

To explore whether the differences in elemental fingerprints between the 20 sampling points were related to geographical location, we set the labels to five cities and visualized the results as shown in Figure 7(a). When the samples were divided based on the city labels, the soils from the five city sources could not be distinguished well. It indicates that the differences between the samples do not lie in the differences between the cities, so further analyses are required. Afterwards, we excluded the macronutrients and only retained the eight heavy metal elements and isotopes of chromium (Cr), cobalt (Co), arsenic (As), silver (Ag), cadmium (Cd) and lead (^{206}Pb, ^{207}Pb, ^{208}Pb) for analysis again. The results of the downscaling and visualization are shown in Figure 7(b), where sample numbers 1, 2, 3, 4 and 5 represent soil samples from Baicheng, Jilin; Fuyang, Anhui; Xi'an, Shaanxi; Zhengzhou, Henan and Tongzhou, Beijing, respectively.

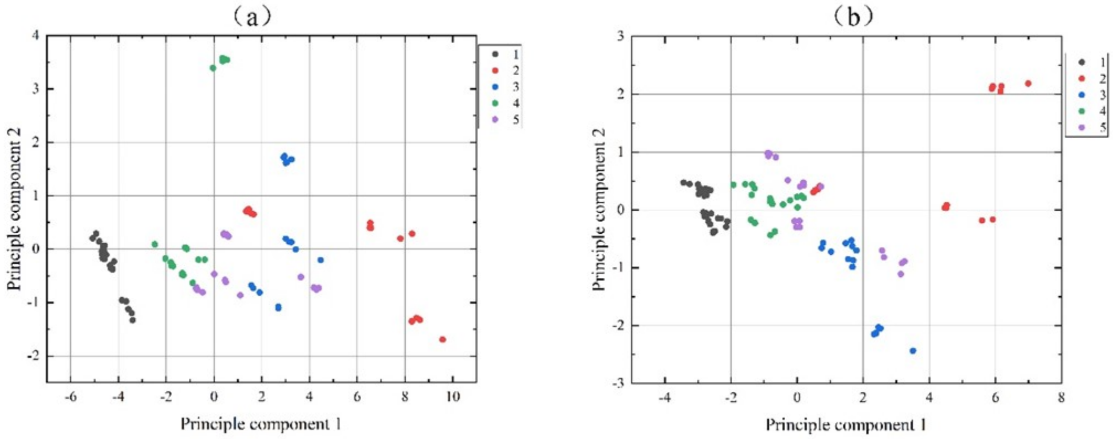

Figure 7: Element fingerprint visualization between cities

(a) Visual distribution of 18 soil elements by two-dimensional principal component analysis; (b) visual distribution of eight heavy metal elements by two-dimensional principal component.Baicheng

Compared with the full range of soil elements, the visual classification results of the screened heavy metal elements showed a large improvement, especially the samples from Baicheng achieved significant clustering. A small number of samples from Xi'an, Zhengzhou and Tongzhou were difficult to distinguish (with the methodology of this study). Based on the above analysis, the subsequent use of machine learning models to distinguish the urban origin of soils may be unreliable, which also illustrates the consistency of soil elemental composition at small scales and the randomness at large scales.

Element Fingerprinting

The 100 samples were divided into training and test sets in a 4:1 ratio randomly, and the SVM models with linear functions, quadratic polynomials and cubic polynomials as kernel functions all had a test set accuracy of 99%, with only one sample in sampling point 6 having a mismatch.

The results show that it is feasible to use all 18 elements and isotopes as features to identify fine location sources of soil evidence, and very high identification accuracy can be achieved using the support vector machine model. By disrupting the sampling point information labels and using the support vector machine model again for identification, the results showed that the identification accuracy of the three support vector machine models dropped significantly to 28%, 16% and 14% respectively, indicating that the method can effectively identify the sampling point sources of soil based on the elemental fingerprint features.

This was followed by training and identification with a partial least squares model based on the same data. The first step was to select the optimal number of retained components, so the effect of the number of principal components on the classification results was traversed from 1 to 18, as shown in Figure 8. The higher the number of retained principal components, the better the identification results. When 14 principal components are retained, the recognition accuracy has reached 100%. This indicates that all 18 elements have a significant influence on soil classification, which is consistent with the results of the factor analysis. The results show that the PLS-DA classification is better than the support vector machine model and can achieve complete classification of soil elemental fingerprints.

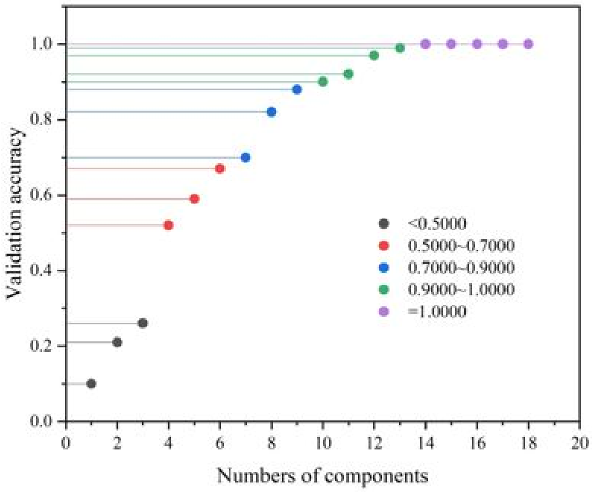

Figure 8: Importance of principal component variables

Hyperspectral Image Analysis

Model Development

For hyperspectral image data format characteristics, the latest deep learning methods, including feedforward neural networks and convolutional neural networks, were selected for model construction and optimization, and compared with the traditional spectral data analysis model of partial least squares discriminant. All hyperspectral images were acquired, background removed, and pixel fused to obtain a total of 392,250 observation samples, which were divided into training set, validation set and test set in the ratio of 7:1.5:1.5. The performance parameters of all model development sessions were optimized based on the results of the validation set.

Each observation sample contains 101 data points, and the feature factors obtained by using partial least squares to reduce the dimensionality are the most influential metrics on the classification results. Therefore, we evaluated the impact of the number of factors after dimensionality reduction on the chase removal amount of the validation set and selected the main adult score with relatively high recognition accuracy and relatively few retained factors. The optimization results show that the validation set model accuracy is gradually increasing with the increase of retained master fractions and remains stable at around 86%. The results of the PLS-DA model parameters evaluation are shown in Figure 9(a).

BPNN is one of the most fundamental network models in deep learning. During the optimization process, the number of fully connected layers, the batch size for training, and the choice of activation function have a great impact on the performance of the final model. The number of fully connected layers will affect the depth and number of parameters of the overall deep learning model, and for complex learning objects, deeper fully connected layers are often required to achieve this. For 1*101 spectral data, a shallow neural network is already sufficient for feature learning and extraction. Boosting the number of neural network layers may lead to situations such as gradient disappearance. The activation functions of classification models mainly include Tanh and Relu. Tanh can retain more information of the original parameters and is more suitable for refinement differentiation of small-scale similar samples; Relu function can significantly improve the training speed and mitigate the negative impact of gradient disappearance in deeper networks [54]. The batch size of training also has an important impact on model training accuracy. In general, as the batch size increases, the time required for one Epoch training decreases, while the Epoch required to train to the same accuracy increases accordingly. The number of Epochs was fixed in order to evaluate the performance of different model methods, so that a large batch size would reduce the training time and a small batch size would improve the training accuracy. The results of the feedforward neural network model optimization are shown in Figure 9(b). CNNs are the most important deep learning models developed in recent years and are widely used in image processing and are active in the field of artificial intelligence algorithms as the basis for several derivative algorithms. Advanced convolutional neural network models such as VGGNet and ResNet outperform the traditional LeNet for recognition of complex 2D and 3D objects, but shallow LeNet still has the best model performance in modelling thought spectral data. In this study, a three-layer one-dimensional convolutional neural network was constructed and the model parameters (number of convolutional layers, number of fully connected layers, activation function, etc.) were optimized, and the results are shown in Figure 9(c). The finalized feedforward neural network and convolutional neural network model parameters are shown in Table 8.

Table8: Hyperparameters of CNN model

Layer	Input shape	Output shape	Filter setting
Conv1d	(101,1)	(51,32)	Size=3;Strides=2
MaxPool1d	(51,32)	(26,32)	Size=3;Strides=2
Conv1d	(26,32)	(26,64)	Size=3;Strides=2
MaxPool1d	(26,64)	(13,64)	Size=3;Strides=2
Conv1d	(13,64)	(7,128)	Size=3;Strides=1
MaxPool1d	(7.128)	(4,128)	Size=3;Strides=2
Flatten	(4,128)	512	-
Dense	512	256	-
Dense	256	18	-
Activation(Softmax)	18	18	-

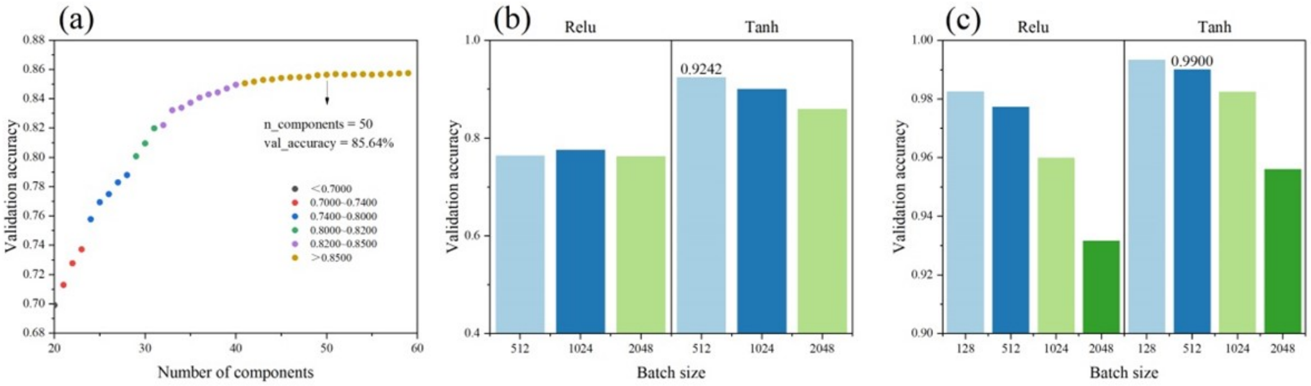

Figure 9: Optimization of three models.(a) PLS-DA;(b) BPNN;(c)CNN

Hyper Spectral Image Recognition

Hyperspectral images were recognized using the optimized model. the PLS model took the shortest time from training to recognition, but the final validation set accuracy was only 85.25% with the inclusion of 100 principal components. This indicates that the PLS model is no longer able to achieve accurate classification under the condition of large sample size. Therefore, deep learning methods were used to further explore the spectral data features. The changes in accuracy and loss function during the training of the feedforward neural network model and the convolutional neural network model are shown in Figure 10. As can be seen from the figure, the training process of the CNN model was smoother, both in terms of loss function convergence and recognition accuracy improvement. At around round 80, the BPNN model showed a serious oscillation, which was intermittent and irregular, caused by local anomalous data, and seriously affected the training efficiency. In terms of the training process, the CNN significantly outperformed the BPNN.

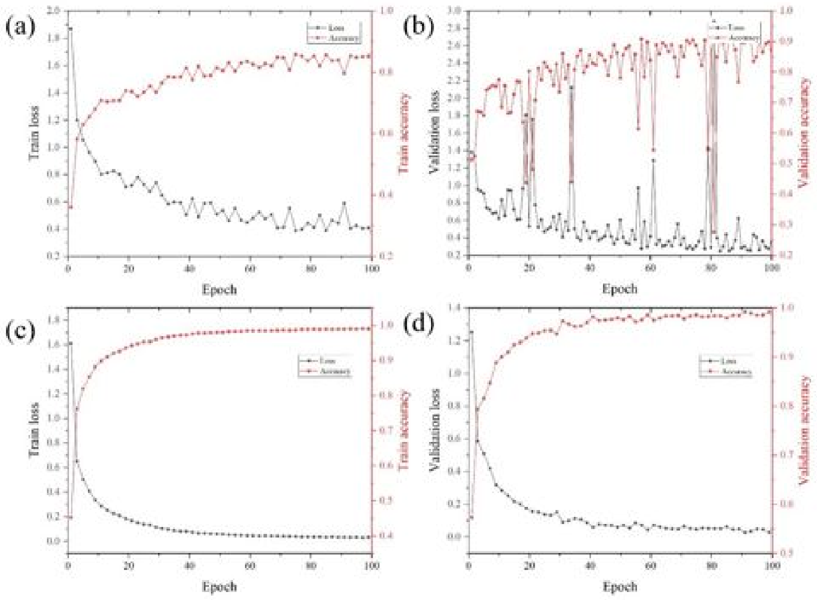

Figure 10: Training process.(a) is the accuracy and loss function of the training set of the BPNN during training; (b) is the accuracy and loss function of the test set of the BPNN during training; (c) is the accuracy and loss function of the CNN in the training set during training; (d) is the accuracy and loss function of the test set of the CNN during training

After training, the test set data was substituted into the three models for prediction. The results showed a final recognition accuracy of 85.25% for the PLS-DA model, 91.65% for the BPNN model and 99.19% for the CNN model. The evaluation parameters for the classification results of the three models are shown in Table 6, and the confusion matrix of the results is shown in Figure 11. F1 score balances the two metrics of recall and accuracy, which is the most used evaluation metric in machine learning classification problems. The F1 score of the CNN model reached 0.9908, which was able to distinguish significantly between the different soil hyperspectral images between the 20 sampling points. As shown in the confusion matrix, more and darker red parts represent more misclassified samples. It is clear that the PLS-DA model classification results have the most pronounced confusion matrix in red, followed by BPNN, and the CNN images have the least amount of red parts.

Table9: Model evaluation

Model	**Accuracy/%**	**Precision**	**Recall**	**F1-score**
PLS-DA	85.25	0.8629	0.839	0.8437
BPNN	91.65	0.9104	0.9089	0.9083
CNN	99.19	0.9909	0.9908	0.9908

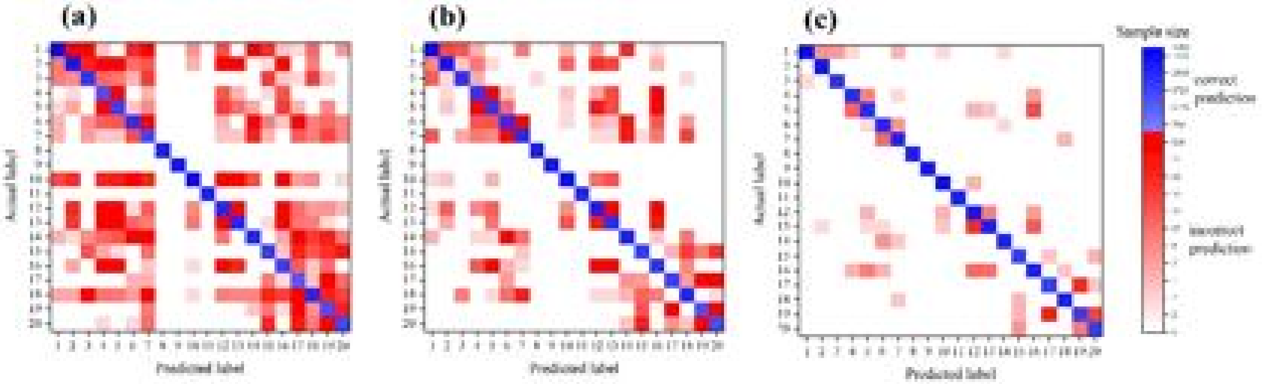

Figure 11: Confusion matrix for classification results.(a), (b) and (c) represent the confusion matrix of hyperspectral classification results for PLS, BPNN and CNN models respectively

Discussion

The two soil traceability analysis methods distinguish between the microchemical composition and macro-morphological characteristics of the soil. In terms of accuracy of the results, the soil elemental fingerprinting method combined with an appropriate machine learning model has better accuracy. That is, when combined with partial least squares discriminant analysis, it can distinguish 100% of the information from 20 soil collection points. Another advantage of using ICP-MS is the low volume of material needed, requiring only 0.25g of pre-processed soil per sample, which can be very helpful in the analysis of trace evidence [57].

The use of hyperspectral soil images and mathematical modelling to distinguish the source of the soil is superior to the use of elemental fingerprints for classification. First, elemental analysis instruments such as ICP-MS are expensive and require more complex pre-treatment processes, such as microwave digestion, autoclave digestion, and wet digestion. Moreover, it is difficult to achieve the sample size required for machine learning because of the small amount of analyzable data obtained per test. In contrast, hyperspectral techniques are convenient, fast and non-destructive, which can meet the time-sensitive requirements of detection [58]. However, hyperspectral detection requires a certain amount of sample material. For example, each image acquisition needs to consume about 2g of soil; otherwise, the color of the substrate could not be completely covered under the acquisition lens which, in turn, would have an impact on the next data processing step. The hyperspectral detection of thinner (i.e., smaller) soil layers needs to be further investigated and may require a more optimized background removal algorithm.

Conclusion

In this study, the elemental fingerprints and hyperspectral images of soil samples from five cities and 20 sampling sites in China were analyzed using ICP-MS and visible-NIR hyperspectroscopy, respectively, to compare two ideas of soil physical evidence traceability based on micro-composition and macro-morphology. In the elemental analysis, principal component analysis was used to downscale and visualize the elemental fingerprint information, and then SVM and PLS-DA were used to identify the soil sampling points. In the hyperspectral image analysis, the images were processed using pixel fusion and background removal. Thereafter, the classification efficacy of the PLS-DA, BPNN, and CNN models were compared. The results show that the ICP-MS-based soil elemental fingerprinting data provide more accurate classification results and even achieve 100% complete classification when combined with PLS-DA. However, the analysis is more costly and sample pre-processing is more complex. The hyperspectral imaging-based soil morphology analysis also achieves higher identification performance, especially when combined with the CNN model, which increases the classification accuracy to 99.19%.

For small amounts (<1g) of evidence at crime scenes, ICP-MS can be used to obtain fingerprinting information on the extracted soil material. For larger residual amounts of soil evidence, hyperspectral image acquisition is recommended, which requires less soil pre-processing, is nondestructive, and allows for subsequent destructive analysis. Thus, the soil evidence collected at the scene can be effectively used to compare suspicious samples and make a comprehensive study in conjunction with the case to obtain key clues to solve the case. For the geographical traceability of soil physical evidence, future research needs to expand the sample size and apply more machine learning methods in the actual investigation.

Funding

This work was supported by Double First-Class Innovation Research Project for People's Public Security University of China（No.2023SYL06）

References

1. Werner D, Burnier C, Yu Y, Marolf AR, Wang Y, et al. (2019) Identification of some factors influencing soil transfer on shoes. Science & justice: journal of the Forensic Science Society, 59: 643-53.

2. APS (2023) Environmental Variables in Predictive Soil Mapping: A Review[J]. Eurasian Soil Science, 56.

3. Jelena M, Carlos V, Manuel A (2022) Recent advances in multivariate analysis coupled with chemical analysis for soil surveys: a review[J]. Journal of Soils and Sediments, 23.

4. Elijah AA (2021) A Review of the Environmental Impact of Gas Flaring on the Physiochemical Properties of Water, Soil and Air Quality in the Niger Delta Region of Nigeria[J]. Earthline Journal of Chemical Sciences, 7.

5. Ting Z, Linhao W, Jianzhu L, et al. (2023) Prediction of the soil water content in the Luanhe river basin based on CMIP6[J]. Journal of Cleaner Production, 425.

6. Sun Z, Zizhong L, Zhang R , et al. (2023) Spatial variability of soil water content at the crop row scale with and without straw mulch inside a corn field in semi-humid Northeastern China[J]. Arid Land Research and Management, 37.

7. Guanshi L, Shengkui T, Guofang X, et al. (2023) Combination of effective color information and machine learning for rapid prediction of soil water content[J]. Journal of Rock Mechanics and Geotechnical Engineering, 15.

8. Jia L, Zu W, Yang F, et al. (2023) Estimating Organic Matter Content in Hyperspectral Wetland Soil Using Marine-Predators-Algorithm-Based Random Forest and Multiple Differential Transformations[J]. Applied Sciences, 13.

9. Wu M, Dou S, Lin N, et al. (2023)Estimation and Mapping of Soil Organic Matter Content Using a Stacking Ensemble Learning Model Based on Hyperspectral Images[J]. Remote Sensing, 15.

10. Jintao Y, Xican L, Shuang C, et al. (2023) Grey fuzzy prediction model of soil organic matter content using hyper-spectral data[J]. Grey Systems: Theory and Application, 13.

11. Minh HN, Kim TNH, Thi LH, et al. (2023) Utilizing X-ray fluorescence (XRF) method to evaluate the content of metal elements in soil and their effects on the total phenolic and flavonoid contents of some medicinal plants.[J]. Environmental monitoring and assessment, 195.

12. Said N, Said EM, Safa SEE, et al. (2023) Estimation of key potentially toxic elements in arid agricultural soils using Vis-NIR

spectroscopy with variable selection and PLSR algorithms[J]. Frontiers in Environmental Science, 11.

13. Tiruneh GA, Alemayehu TY, Meshesha DT, Vogelmann ES, Reichert JM, et al. (2021) Spatial variability of soil chemical properties under different land-uses in Northwest Ethiopia. .PLoS One, 16: e0253156.

14. Jelena M ,Carlos V ,Manuel A (2022) Recent advances in multivariate analysis coupled with chemical analysis for soil surveys: a review[J]. Journal of Soils and Sediments, 23.

15. Elijah AA (2021) A Review of the Environmental Impact of Gas Flaring on the Physiochemical Properties of Water, Soil and Air Quality in the Niger Delta Region of Nigeria[J]. Earthline Journal of Chemical Sciences, 7.

16. Hu C, Mei HC, Guo HL, et al. (2020) The analysis of soil evidence to associate criminal tool and location[J]. Forensic Science International, 309: 110231.

17. Anonymous (2018) Spectroscopy in Real-World Applications: Current Trends in ICP-MS, Raman, NIR, LIBS, and XRF[J]. Spectroscopy, 33.

18. Montoriol Romain (2021) Gunshot residue detection in stagnant water: SEM-EDX or ICP-MS? A preliminary study.[J]. Journal of forensic sciences.

19. Bao W, Jianfei S, Lianming X (2023) The Applications of Hyperspectral Imaging Technology for Agricultural Products Quality Analysis: A Review[J]. Food Reviews International, 39.

20. Yi G, Tao H, JingRu H (2022) Research Progress of Hyperspectral Imaging Technology in Biological Evidence.[J]. Fa yi xue za zhi, 38.

21. Yu Jiaxin, Zhao He, He Tao, Luo Tao, Zhang Wen, et al. (2021) A high-performance method for direct determination of ultra--trace REEs in geological samples by ICP-MS using a designed heating-condensing system[J]. JOURNAL OF ANALYTICAL ATOMIC SPECTROMETRY, 36.

22. Christopher SJ, Ellisor DL, Davis WC (2021) Investigating the feasibility of ICP-MS/MS for differentiating NIST salmon reference materials through determination of Sr and S isotope ratios[J]. Talanta, 231.

23. Yang Rui, Li Qingcun, Zhou Wenjing, Yu Sujuan, Liu Jingfu (2022) Speciation Analysis of Selenium Nanoparticles and Inorganic Selenium Species by Dual-Cloud Point Extraction and ICP-MS Determination.[J]. Analytical chemistry.

24. Sichler Theresa Constanze, Becker Roland, Sauer Andreas, Barjenbruch Matthias, Ostermann Markus, et al. (2022) Determination of the phosphorus content in sewage sludge: comparison of different aqua regia digestion methods and ICP-OES, ICP-MS, and photometric determination[J]. Environmental Sciences Europe, 34.

25. Clases David, Gonzalez de Vega Raquel (2022) Facets of ICP-MS and their potential in the medical sciences-Part 2: nanomedicine, immunochemistry, mass cytometry, and bioassays.[J]. Analytical and bioanalytical chemistry.

26. Sanchez Raquel, Chainet Fabien, Souchon Vincent, Carbonneaux Sylvain, Lienemann Charles Philippe, et al. (2020) Silicon speciation in light petroleum products using gas chromatography coupled to ICP-MS/MS[J]. JOURNAL OF ANALYTICAL ATOMIC SPECTROMETRY, 35.

27. Ulanova Tatyana S, Volkova Marina V, Stenno Elena V, Nedoshitova Anna V, Veikhman Galina A (2019) Assessment of rare earth elements by ICP-MS in workplace air of metallurgical enterprise[J]. Occupational Health and Industrial Ecology.

28. He, Qian, Xing, Zhi, Zhang (2015) ICP-MS/MS as a tool to study abiotic methylation of inorganic mercury reacting with VOC-s[J]. Journal of Analytical Atomic Spectrometry, 30.

29. 29] A Khan, M Munir, W Yu (2020) A review towards hyperspectral imaging for real time quality control of food products with an illustrative case study of milk powder production[J], Food Bioprocess Technol, 13: 739-752.

30. W Che, L Sun, Q Zhang (2018) Pixel based bruise region extraction of apple using Vis-NIR hyperspectral imaging[J], Comput. Electron. Agric, 146: 12-21.

31. Y Zhao, C Zhang, S Zhu (2020) Shape induced reflectance correction for non-destructive determination and visualization of soluble solids content in winter jujubes using hyperspectral imaging in two different spectral ranges[J], Postharvest Biol. Technol. 161: 111080.

32. A Gong, S Zhu, Y He (2017) Grading of Chinese Cantonese sausage using hyperspectral imaging combined with chemometric methods[J], Sensors, 17: 1706.

33. C Zhang, H Jiang, F Liu (2017) Application of near-infrared hyperspectral imaging with variable selection methods to determine and visualize caffeine content of coffee beans[J], Food Bioprocess Technol, 10: 213-21.

34. W He, H He, F Wang (2021) Non-destructive detection and recognition of pesticide residues on garlic chive (Allium tuberosum) leaves based on short wave infrared hyperspectral imaging and one-dimensional convolutional neural network [J], J. Food Meas. Char, 1-11.

35. M Zulfiqar, M Ahmad, A Sohaib (2021) Hyperspectral imaging for bloodstain identification[J], Sensors, 21: 3045.

36. MAF De La Ossa, JM Amigo, C García-Ruiz (2014) Detection of residues from explosive manipulation by near infrared hyperspectral imaging: a promising forensic tool[J], Forensic Sci. Int, 242: 228-35.

37. CS Silva, MF Pimentel, JM Amigo. Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models[J], TrAC, Trends Anal. Chem. 95: 23-35.

38. S Wang, H He, R Lv (2021) Classification Modeling Method for Hyperspectral Stamp-pad Ink Data Based on One-dimensional Convolutional Neural network[J], Journal of Forensic Sciences.

39. Koch Markus (2021) Effects of water tension and surface roughness on soil hyperspectral reflectance[J]. Geoderma, 385.

40. Xu Xitong (2020) Exploring Appropriate Preprocessing Techniques for Hyperspectral Soil Organic Matter Content Estimation in Black Soil Area[J]. Remote Sensing, 12 : 3765-3765.

41. Cássia DKM, Scorsatto RO, Flôres MF (2023) Hyperspectral imaging in forensic science: An overview of major application areas[J]. Science & Justice, 63.

42. Corrêa R, Melo V, Abreu G (2018) Soil forensics: How far can soil clay analysis distinguish between soil vestiges?[J]. Science & Justice, 58.

43. Bing Lu (2020) Recent Advances of Hyperspectral Imaging Technology and Applications in Agriculture[J]. Remote Sensing, 12.

44. Ewing Jordan (2020) Utilizing Hyperspectral Remote Sensing for Soil Gradation[J]. Remote Sensing, 12: 3312-3312.

45. Qian Lu (2019) Rapid inversion of heavy metal concentration in karst grain producing areas based on hyperspectral bands associated with soil components[J]. Microchemical Journal, 148 : 404-11.

46. Barra Issam (2020) Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances -A review[J]. TrAC Trends in Analytical Chemistry, 135: 116166.

47. Iman Tahmasbian (2018) Laboratory-based hyperspectral image analysis for predicting soil carbon, nitrogen and their isotopic compositions[J]. Geoderma, 330: 254-63.

48. 48. Shiwen Wu (2018) Mapping the Salt Content in Soil Profiles using Vis-NIR Hyperspectral Imaging[J]. Soil Science Society of America Journal, 82: 1259-69.

49. LeCun Yann, Bengio Yoshua, Hinton Geoffrey (2015) Deep learning.[J]. Nature, 521: 436-44.

50. FM Riese, S Keller (2019) SOIL TEXTURE CLASSIFICATION WITH 1D CONVOLUTIONAL NEURAL NETWORKS BASED ON HYPERSPECTRAL DATA[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-2/W5 : 615-21.

51. Savignan Lionel (2021) Spatial distribution of trace elements in the soils of south-western France and identification of natural and anthropogenic sources[J]. Catena, 205.

52. Zhang Xiang (2021) Source identification of soil elements and risk assessment of trace elements under different land uses on the Loess Plateau, China.[J]. Environmental geochemistry and health, 43: 2377-92.

53. Cheng Hang, Wang Jing, Du Yingkun (2021)Combining multivariate method and spectral variable selection for soil total nitrogen estimation by Vis–NIR spectroscopy[J]. Archives of Agronomy and Soil Science, 67: 1665-78.

54. FY Lian (2019) Quantitative Analysis of Trans Fatty Acids in Cooked Soybean Oil Using Terahertz Spectrum[J]. Journal of Applied Spectroscopy, 86: 917-24.

55. Wang Shuyue, He Hongyuan, Lv Rulin, He Weiwen, Li Chunyu (2021) Classification modeling method for hyperspectral stamp-pad ink data based on one-dimensional convolutional neural network.[J]. Journal of forensic sciences.

56. Yuhang Pan (2020) Activation functions selection for BP neural network model of ground surface roughness[J]. Journal of Intelligent Manufacturing, 1-12.

57. Desiderio Vincent J, Taylor Chris E, Daéid Niamh Nic (2020) Handbook of Trace Evidence Analysis[M]. Chichester, UK : John Wiley & Sons, Ltd.

58. Crowther M, Li B, Thompson T, Islam MA (2021) comparison between visible wavelength hyperspectral imaging and digital photography for the detection and identification of bloodstained footwear marks. J Forensic Sci, 66: 2424-37.