INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Discrimination of Soil Samples Collected from Haryana (India) Using Non-destructive ATR-FTIR Spectroscopy Coupled with Multivariate Statistical Analysis

Preeti Sangwan[1], Chongtham Nimi[2], Tarsem Nain[3], Rajinder Singh[2], Neelkamal Sharma[4]*

1 Department of Forensic Science, Maharshi Dayanand University, Rohtak, 124001, Haryana, India
2 Department of Forensic Science, Punjabi University, Patiala, 147002, Punjab, India
3 Department of Genetics, Maharshi Dayanand University, Rohtak, 124001, Haryana, India
4 Associate Professor, Department of Forensic Science, Maharshi Dayanand University, Rohtak, 124001, Haryana, India

OPEN ACCESS

Received: 19-11-2023
Accepted: 09-02-2024
Published: 07-03-2024

Citation: Sangwan P, Nimi C, Nain T, Singh R, Sharma N (2024) Discrimination of Soil Samples Collected from Haryana (India) Using Non-destructive ATR-FTIR Spectroscopy Coupled with Multivariate Statistical Analysis. Indian Journal of Science and Technology 17(11): 1087-1096. https://doi.org/10.17485/IJST/v17i11.2930

* **Corresponding author**.

neelforensics@gmail.com

Funding: University Grants Commission (UGC), New Delhi for providing financial assistance

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To discriminate and classify soil samples collected from different regions of Haryana, India. **Methods:** Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) spectroscopy with multivariate statistical tools is employed. A total of 232 samples were collected. A composite mixture of all districts was prepared, having twenty-nine top and twenty-nine depth soil samples. Chemometric methods, namely, PCA (Principal Component Analysis) and PCA-LDA (Principal Component Analysis-Linear Discriminant Analysis) were used to interpret the data. **Findings:** Soil samples are well characterized by their organic and inorganic contents. Sample clustering due to similarity in chemical composition was visualized using PCA. PCA-LDA resulted in 100% classification accuracy for top soil and 98.85% classification accuracy for depth soil. Blind test validation was carried out, which resulted in 100% and 80% prediction accuracies for top soil and depth soil respectively. The present research methodology effectively discriminated soil samples and can be utilized by forensic investigators dealing with cases that involve soil as vital evidence. **Novelty:** Study reveals novel unexplored geographical location, local soil variability, practical implications of non-destructive analytical technique combined with chemometrics methods, contextualization with the previous studies and the potential policy field relevance.

Keywords: Soil forensics; ATRFTIR; PCA; LDA; Discrimination

1 Introduction

Soil is valuable to trace evidence commonly recovered from clothes, vehicles, footwear, or crime scenes. It has a significant value because of its complex, heterogenous, and transferable nature. Using analytical techniques to analyze soil samples collected from crime scenes helps the forensic expert evaluate whether the samples originate from the same geographic place. Moreover, soil examination provides valuable information about Hit-and-Run cases, wildlife crimes, sexual assaults, murder, missing cases, etc. As most crimes are committed outside, and soil frequently transfers from one location to another [1].

Characterization and discrimination of soils are two main parts of soil examination. Both features aid in linking crime with suspects, victims, and objects and determining the crime's origin. There are several analytical techniques being employed for characterization of inorganic components in soil samples [2,3]. In addition to being expensive and destructive, these techniques are not common in every forensic science laboratory. To overcome these limitations, non-destructive, cost-effective, fast, and requires little to no sample preparation are available via spectroscopic techniques such as Attenuated Transmission Reflectance Fourier Transformed Infrared Spectroscopy (ATR-FTIR), Raman spectroscopy, and diffuse reflectance UV-Vis-NIR spectroscopy.

However, applying these analytical techniques for forensic purposes could be challenging because forensics demands a high degree of accuracy in data analysis compared to agricultural or environmental programs, where similar techniques are also routinely used. Moreover, it is also problematic because soil transferred during criminal events is typically small. This trace amount might not be sufficient for analysis by destructive methods. That is why we require non-destructive, cost-effective, fast, no-sample preparation, reproducible results, and reliable spectroscopic methods, thus obtaining results that can efficiently differentiate and characterize samples from one location to another or show similarity in samples originating from same location.

Since the last few decades, FTIR spectroscopy has emerged as one of the most useful spectroscopic approaches. It is a handy method for analyzing soil and sediment samples as IR spectra determine the presence of soil organic matter and minerals. Xing et al. reported the compositional and structural change in SOM of different depth soils by FTIR-PAS with principal component analysis (PCA). Goydaragh et al. [4] investigated SOM by using a combination of environmental variables and FTIR spectroscopy using tree-based models. Kocak et al. [5] studied feasibility of vibrational spectroscopy for the analysis of soil samples in a single location without focusing on the general intrinsic components of soil. Parnpuu et al. [6] reported using FTIR spectroscopy technique to estimate SOM in different soils.

These studies are carried out on a minimal sample size to quantify organic and inorganic contents in soil samples. There hasn't been a study to date that examines the geographical restricted representation of soil types, depth variation, comparison with advanced techniques, temporal variability, integration with field observations, impact of anthropogenic activities, economic feasibility, spatial resolution and scale, insufficient validation. Therefore, there is a strong need to generate data on soil of different areas of Haryana, India, which will help the forensic expert find out the culprit. This will reduce their fieldwork and prove a boon for them.

2 Methodology

2.1 Study area

All soil samples used in the current study were collected from various areas of Haryana, India, located in the northern part of the country (29° 3' 56.7828" N and 76° 2' 25.7892" E) (Figure 1).

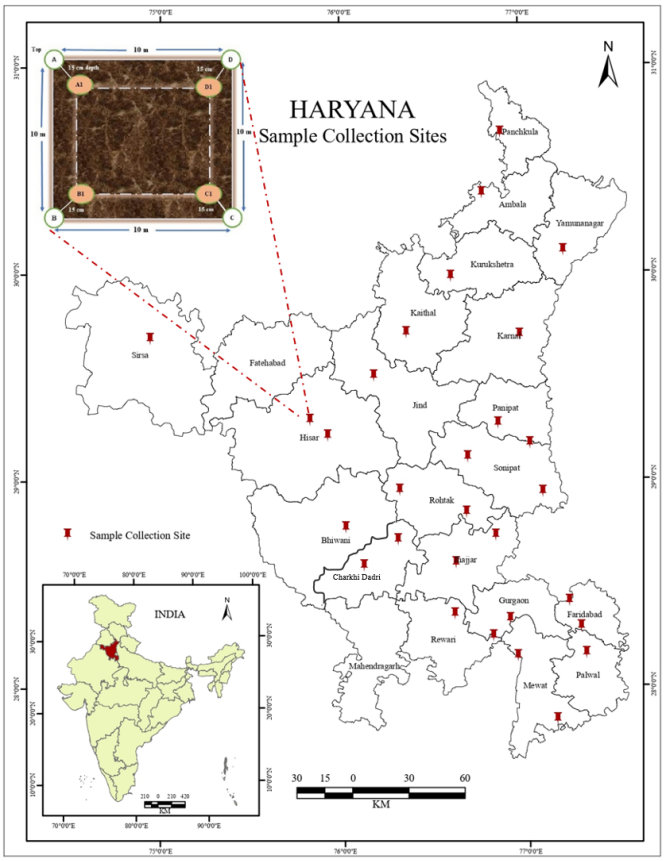

Fig 1. Sample collection site

2.2 Sample collection

To collect samples, a 10x10 meter grid was marked. Debris and other foreign contaminants were partially removed from the earth's surface. Then, soil samples (approx. 400 grams) were taken from four locations representing each of the four corners. Similarly, four samples from about 15cm depth were collected with the same approach. The collected samples were stored in plastic zipper bags with proper markings (sample ID and coordinates) (Table 1).

Table 1. Sample IDs and their coordinates

S.No.	Districts	Topsoil sample ID	Depth soil sample ID	Latitude (N)	Longitude (E)
1	Jind	T1	D1	29°30'36"	76°10'33"
2	Rewari	T2	D2	28°21'23"	76°36'31"
3	Rohtak	T3	D3	28°50'49"	76°40'39"
4	Faridabad	T4	D4	28°24'47"	77°13'52"
5	Bhiwani	T5	D5	28°46'34"	76°00'59"
6	Charkha Dadri	T6	D6	28°35'29"	76°06'47"
7	Charkha Dadri	T7	D7	28°43'01"	76°18'02"
8	Sonipat	T8	D8	29°06'49"	76°41'14"
9	Panipat	T9	D9	29°16'29"	76°51'29"
10	Jhajjar	T10	D10	28°36'11"	76°37'02"
11	Jhajjar	T11	D11	28°44'00"	76°50'14"
12	Nuh	T12	D12	28°09'02"	76°56'53"
13	Rohtak	T13	D13	28°57'23"	76°18'44"
14	Hisar	T14	D14	29°13'19"	75°55'9"

Continued on next page

Table 1 continued

15	Gurugram	T15	D15	28°19'41"	76°54'29"
16	Gurugram	T16	D16	28°14'45"	76°48'57"
17	Hisar	T17	D17	29°17'56"	75°49'15"
18	Kaithal	T18	D18	29°43'00"	75°21'28"
19	Kurukshetra	T19	D19	29°59'11"	76°36'28"
20	Ambala	T20	D20	30°23'13"	76°47'08"
21	Sonipat	T21	D21	28°56'26"	77°05'52"
22	Panipat	T22	D22	29°10'36"	77°01'46"
23	Nuh	T23	D23	27°50'32"	77°09'29"
24	Palwal	T24	D24	28°09'29"	77°19'16"
25	Faridabad	T25	D25	28°17'19"	77°17'48"
26	Sirsa	T26	D26	29°41'24"	74°56'28"
27	Karnal	T27	D27	29°42'02"	76°59'05"
28	Yamunanagar	T28	D28	30°06'25"	77°13'55"
29	Panchkula	T29	D29	30°40'40"	76°53'27"

2.3 Sample pretreatment

After sample collection, they were allowed to dry at room temperature for at least five days before being crushed using a pestle and mortar and sieved through a 2 mm standard sieve to remove plastics, stones, leaves, and other foreign objects. Then samples were stored in airtight specimen tubes. A total of 232 samples were taken. A composite mixture of all districts was prepared, giving twenty-nine top and twenty-nine depth soil samples.

2.4 ATR-FTIR setup

ATR-FTIR spectra were carried out by FTIR spectrophotometer (Perkin Elmer SPECTRUM TWO) in the range of 4000 to 400 cm^{-1} at 4 cm^{-1} resolution, and 16 scans per sample were preferred to avoid any error or noise in the spectra. A background scan was done for calibration by running a spectrometer without a sample. As soon as the background scan was completed, dried and sieved soil samples were placed directly on detection window. Then, ATR crystal was pressed on the soil samples. All the 58 composite samples (29 samples for top soil and 29 samples for depth soil) were analyzed in triplicates to be able to consider sample variation. To check the reproducibility of the FTIR instrument, same sample was analyzed three times. The Origin 2019b software was used to plot data and visually assess the spectra.

2.5 Multivariate data analysis

Multivariate analysis is a key method to evaluate large data sets that contain more than one variable[7].

2.5.1 Principal Component Analysis (PCA)

PCA is an unsupervised pattern recognition tool. Using PCA, a large number of interrelated variables can be simplified to a few PCs. The first few PCs in the set, which are uncorrelated from one another, account for the majority of the variations seen in the datasets. The PCs with eigenvalues greater than one are chosen from among all PCs. Additionally, it is a handy method to illustrate the significant correlations between selected samples and helps to summarize and presents the original data. In the present study, PCA was used to visualize the trends in the dataset and to identify clustering of the samples due to their similarity.

2.5.2 Linear Discriminant Analysis (LDA)

LDA is the most extensively studied supervised pattern recognition technique. It is primarily used to characterize or separate two more classes of events and to make decisions among the specified classes without altering the shape or location of the original datasets. While LDA and PCA both create new variables from the original datasets, the primary distinction between the two is that LDA achieves maximal separation in comparison to PCA. The new variables are known as discriminant functions and are orthogonal to one another. The efficiency of classification is increased by using a combination of PCA and LDA by automatically selecting the most significant features to build the classification model[8]. In the present study, PCA-LDA was carried out using Unscrambler X software (Version 10.5.1, 64-bit, CAMO, AS, Norway). The first 3 PCs derived from PCA were employed for PCA-LDA.

2.5.3 Data pre-processing

Before performing the main data analysis, the data was mathematically transformed to reduce and eliminate any noise and variation from extraneous factors. All spectra were subjected to ATR correction prior to the application of any pre-processing methods. To perform data pre-processing and, for ATR correction, Unscrambler X software (Version 10.5.1, 64-bit, CAMO, AS, Norway) was used.

Pre-processing methods such as baseline offset and linear baseline correction, smoothing with Savitzky–Golay algorithm with 3 smoothing points and 2 polynomial orders in symmetric kernel and normalization by range were performed on all the spectra acquired in the present study.

2.5.4 Discrimination power (DP)

DP was first calculated by Smalldon and Moffat formula[9]

$$DP = \frac{Total\ no.\ of\ discriminating\ pair\ of\ samples}{Total\ no.\ of\ possible\ pair\ of\ samples} X\ 100$$

Total number of possible sample pairs $(n) = \frac{n(n-1)}{2}$, Where n is the total number of samples.

3 Results and Discussion

3.1 Characterization of soil samples

The obtained spectra of soil samples in the mid-infrared region (4000-400 cm^{-1}) were used to check the compositional differences between samples taken from different locations. The spectra of all 58 samples (29 from top surface and 29 from 15cm depth) are qualitatively examined. Many characteristic peaks were identified in the fingerprint region from 1800-400 cm^{-1}, which played an influential role in soil sample discrimination. The representative spectra of surface soil samples are shown in Figure 2.

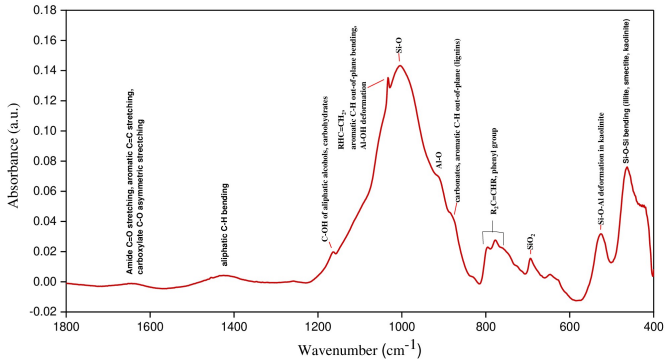

Fig 2. Mean ATR-FTIR spectra of collected soil samples in fingerprint area

The obtained ATR-FTIR spectra were further classified into eight regions. In the first region, ranges from 460-470 cm^{-1} showed $v_{Si-O-Si}$ symmetrical bending vibrations in kaolinite, illite, and smectite. The second region ranging from 525-535 cm^{-1} is caused by bending vibrations of $v_{Al-O-Si}$ in kaolinite[10]. The third region, i.e., 690-700 cm^{-1} showed metal oxides and carbonates deformation bands. The absorbance at 775-785 cm^{-1} attributed to the fourth region corresponds to v_{Si-O} symmetrical stretching in quartz, calcite, and as well as the presence of R_2C=CHR groups[11]. The bands in the fifth wavelength region that appeared around 995-1005 cm^{-1} are due to aromatic v_{C-H} and $v_{C=C}$ from polysaccharides[11]. In the sixth region, absorbance at 1160-1170 cm^{-1} showed v_{C-O} stretching of polysaccharides, alcohol, ester, and ether-like groups. The absorbance at 1430-1435 cm^{-1} attributed to the seventh region showed aliphatic v_{C-H} bending of CH_2 and CH_3 groups[12]. Lastly, 1635-1650 cm^{-1} is associated with asymmetrical stretching of metal carboxylate as well as presence of humic acids, proteins and lignin[12].

Detailed information about the absorption bands resembled the organic and inorganic constituents detected in all top surface and depth samples collected from different sites (Table 2). Different types of organic and inorganic i.e., quartz, calcite, kaolinite, aragonite, hematite, and bentonite minerals are identified.

Table 2. ATR-FTIR organic and inorganic constituents' identification of soil samples

Frequency in cm^{-1}		Band assignment		Samples studied	References
Peaks	**Reported region**	**Inorganic constituents**	**Organic constituents**		
3621	3625-3615	Si-O-H vibrations of clays, gibbsite, iron oxides, kaolinite, free O-H, N-H stretching	Moisture and oxygen-containing organic compound	T7, T9, T14-T18, T20-T22, D4, D25, D27, D28	[13]
	3440-3320	Hydrogen bonded O-H and N-H stretching	n/a	T3, T14, T15, T17, D5, D7	[11]
	2965-2853	n/a	Aliphatic C-H symmetric and asymmetric stretching	T20-T26, D5	[12]
1642	1660-1640	O and N- containing polar functional group	Amide C=O stretching, aromatic C=C stretching, carboxylate C-O asymmetric strectching, C=N stretching, conjugated ketone C=O stretching	T1, T3-T7, T9, T11, T13-T17, T19-T28, D1-D4, D7, D8, D10, D12-D20, D22-D25, D27, D28	[12]
1434	1444-1408	n/a	Aliphatic C-H bending	T1, T5, T7, T9, T11, T14, T17, T20-T23, T25-T28, D1, D5, D7, D9, D10, D14, D17, D20, D22, D25, D26	[12]
	1403-1354	n/a	C-O stretching and O-H deformation of COOH, phenolic C-O stretching	T12	[14]
~ 1161	1185-1144	n/a	C-OH of aliphatic alcohols, carbohydrates	T_1-T_{25}, D_1-D_{25}	[15]
1033	1045-1010	Al-OH deformation (kaolinite)	RHC=CH_2, aromatic C-H out-of-plane bending	T1, T2, T4, T10, T14-T16, T19, T20, T23-T26, T29, D1-D7, D10, D16, D20, D23, D24	[16]
1000	1005-995	SiO_2 Si-O stretch lattice	Polysaccharides aromatic =CH and C=C groups	T1, T4-T16, T18-T28, D1-D5, D7-D16, D18-D29	[11]
	945-870	n/a	Benzoic acid, pyranose ring (carbohydrates), cellulose, RHC=CH_2, R_2C=CH_2	T1, T4, T8, T9, T14-T23, T27, T28, D1-D4, D8, D11, D15-D24, D27, D28	[17]
872	870-890	Carbonates	Lignins (Aromatic C-H out-of-plane)	T5, T14, T23, T25, T26, D7, D9, D14, D22, D25, D26	[17]
794	820-752	Inorganic materials (clay and quartz), carbonates, kaolinite	R_2C=CHR, phenyl group	T_1-T_{25}, D_1-D_{25}	[11]
720	725-720	Calcite	Long chain alkanes (C-H rock methyl)	T23, T25, T26, D12, D13, D25, D26	[18]
695	697-690	SiO2	n/a	T1-T25, D1-D25	[19]
650		Bentonite	n/a	T1, T3, T5 -T7, T25, T26, D1, D5, D6, D11-D13, D26	[20]
525	535-525	Si-O-Al deformation in kaolinite	n/a	T1-T29, D1-D29	[10]
464	470-460	Si-O-Si bending (illite, smectite, kaolinite)	n/a	T1-T29, D1-D29	[21]

Furthermore, three separate checks of soil samples number T7 and T12 were conducted to determine intra-location variations and reproducibility, and no significant differences were found, as shown in **Figure S1 and S2**, respectively.

3.2 Visual inspection and Preliminary Discrimination of spectra

The visual inspection was carried out by comparing the spectra of each soil sample. For discrimination purposes, only the presence or absence of peaks was considered as it makes the spectra more distinguishable. Detailed information regarding the presence or absence of peaks at various wave number ranges is shown in **Tables S6 and S7**. Further, pairwise discrimination reveals that most of the samples are discriminated from each other except 108 pairs and 116 pairs of soil samples in the case of surface and depth samples respectively as shown in **Table S1**. This might be due to similar geochemical composition during their formation. The DP is calculated according to Smalldon and Moffat formula. Thus, this method shows 73.39 % and 71.42 % discriminating power for surface and depth soil samples respectively.

This approach is helpful for the discrimination of small number of samples. Increasing sample size also increases the possibility of error because manual comparison of spectral data is extremely time-consuming and labor-intensive. Therefore, there is a need for a chemometrics statistical analysis method that can provide accurate and trustworthy results in a shorter time.

3.3 Chemometric Discrimination

3.3.1 Top Soil

3.3.1.1 PCA. PCA was performed using the ATR-FTIR spectra in the range of 4000-400 cm^{-1} to visualize the clustering of all the samples. The explained variances of the first (PC1), second (PC2), third (PC3), and fourth (PC4) principal components were 76%, 16%, 3%, and 2%, respectively. The total explained variance was 97%, accounted by the first four PCs. The first three PCs were used to make a score plot as they contained the most variances. The score plot using the first 3 PCs is shown in Figure 3. The top soil samples formed a cluster on the PCA score plot, with a few samples separated from the cluster. Samples T2 (Rewari) and T3 (Rohtak) were separated from the cluster along PC1. Sample T29 (Panchkula) was separated from the cluster along PC2 whereas sample T17 (Hisar) was separated along PC3. In the score plot, visually distinct samples were grouped together along PC1 and PC3. Minor peaks that allowed for visual differentiation between particular samples were probably deemed to be of low significance by PCA and subsequently incorporated into later PCs that weren't used for plotting the data. However, PCA is only a tool used for visualizing the trend in the dataset, therefore, a supervised discrimination tool is used in the further section.

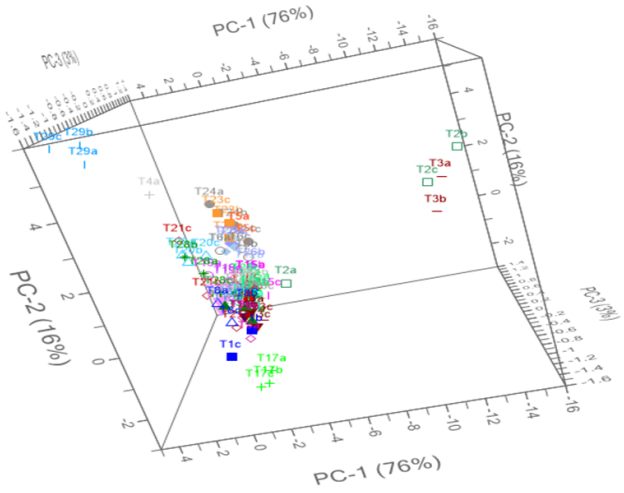

Fig 3. PCA score plot of all investigated top-surface soil samples using 3 PCs

The factor loadings of the top samples are given in **Figure S3**, PC1 is positively correlated at 2986-2920 cm^{-1}, 1072-990 cm^{-1} and negative loading at 4000-2991 cm^{-1}, 2918-1073 cm^{-1}, 988-469 cm^{-1}. PC2 showed positive loading at 2214-1036 cm^{-1}, 867-558 cm^{-1}, 499-445 cm^{-1} and negative loading at 4000-2399 cm^{-1}, 1036-876 cm^{-1}, 439-400 cm^{-1}. PC3 showed positive loading at 4000-2810 cm^{-1}, 1534-1305 cm^{-1}, 1181-463 cm^{-1} and negative loading at 2804-1539 cm^{-1}, 457-438 cm^{-1}.

3.3.1.2 PCA-LDA. All the samples in the present study were subjected to PCA-LDA to get an objective classification of the samples. To perform PCA-LDA, all replicate spectra that were registered between 4000-400 cm^{-1} were used. With the help of the first three PCs, the LDA model was developed for the topsoil samples. Each sample was treated as a separate class. All the

samples were correctly classified with a classification accuracy of 100%. No misclassification was observed as shown in **Table S2**. The PCA-LDA model was able to correctly classify all the samples which could not be separated on the PCA plot. The model was then applied to the prediction of unknown samples. Additional spectra (newly analyzed) from 5 samples were used to validate the model. The samples for the validation were randomly chosen and were given the sample codes U1, U2, U3, U4, and U5. The identity of these samples was not informed to the researcher. U1, U2, U3, U4 and U5 were correctly predicted as belonging to samples T26, T29, T22, T4 and T3 respectively. A validation accuracy of 100% was obtained. Results acquired from PCA-LDA model are illustrated in **Figure S4 and Table S3**.

3.3.2 Depth Soil

3.3.2.1 PCA. The loading plot provided information regarding the number of PCs to be considered to describe the variances. Explained variances of first (PC1), second (PC2), third (PC3), fourth (PC4), fifth (PC5), and sixth (PC6) principal components were 47%, 25%, 13%, 5%, 4%, and 2% respectively. The total explained variance was 96%, accounted by the first six PCs. The first three PCs were used to make a score plot as they contained the most variances. The score plot using the first 3 PCs is shown in Figure 4. Various samples formed a cluster on the PCA score plot, with a few samples separated from the cluster. Samples D7 (Charkha Dadri) and D29 (Panchkula) were separated from the cluster along PC3. Sample D26 (Sirsa) was separated along PC2. Samples D9 (Panipat) and D25 (Faridabad) formed a sub-cluster.

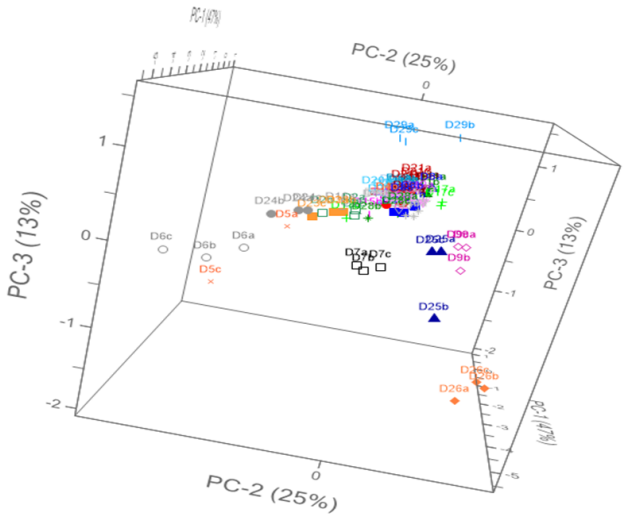

Fig 4. PCA score plot of all investigated depth soil samples using 3PCs

As observed in the case of top soil, the PCA plot of depth soil also showed a close grouping of samples which showed dissimilar spectral profiles. Minor peaks that allowed for visual differentiation between particular samples were probably deemed to be of low significance by PCA and subsequently incorporated into later PCs that were not used for plotting the data. However, PCA is only a tool used for visualizing the trend in the dataset, therefore, a supervised discrimination tool is used in the further section.

The factor loadings of depth soil samples are given in **Figure S5**, PC1 showed positive loading at 4000-1536 cm^{-1}, 1008-812 cm^{-1}, 686-497 cm^{-1}, 425-400 cm^{-1} and negative loading at 1524-1009 cm^{-1}, 810-700 cm^{-1}, 497-420 cm^{-1}. PC2 showed positive loading at 4000-2895 cm^{-1}, 1676-1573 cm^{-1}, 1232-877 cm^{-1}, 867-403 $cm^{-1}$and 2841-1694 cm^{-1}, 1560-1249 cm^{-1}, 875 cm^{-1}, 400 cm^{-1}. PC3 showed positive loading at 2894-1542 cm^{-1}, 1270-1070 cm^{-1}, 812-751 cm^{-1}, 474-426 $cm^{-1}$and negative loading at 4000-3212 cm^{-1}, 1540-1276 cm^{-1}, 1070-810 cm^{-1}, 749-480 cm^{-1}, 418-400 cm^{-1}.

3.3.2.2 PCA-LDA. All the samples in the present study were subjected to PCA-LDA to get an objective classification of all samples. To perform PCA-LDA, all replicate spectra of samples registered between 4000-400 cm^{-1} were used. With the help of the first three PCs, the LDA model was developed for depth soil samples. Each sample was treated as a separate class. A PCA-LDA classification accuracy of 98.85% was achieved. A misclassification was observed as shown in **Table S4**. A single replicate of sample D5 was misclassified as D4. Apart from a single replicate of sample D5, PCA-LDA model was able to correctly classify all samples which showed close clustering in the PCA plot into their groups. The model was subsequently used for prediction

of unknown samples. The additional spectra (newly analyzed) from 5 samples were used to validate the model. The samples for the validation were randomly chosen and were given the sample codes U1, U2, U3, U4, and U5. The identity of these samples was not informed to the researcher. U1, U2, U3, and U5 were correctly predicted as belonging to samples D11, D19, D13, and D29 respectively. Sample U4 was incorrectly predicted as belonging to D4. A validation accuracy of 80% was obtained. Results acquired from PCA-LDA model are illustrated in **Figure S6 and Table S5**.

In a previous study reported by Chauhan et al., [22] depth soil was more accurately classified than surface soil. In the present study, topsoil (100%) showed a higher classification accuracy than depth soil (98.85%). The additional classification of the topsoil could be attributed to organic matter on the topsoil, which contributed to greater variability in the chemical composition, allowing for more accurate classifications.

- **Potential benefits of ATR-FTIR/ PCA/LDA over other methods**

The advantages of using ATR-FTIR with PCA and LDA for soil sample discrimination in forensic protection include the ability to integrate with databases for effective matching, multivariate analysis, dimensionality reduction, improved accuracy and sensitivity, statistical confidence, automated pattern recognition, and interpretable results. Table 3 presents a comparison between the research approach used in this study and its conclusions with previous studies.

Table 3. Comparative analysis of soil samples

Analytical methods	Year	Statistical methods used	Area of Study	Findings	References
SEM-EDS	2019	Chi-square test	Japan	Discrimination of soil samples	[23]
EDXRF &FTIR	2019	PCA	Brazil	98.6 % similarity found in soil samples	[24]
UV-visible spectroscopy	2020	Correlation	China	Soil organic matter is more favorable to bind with Pb^{+20} ions than Cd^{+2} ions	[25]
Color, UV-NIR	2020	PCA	China	Quantitative measurement based on the similarity and dissimilarity of soil samples	[26]
TGA	2020	PCA, HCA, PCA-LDA	India	100% discrimination and classification of soil samples	[27]
UV-visible	2021	PCA, PC-LDA Clustering	India	Characterization of soil from different areas, 95% correct classification by leaving out chain validation	[28]
Vis-NIR	2022	PLSR, PCR, LDA	New York	Collect and model data carefully using visible and near-infrared reflectance spectroscopy to detect distinct types of lead in soil	[29]
FTIR, XRD	2023	PCA, LDA	Western Australia	Arid, sandy soils, accurately distinguish and associate an unknown ”recovered” sample with a single reference soil	[30]

4 Conclusion

In the present study, ATR-FTIR spectroscopy was used to characterize and discriminate between 58 (29 top and 29 depth surface) soil samples. These samples were analyzed in replicates in order to consider sample variation. The ATR-FTIR spectroscopy method delivered a non-destructive analysis of trace samples in a lesser time. There was no sample preparation and results were reproducible. Multivariate statistical tools such as PCA and PCA-LDA were employed for better discrimination and classification between samples. PCA was used to visualize the trends in the dataset. PCA-LDA resulted in a classification accuracy of 100% and 98.85% in top-surface and depth samples respectively. For top-surface and depth soil samples, blind test validation showed 100% and 80% accuracy, respectively. Expanded geographic coverage, temporal monitoring, integration with remote sensing, correlation with multiple crimes, multiple depth analysis, combined application of analytical techniques, validation through field observations, cost-benefit analysis, and cooperation with interdisciplinary fields are the major future prospects and research directions that will be investigated.

References

1) Sangwan P, Nain T, Singal K, Hooda N, Sharma N. Soil as a tool of revelation in forensic science: a review. *Analytical Methods*. 2020;12(43):5150–5159. Available from: https://doi.org/10.1039/D0AY01634A.
2) Mishra AC, Gupta S. Analysis of Heavy Metal in Industrial Soil Through Atomic Absorption Spectroscopy and its Relationship with Some Soil Properties. *Journal of Materials & Metallurgical Engineering*. 2021;11(2). Available from: https://engineeringjournals.stmjournals.in/index.php/JoMME/article/view/5865.
3) Allegretta I, Legrand S, Alfeld M, Gattullo CE, Porfido C, Spagnuolo M, et al. SEM-EDX hyperspectral data analysis for the study of soil aggregates. *Geoderma*. 2022;406:115540. Available from: https://doi.org/10.1016/j.geoderma.2021.115540.
4) Goydaragh MG, Taghizadeh-Mehrjardi R, Jafarzadeh AA, Triantafilis J, Lado M. Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. *CATENA*. 2021;202:105280. Available from: https://doi.org/10.1016/j.catena.2021.105280.
5) Koçak A, Wyatt W, Comanescu MA. Comparative study of ATR and DRIFT infrared spectroscopy techniques in the analysis of soil samples. *Forensic Science International*. 2021;328:111002. Available from: https://doi.org/10.1016/j.forsciint.2021.111002.
6) Pärnpuu S, Astover A, Tõnutare T, Penu P, Kauer K. Soil organic matter qualification with FTIR spectroscopy under different soil types in Estonia. *Geoderma Regional*. 2022;28:e00483. Available from: https://doi.org/10.1016/j.geodrs.2022.e00483.
7) Sauzier G, Van Bronswijk W, Lewis SW. Chemometrics in forensic science: approaches and applications. *The Analyst*. 2021;146(8):2415–2448. Available from: https://doi.org/10.1039/D1AN00082A.
8) Gautam R, Vanga S, Ariese F, Umapathy S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*. 2015;2(1):1–38. Available from: https://doi.org/10.1140/epjti/s40485-015-0018-6.
9) Lei L, Massonnet G. Forensic analysis of white automotive paint of same manufacturer with Raman spectroscopy and chemometrics. *Journal of Raman Spectroscopy*. 2024;55(2):148–160. Available from: https://doi.org/10.1002/jrs.6626.
10) Madejova J, Komadel P. Baseline studies of the clay minerals society source clays: infrared methods. *Clays and clay minerals*. 2001;49(5):410–432. Available from: https://doi.org/10.1346/CCMN.2001.0490508.
11) Volkov DS, Rogova OB, Proskurnin MA. Organic Matter and Mineral Composition of Silicate Soils: FTIR Comparison Study by Photoacoustic, Diffuse Reflectance, and Attenuated Total Reflection Modalities. *Agronomy*. 2021;11(9):1–30. Available from: https://doi.org/10.3390/agronomy11091879.
12) Calderón F, Haddix M, Conant R, Magrini-Bair K, Paul E. Diffuse-Reflectance Fourier-Transform Mid-Infrared Spectroscopy as a Method of Characterizing Changes in Soil Organic Matter. *Soil Science Society of America Journal*. 2013;77(5):1591–1600. Available from: https://doi.org/10.2136/sssaj2013.04.0131.
13) Tinti A, Tugnoli V, Bonora S, Francioso O. Recent applications of vibrational mid-Infrared (IR) spectroscopy for studying soil components: a review. *Journal of Central European Agriculture*. 2015;16(1):1–22. Available from: https://doi.org/10.5513/JCEA01/16.1.1535.
14) Dhillon GS, Gillespie A, Peak D, Van Rees KCJ. Spectroscopic investigation of soil organic matter composition for shelterbelt agroforestry systems. *Geoderma*. 2017;298:1–13. Available from: https://doi.org/10.1016/j.geoderma.2017.03.016.
15) Pedersen JA, Simpson MA, Bockheim JG, Kumar K. Characterization of soil organic carbon in drained thaw-lake basins of Arctic Alaska using NMR and FTIR photoacoustic spectroscopy. *Organic Geochemistry*. 2011;42(8):947–954. Available from: https://doi.org/10.1016/j.orggeochem.2011.04.003.
16) Ma F, Du C, Zhang Y, Xu X, Zhou J. LIBS and FTIR–ATR spectroscopy studies of mineral–organic associations in saline soil. *Land Degradation & Development*. 2021;32(4):1786–1795. Available from: https://doi.org/10.1002/ldr.3829.
17) Peltre C, Gregorich EG, Bruun S, Jensen LS, Magid J. Repeated application of organic waste affects soil organic matter composition: Evidence from thermal analysis, FTIR-PAS, amino sugars and lignin biomarkers. *Soil Biology and Biochemistry*. 2017;104:117–127. Available from: https://doi.org/10.1016/j.soilbio.2016.10.016.
18) Smith BC. How to properly compare spectra, and determining alkane chain length from infrared spectra. *Spectroscopy*. 2015;30(9):40–46. Available from: https://www.spectroscopyonline.com/view/how-properly-compare-spectra-and-determining-alkane-chain-length-infrared-spectra.
19) Saikia BJ, Parthasarathy G, Sarmah NC. Fourier transform infrared spectroscopic estimation of crystallinity in SiO2 based rocks. *Bulletin of Materials Science*. 2008;31(5):775–779. Available from: https://doi.org/10.1007/s12034-008-0123-0.
20) Calderón FJ, Reeves JB, Collins HP, Paul EA. Chemical Differences in Soil Organic Matter Fractions Determined by Diffuse-Reflectance Mid-Infrared Spectroscopy. *Soil Science Society of America Journal*. 2011;75(2):568–579. Available from: https://doi.org/10.2136/sssaj2009.0375.
21) Fakhry A, Osman O, Ezzat H, Ibrahim M. Spectroscopic analyses of soil samples outside Nile Delta of Egypt. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. 2016;168:244–252. Available from: https://doi.org/10.1016/j.saa.2016.05.026.
22) Chauhan R, Kumar R, Sharma V. Soil forensics: A spectroscopic examination of trace evidence. *Microchemical Journal*. 2018;139:74–84. Available from: https://doi.org/10.1016/j.microc.2018.02.020.
23) Kikkawa HS, Naganuma K, Kumisaka K, Sugita R. Semi-automated scanning electron microscopy energy dispersive X-ray spectrometry forensic analysis of soil samples. *Forensic Science International*. 2019;305:109947. Available from: https://doi.org/10.1016/j.forsciint.2019.109947.
24) Prandel LV, Vander Freitas Melo, Testoni SA, Brinatti AM, Saab SDC, Dawson LA. Spectroscopic techniques applied to discriminate soils for forensic purposes. *Soil Research*. 2020;58(2):151–160. Available from: https://doi.org/10.1071/SR19066.
25) Chen W, Peng L, Hu K, Zhang Z, Peng C, Teng C, et al. Spectroscopic response of soil organic matter in mining area to Pb/Cd heavy metal interaction: A mirror of coherent structural variation. *Journal of Hazardous Materials*. 2020;393:122425. Available from: https://doi.org/10.1016/j.jhazmat.2020.122425.
26) Zeng R, Rossiter DG, Zhao YG, Li DC, Zhang GL. Forensic soil source identification: comparing matching by color, vis-NIR spectroscopy and easily-measured physio-chemical properties. *Forensic Science International*. 2020;317:110544. Available from: https://doi.org/10.1016/j.forsciint.2020.110544.
27) Chauhan R, Kumar R, Diwan PK, Sharma V. Thermogravimetric analysis and chemometric based methods for soil examination: Application to soil forensics. *Forensic Chemistry*. 2020;17:100191. Available from: https://doi.org/10.1016/j.forc.2019.100191.
28) Chauhan R, Kumar R, Kumar V, Sharma K, Sharma V. On the discrimination of soil samples by derivative diffuse reflectance UV–vis-NIR spectroscopy and chemometric methods. *Forensic Science International*. 2021;319:110655. Available from: https://doi.org/10.1016/j.forsciint.2020.110655.
29) Paltseva AA, Deeb M, Iorio ED, Circelli L, Cheng Z, Colombo C. Prediction of bioaccessible lead in urban and suburban soils with Vis-NIR diffuse reflectance spectroscopy. *Science of The Total Environment*. 2022;809:151107. Available from: https://doi.org/10.1016/j.scitotenv.2021.151107.
30) Newland TG, Pitts K, Lewis SW. Multimodal spectroscopy with chemometrics: Application to simulated forensic soil casework. *Forensic Chemistry*. 2023;33:100481. Available from: https://doi.org/10.1016/j.forc.2023.100481.