Iraqi Journal of Science, 2025, Vol. 66, No. 12, pp: 5817- 5830
DOI: 10.24996/ijs.2025.66.12.43

ISSN: 0067-2904

Spectroscopy Regression Models to Predict Petroleum Contaminants in Soil

Zahraa A. Khaleel [*1], Auday H. Shaban [1], Ali A. Al Maliki [2]
[1] *Department of Remote Sensing & GIS, College of Science, University of Baghdad, Baghdad, Iraq*
[2] *Ministry of Science and Technology, Environment, water and renewable Energy directorate Baghdad, Iraq*

Received: 14/6/2024 Accepted: 3/12/2024 Published: 30/12/2025

Abstract

Hydrocarbon soil pollution is one of the most dangerous pollutants in the world. It occurs for several reasons and increases due to factories not adhering to environmental protection controls, the most prominent of which is oil production. This work used two sets of soil petroleum contamination to demonstrate principal component analysis (PCA) and partial least squares regression (PLS) modeling. To determine the variables adopted in this study based on spectroscopic analysis within the spectrum range of 1700-1800 nm and 2200-2400 nm, the distinct absorption peaks at 1720, 1750, 2220, 2300, and 2350 nm indicated the crude oil content. Chemical analysis of the samples was used to measure the relationship and build a PLS and PC model, which helped obtain a high percentage of match of up to 90%. The work indicates that this technique may enhance field investigation of oil contamination, providing an accurate in-field technique.

Keywords: Hydrocarbon, soil contamination, Remote sensing, PLS

نماذج الانحدار الطيفي للتنبؤ بالملوثات البترولية في التربة

زهراء اياد خليل [*1]،عدي حاتم شعبان [1]،علي عبد الرضا عجيل المالكي [2]
[1]قسم الاستشعار عن بعد ونظم المعلومات الجغرافية، كلية العلوم، جامعة بغداد، بغداد، العراق
[2]وزارة العلوم والتكنولوجيا، مديرية البيئة والمياه والطاقة المتجددة بغداد، العراق

الخلاصة

يعد تلوث التربة الهيدروكربونية من أخطر الملوثات في العالم، والذي يحدث لعدة أسباب ويزداد نتيجة عدم التزام المصانع بضوابط حماية البيئة، وأبرزها مصانع إنتاج النفط. في هذا العمل، تم استخدام مجموعتين من تلوث التربة بالنفط لإظهار تحليل المكونات الرئيسية (PCA) ونمذجة انحدار المربعات الصغرى الجزئية (PLS).

لتحديد المتغيرات المعتمدة في هذه الدراسة اعتماداً على التحليل الطيفي ضمن المدى الطيفي 1700 – 1800 و2200 – 2400، تشير قمم الامتصاص المميزة عند 1720، 1750، 2220، 2300، و2350 نانومتر إلى محتوى النفط الخام والتحليل الكيميائي للتربة. تم تحقيق العلاقة بين العينات الدالة على محتوى النفط الخام وبناء نموذج PLS و PCA مما ساعد في الحصول على نسبة تطابق عالية تصل إلى 90%.

*Email: Zahraa.Ayad2209m@sc.uobaghdad.edu.iq

ويشير العمل إلى أن هذه التقنية قد تعزز التحقيق الميداني للتلوث النفطي، مما يوفر تقنية دقيقة في العمل الحقلي.

1. Introduction

Hydrocarbon pollution is one of the most severe types, threatening environmental life and living organisms by producing diseases and risks. It is produced in commercial factories for various products, and the most prominent causes are oil extraction, production, and refining factories. These oil facilities extract large quantities of crude oil daily in addition to its daily use in daily life, spills, and accidents. Leakage of fuel and oil pipelines, as well as natural causes such as earthquakes and movement of rocks and earth layers [1] [2]. Oil is a group of hydrocarbon compounds consisting of bonded hydrogen and carbon atoms; the threats caused by hydrocarbons must be part of a monitoring plan to maintain the scope of pollution and treat it [3]. It is formed due to oil extraction and transportation, natural leakage, or accidents, which cause environmental destruction in addition to the problem of pollution in isolated places such as refineries and oil fields. The ASD filed spec3 (Analytical Spectral Devices) provides distinct spectral signatures for each substance; a spectroscopic analysis device provides different spectral signatures for each substance. So that the presence or absence of hydrocarbons can distinguish the soil, each giving different spectral signatures. Crude oils and petroleum fuels have absorption around 1725-2310 nm; the NIR-SWIR absorption bands of crude oils and fuels originate in clusters of saturated CH_2 and terminal CH_3 stretching patterns, or aromatic CH_3 groups. Spectral information in the NIR-SWIR band is excellent for qualitative and quantitative analysis of soils contaminated with hydrocarbons. However, the resulting spectral bands obstruct the interpretation and quantification of spectra [4] [5].
Multivariate procedures in spectroscopy or material science involve analyzing multiple variables (like absorbance spectra at different wavelengths) to determine material properties. Principal Component Analysis (PCA) or Partial Least Squares Regression (PLSR) techniques used to extract meaningful information from complex datasets, relating spectral data to material composition, structure, or properties; multivariate calibration generally resolves the problem of interference from compounds bound to the target, thus eliminating the require for selectivity PLS and PCA decompose the spectral data into components that explain the maximum variance in the predictor (spectral) variables and the response (analyte concentration) variables. This decomposition allows the model to extract the relevant information about the analyte of interest from the spectral data, even in the presence of interference from other compounds or matrix effects. Recent studies focused on using remote sensing and statistical spectral techniques to detect hydrocarbon contamination. For example, algorithms and techniques clarify the relationship between spectroscopic and chemical analysis data. The most important of these is PLSR (Partial Least Square Regression), which analyzes and extracts sample files, ensuring quality [6] [7]. Reflectance spectroscopy (RS) has been recognized as a reliable alternative technique for the direct detection of petroleum hydrocarbons (PHCs) [8] [9]. It has been acknowledged as a dependable alternative procedure for directly detecting petroleum hydrocarbons (PHCs)3. Despite not being the most common method for this purpose, RS has also proven to be a simple, quick, and cost-effective strategy for rapidly detecting and characterizing PHC-contaminated soils.

More specifically, RS in the range of (NIR-SWIR, 700-3000 nm) is directly a prevalent method for speedy recognizable proof and quantification of PHCs in contaminated soils, with sensible levels of accuracy, particularly due to the transportability of the devices and least or no planning and pre-treatments required for the samples studied used PLS in the process of analyzing data for 98 samples of contaminated soil under laboratory conditions to evaluate

hydrocarbon contamination in soil. Because spectroscopic analysis has become the focus of interest for soil scientists [10] [11]. As well as the PLS technique in detecting hydrocarbon contamination, a researcher used 150 samples from four different regions in the UK and made it possible to predict the relationship between soil, water, hydrocarbons, and soil [12] [13]. A researcher used PCA (Principal Component Analysis) and PLS with Fourier infrared analysis using 50 samples for calibration and 37 samples for verification in Australia [14] [15]. In Brazil, a researcher used two sets of samples, the first set of 3 samples and the second set of 4 samples with different quantities of oil in laboratory conditions. The soil's spectral characteristics were analyzed using LS and PCA as statistical tools to create qualitative models useful in detecting hydrocarbon pollution in areas close to oil facilities most Affected by pollution [16]. In area ONERA in France used an ASD Field Spec3(Analytical Spectral Devices) spectroradiometer and used ENVI software to calculate spectral indices for detecting the spectral that were not normal for hydrocarbon, used four boxes have soil and sand with oil where the results of the spectral signature were between 1700 and 2300 nm [17]. In another proposal to analyze soil and detect hydrocarbon contamination, infrared spectroscopy was used, along with statistical models to predict hydrocarbon content, by testing 72 samples in Australia, where results were obtained by sensing hydrocarbons in the range 2340-2300nm. [18]. In Brazil, analysis using techniques (PCA) and (PLS) for samples from two Gather I is composed of three samples of crude oils. Bunch II comprises six samples of mineral substrates (MS) research utilizing FieldSpec3 spectrometer (ASD), ENVI software, and the Unscrambler X 10.1 software, using the Savitzky-Golay filter [19]. The study aims to identify hydrocarbon contamination accurately and its presence to aid in prediction and to study the variation and correlation between spectroscopic analysis data and chemical analysis data using PLS and PCA techniques to serve as a quick monitoring of soil in areas close to oil facilities.

2.1 Study area

The soil samples were collected from the North Rumaila Oil field in southern Iraq at Basra Governorate, at the coordinates from 47° 16' 23.271"E, 30° 42' 45.769"N to 47° 21' 55.343"E, 30° 35' 10.267"N. Lime content and oil activity characterize the soil of the study area due to the presence of oil and the processes of extracting and developing oil facilities. It produces high concentrations of hydrocarbons and their various compounds. The map of the study area and sampling locations are shown in Figure 1.

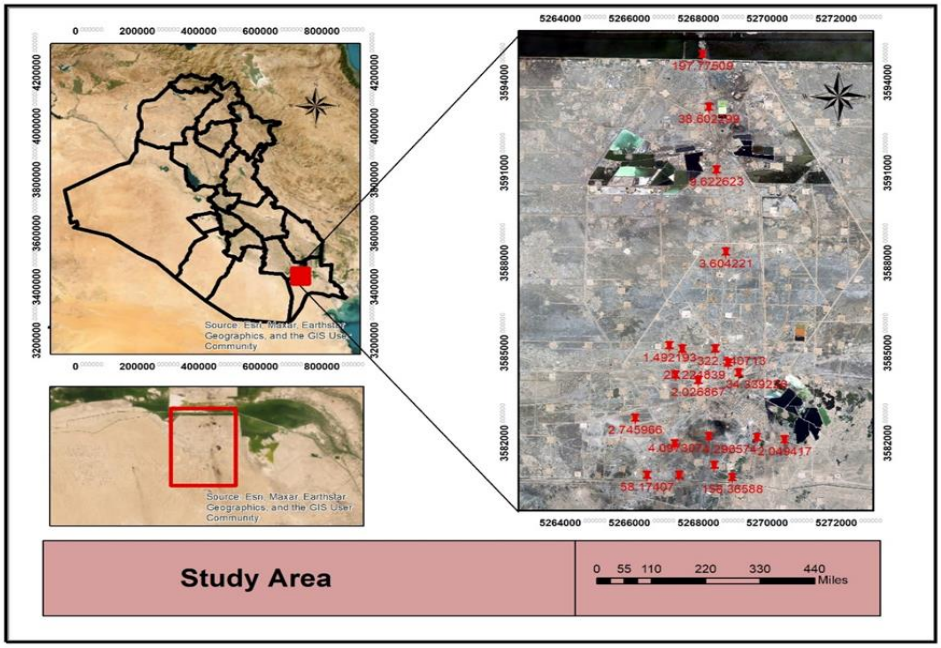

Figure 1: Study area and sampling locations

2.2 Soil samples and mixtures

This study utilized two samples; the training set consists of 22 samples taken directly from the study area's top surface soil (0-10 cm). The site location and total hydrocarbon in ppm according to gas chromatography GC mass device instrument appears in Table 1.

Table 1: Filed sample's (F) location and Hydrocarbon Concentration (training set)

Sample	Long E	Lat N	HCs (ppm) (GC result)
F 1	47 19 29	30 40 32	9.62
F 2	47 19 38	30 39 16	3.60
F 3	47 19 28	30 37 46	322.54
F 4	47 19 12	30 37 17	2.02
F 5	47 19 12	30 37 17	27.60
F 6	47 18 51	30 37 22	20.33
F 7	47 19 22	30 36 25	4.09
F 8	47 19 27	30 35 58	31.36
F 9	47 19 44	30 35 47	158.36
F 10	47 20 07	30 36 24	4.29
F 11	47 20 33	30 36 22	2.04
F 12	47 18 50	30 36 18	213.93
F 13	47 18 13	30 36 42	2.74
F 14	47 18 54	30 35 49	126.86
F 15	47 18 24	30 35 49	58.17
F 16	47 19 40	30 37 33	20.22
F 17	47 19 50	30 37 24	146.85
F 18	47 18 57	30 37 46	0.78
F 19	47 18 45	30 37 49	1.49
F 20	47 19 22	30 41 30	38.60
F 21	47 19 16	30 42 19	197.77

The second is the test set, a set of samples prepared in a laboratory from crude oil added to soil free of hydrocarbons. Table 2 shows the calibration experiments that were carried out for the soil test set. This dynamic included the addition of 5 mL of crude oil up to a fixed 20 mL volume of each dry soil mixture coming about in a range of oil concentrations from 5-30%. The oil-containing soil was thoroughly mixed using a glass bar at each dosing, and the surface was straightened before spectral filtering. Readings were taken utilizing an ASD trumpet fore optic at three separate points on the surface, and the middle value was found to get a representative spectrum for a given concentration for each sample, Table 1. The proportions of crude oil in the laboratory-which combined with uncontaminated soil were converted from mg to ppm to standardize the units in the following steps:

- Convert the additive volume (5,10,15,20 ml) into weight using a specific density.
- Mass of substance = volume × specific density.
- Now, the concentration of the substance in ppm can be calculated using the mass of the substance and the mass of the soil (500 g).
- $ppm\ concentration\ =\ (mass\ of\ substance\ /\ mass\ of\ soil)\ x\ 10\text{^}6$ (1)
- calculated to obtain the final value of the concentration of the substance in ppm.

Table 2: Test (lab) samples set

Lab samples	Soil weight	Crude Oil To add	Hydrocarbon concentration in ppm
S1	500 g	20 ml	34
S2	500 g	15 ml	25.5
S3	500 g	10 ml	17
S4	500 g	5ml	8.5

2.3 Spectral Data Acquisition

The soil samples passed through a series of pre-processing operations where they are Samples were dried at 150°c measuring soil reflectance; the soil was homogenized, beat with a mortar to remove any wetness impact, sieved with a 2 mm work to remove any unpleasantness that would affect the soil's total reflectance, and then put into Petri dishes with an 8 cm breadth and 1.5 cm thickness. With a wavelength range of 350–2500 nm, the portable spectroradiometer ASD Field Spec3, Figure (2) was utilized to require spectral measurements. The spectroradiometer's spectral arrangement was separated into the VIS (350–700 nm), NIR (700–1300 nm), SWIR1 (1300–1800 nm), and SWIR2 (1800–2500 nm) spectral regions.

Figure2: Sample identification of the samples and spectroradiometer device were used to detect signatures for soil samples' reflectance

2.4 Sampling analysis

Two kinds of analyses were applied: spectroscopy analysis and chemical analyses. The ASD filed spec3 device works within a spectral range between 400 and 2500 nm and was used to detect the reflectance spectroscopy of each set of samples. Two spectra ranges were used to detect hydrocarbon signatures within 1700-1800 and 2200-2400 nm, clarified in past research [4]. However, for chemical analysis, soil samples are extracted using hexane and transformed into liquid. Second, they were injected into a GC mass device to give the concentration of the samples in a chart that contains a table for each sample consisting of the hydrocarbon compounds in the sample, which were calculated using an analytical equation given at:

$$C_{sample} = \frac{A_{sample}}{A_{st}} * C_{st} \quad (2)$$

2.5 Pre-processing methods

Data processing was used to reduce the physical effects, remove spectral data variation, and treat scattering in light. Log10 was utilized to convert spectral reflectivity values to more accurate data; log transformation compresses large values more than small ones, reducing the impact of extreme values. In datasets with a wide range of values, this diminishes the contrasts between high and low factors, making the data less dominated by extreme points. Free of distortions and more reasonable for dealing with statistical operations. Data were log10 transformed prior, and the new matrix was analyzed. Changing overall data to a log scale diminished contrasts between the factors due to the estimation units and result ranges. This can result in more stable and interpretable models, mainly when there is skewness in the distribution. The second processing of data is Orthogonal Signal Correction (OSC), which is a transformation and pre-processing technique for analysis operations that deal with spectral data to ensure the quality and accuracy of the results; the OSC transformation is applied to the test and training data, it works to remove the discrepancy between the x matrix variables that are orthogonal to the y variables. The main idea is to remove the variance unrelated to the main variables in the work, which leads to clarity and stability in regression models. It was used to make the PLS model more accurate as it is used in applications on near-infrared data [20]. OSC enhances the signal-to-noise ratio, allowing the PLS model to capture the true relationships between the predictors and the response variable. This is particularly important when dealing with noisy data, where unrelated factors might overshadow the predictive signal [21].

When light interacts with a sample during measurement, light scatters and spreads in different directions. Another processing method was the Multiplicative Scatter Correction MCS process, which analyzed multivariate data dealing with chemical and spectroscopic data, especially infrared and near-infrared data. It works to remove differences associated with scattering in the data that affect the quality of the analysis, which allows for more accurate analysis and facilitates the identification of relevant patterns.

2.6 Regression models

The Unscrambler X 10.5.1 (Camo Software, 2022) was used in this research, a comprehensive software package for applied multivariate data analyses and evaluation regression models. The regression used was Partial least square regression PLSR, which consists of an algorithm that designs a matrix between input variables x (spectral data) and output variables y (hydrocarbon concentration). It is based on modeling data relationships. PLSR technology assists in predicting hydrocarbons in soil. It is one of the modern methods for dealing with linear data, as it works to simplify the relationship between variables because

it depends on the inherent change between the matrices. Where x represents spectral data, and y represents hydrocarbon concentration.

Hydrocarbons show absorption properties in the spectral range 1700-1800 2200-2400 nm. Chemical analysis techniques were often used to create an analysis model to link it with the spectral data to know the concentrations of hydrocarbons in the soil through the spectral signature of each sample, as the models indicate a link between two matrices. The first represents the spectral data of the samples, which is complex and large, and the second matrix of chemical analysis data forms a relationship linking the smallest number of equivalent factors between X and Y [23]. PLS is a powerful multivariate analysis technique utilized for modeling relationships between sets of variables. Understanding the process of explained variance is fundamental for interpreting the efficacy of PLS models.

Principal Component Analysis (PCA) is a statistical technique for simplifying complex datasets. It transforms the data into a new coordinate system where the directions (called principal components) capture the most significant variations within the data. These principal components represent the most important underlying factors explaining the original variables' relationships. PCA was achieved by calculating the eigenvectors and eigenvalues of the covariance matrix of the main variables. Variability in Hydrocarbon Concentrations: Soil samples often contain multiple hydrocarbon compounds, each varying in concentration across different samples. PCA reduces the dimensionality by identifying patterns or principal components that capture the greatest variation in these concentrations. For example, one principal component might represent the collective variation of heavier hydrocarbons, while another might capture the behavior of lighter hydrocarbons. The method effectively reduces dimensionality while preserving the most significant information, facilitating the visualization of differences between groups of samples through charts.

2.7 Validation

Regression coefficient or the linear regression slope, standard errors (SE), correlation R^2 (Pearson), Root mean squire error RMSE, and standard deviation (SD) are the essential statistics tools used to verify and validate the regression analyses. Statistical variants are significant when applying the most acceptable criteria for the model. Another measurable parameter of expectation accuracy considered was the proportion of execution to deviation (RPD), which is the proportion of the standard deviation (SD) of the reference values to the RMSE (Eq. (2))

$$RPD = SD/RMSE \quad (3)$$

Some statistical regression was used to evaluate the relative accuracy of the models used in hydrocarbon analysis. In common, the lowest RMSE, SE, inclination, balanced (intercept) with higher RPD and an R^2 coefficient of regression near 1.0 were utilized as pointers of the foremost accurate regressions.

3. Results and Discussion

3.1 Spectral signatures

Spectral analysis of petroleum-contaminated soil samples is primarily focused on two spectral ranges: 1700–1800 nm and 2200–2400 nm. The spectral curves for the laboratory test set and the field training set are shown in Figures 3 and 4, respectively. These two ranges' specified spectralges agree with [4] and [22]. They found strong relationships between hydrocarbon contamination and spectra in these regions. It is pertinent to explain that wide absorption peaks around 1900–2100 nm are typically associated with the presence of water. As appeared in Figure 3, they were recognized primarily by the spectral geometry between 1700 and 2400 nm. The characteristic absorption top at 1720, 1750, 2220, 2300, and 2350 nm indicates crude oil content.

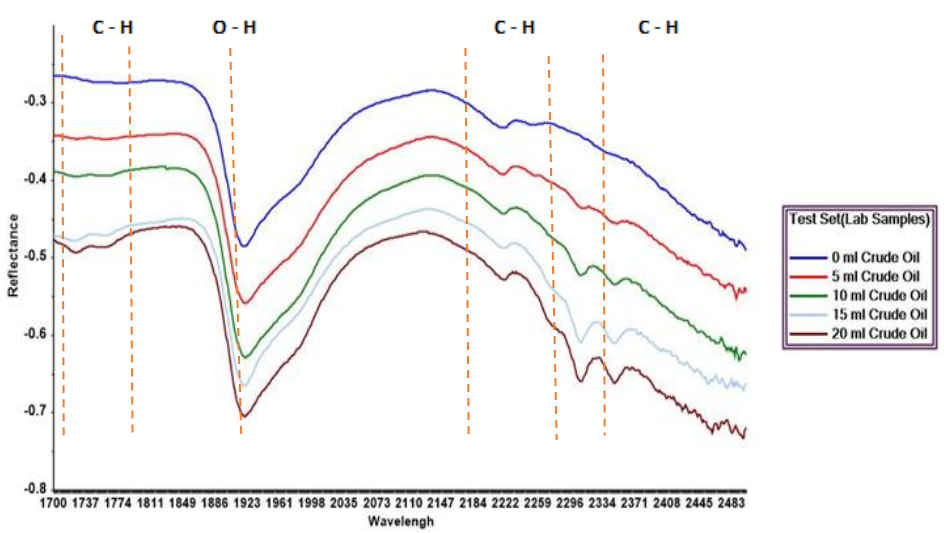

Figure 3: Spectral reflectivity of hydrocarbons in LAP sample

The spectra in Figure 3 illustrate the calibration tests carried out in both free and mixed soils, with sample contents indicating an inverse relationship between spectra and crude oil content, meaning the sample with high crude oil concentration shows greater absorption**.**

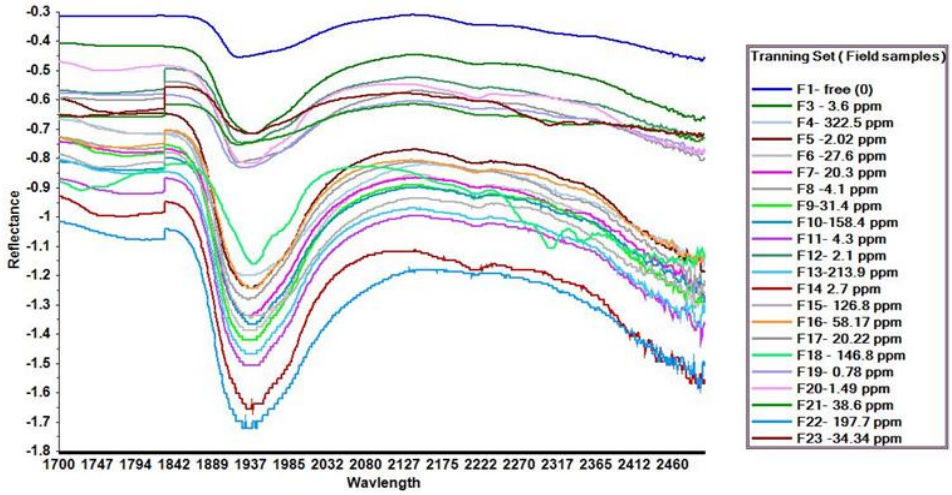

Figure 4: Spectral reflectivity of hydrocarbons in field sample

Figure 4 illustrates the irregularity of the spectral curve for hydrocarbon absorption in field soil. The assumption about soil models with higher hydrocarbon concentrations having a higher absorption spectrum curve is interesting [23]. Irregularity of the absorption spectra in soil can have several causes, even in soils of the same quality and texture. Among these reasons:

- Uneven distribution of hydrocarbons: Hydrocarbons may not be evenly distributed in different soil samples, leading to differences in spectral absorption.
- Interference with other materials: Different materials in the soil, such as organic materials or metals, can interfere with the absorption of hydrocarbons, causing irregularities in the spectral curve.
- Humidity and environmental changes: Humidity and environmental changes, such as temperature, can affect the spectral absorption of hydrocarbons.
- Homogeneity in sampling: The sampling process may be heterogeneous, or the samples may not represent the whole soil, leading to differences in spectral measurements.

• Cross-contamination: Contamination can occur between samples during collection or analysis, leading to irregular results in the spectrum.
• Chemical Changes: Hydrocarbons can change chemically over time or due to interactions with other components in the soil, affecting spectral absorption.
To address these issues, sample collection and analysis procedures can be improved, and advanced techniques can be used to ensure the accuracy of measurements and reduce the influence of interfering factors**.**

3.2 Principal Component Analysis (PCA)

A Principal Component Analysis (PCA) was conducted to distinguish between crude oil-contaminated and uncontaminated soil samples. The PCA results demonstrated that the models accounted for a significant proportion of the variance. Specifically, the independent data matrix (spectral data) showed that the principal components (PC1 and PC2) clarified most of the variance within the soil samples in both sets. This spectral differentiation effectively separated the two components, indicating that soil samples inside a cluster were similar in terms of soil content and soil texture, Figure 5.
For the test set (Lab samples) in Figure 5a, the PCA results revealed that PC1 (plotted on the x-axis) accounted for 99% of the variance, while PC2 (plotted on the y-axis) accounted for 1%. Thus, the combined representation of these two components explained 100% of the total variance. For the training set (field samples) in Figure 5b, the PCA results accounted for 98% of the variance, while PC2 for 1%. Thus, the combined representation of these two components explained 99% of the total variance. Figure (5) shows distinct segregation and clustering between the two groups of samples; the first PCs described most of the observed variance.

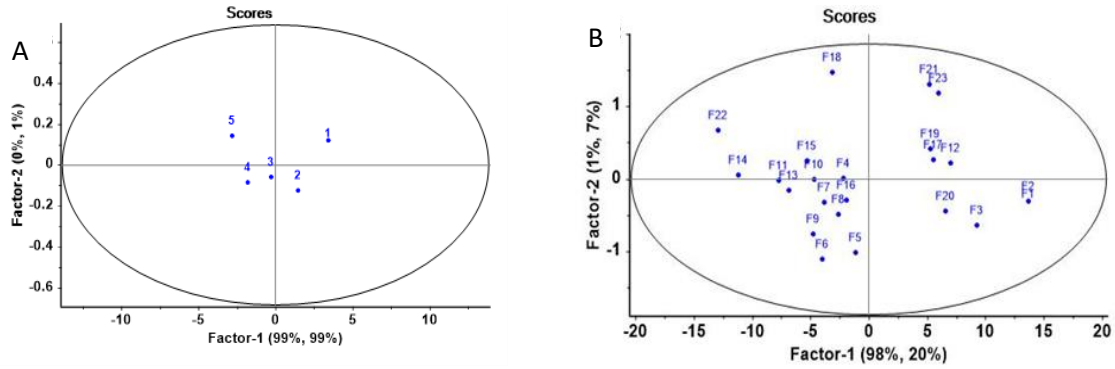

Figure 5: PCA score plots with samples categorized: (a) for test (lab) set, (b) for training (field) set

If the principal components separate the contaminated from uncontaminated samples, PCA can be an efficient, unsupervised method for identifying contamination patterns. However, if there is overlap between the groups, it may suggest that additional variables, more sensitive
techniques, or further preprocessing of the data are needed to improve the differentiation. In conclusion, the effective separation of samples via PCA indicates that it is a powerful tool for identifying contamination-related patterns.

3.3 PLS regression models

The Explained Variance in PLS modeling within PLS modeling using Unscrambler X is a powerful tool for understanding, interpreting, and refining predictive models for hydrocarbon

contamination in soil. By unraveling the intricate relationships between predictor variables and contamination levels, the results obtained from the spectral analysis and the regression model are consistent with the observed hydrocarbon contamination in the soil.

To correlate the spectral analysis and regression model results with observed hydrocarbon contamination, the following key steps and findings can be considered:

1. Spectral Analysis Findings:

Peak Identification: Spectral analysis identifies specific wavelengths or frequency ranges where hydrocarbons absorb or reflect electromagnetic radiation. Specific absorption bands in the infrared or UV spectra (e.g., around 3.4 μm for CH stretching in hydrocarbons) could be identified for hydrocarbon contamination. Signal Strength and Contamination Levels: the intensity of these peaks could be directly correlated with the concentration of hydrocarbons. Higher absorption indicates higher levels of contamination.

Spatial Distribution: spectral data often reveal the spatial distribution of contamination. By mapping the intensity of specific hydrocarbon absorption features, areas with high contamination can be visually identified.

2. Regression Model Findings:

Predictive Power (R^2 Value): The R^2 value from the regression model quantifies how well the model predicts hydrocarbon contamination based on spectral data. A high R^2 value (close to 1) indicates a strong correlation between the spectral features and contamination levels.

Significant Predictors: In partial least squares regression (or similar models), the loading weights can highlight which spectral bands most predict contamination. This identifies the key wavelengths related to hydrocarbon absorption.

Model Performance Metrics: Additional metrics like RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) help assess how accurately the model predicts contamination levels across different areas or datasets.

3. Correlation with Observed Hydrocarbon Contamination:

Comparison with Field Data: The spectral analysis results and the regression model should be validated against field measurements of hydrocarbon concentrations. The strength of the correlation was determined by comparing predicted contamination levels with actual samples from affected sites.

Temporal Trends: If temporal data is available, it can show how contamination evolves over time and whether the spectral features and model predictions are sensitive to these changes.

Summary of Key Findings:

Strong correlations between specific spectral bands (e.g., near 3.4 μm) and hydrocarbon contamination were observed.

The regression model demonstrated high predictive accuracy with an R^2 value of X (insert specific value) and low RMSE of Y (insert specific value), indicating reliable prediction of hydrocarbon levels.

The analysis revealed hotspots of contamination that matched with observed field data, reinforcing the link between spectral signatures and contamination intensity.

This strengthens the argument that spectral analysis combined with regression modeling can effectively detect and predict hydrocarbon contamination in environmental settings.

The PLS technique can help accurately and accurately detect polluted sites on a large scale, especially in areas close to oil facilities and oil production and refining plants; it empowers stakeholders to make informed decisions crucial for environmental management and remediation efforts.

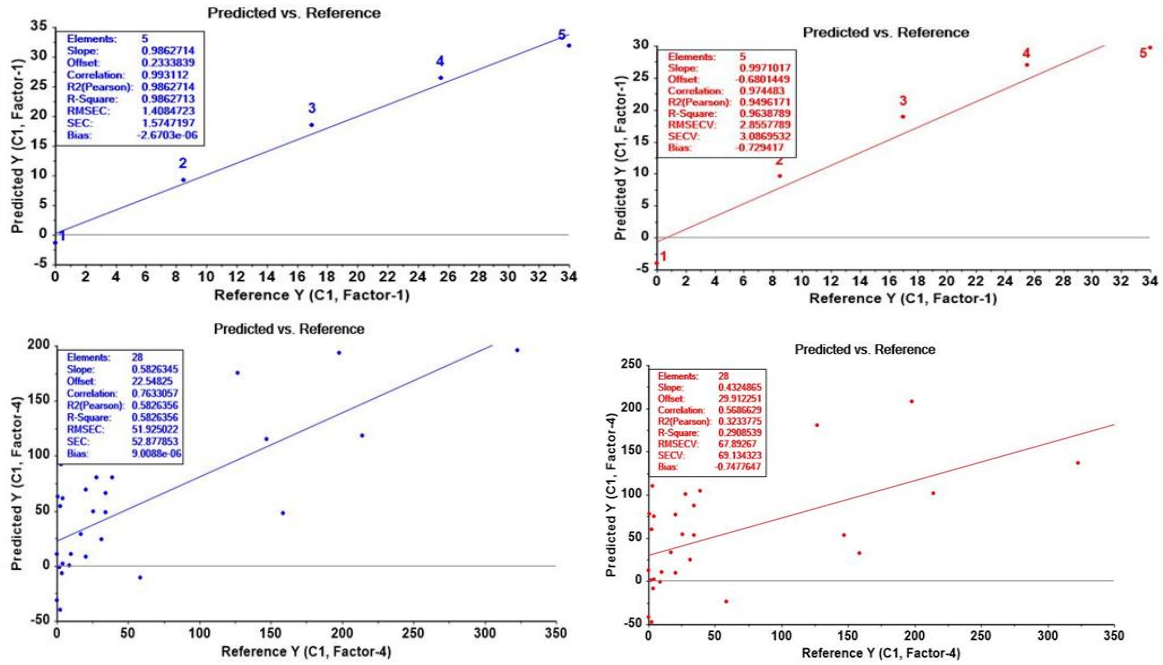

Figure 6: Relationship between anticipated and HCs concentration to PLSR models for Test Set (lab samples) (upper plots) and training (field samples) set (lower plots) within the calibration (cleared out) and validation (right) models.

Correlation coefficients were calculated to assess the relationship between hydrocarbon contamination in soil samples and their corresponding spectral data. The relationship between the normalized spectral information and hydrocarbon substances, which have distinct spectral signatures within the VNIR locale, appears in Table 3 Significant Pearson correlations (R^2). Since HCs are not spectrally active within the reflectance spectra for the overall set of soils considered here (n = 23), a poor expectation of HCs concentration, showing low R^2 values and RMSE > 50%, was obtained utilizing direct reflectance spectra.
However, the Log10 algorithms inside the Unscrambler X computer program were utilized to identify and remove calibration exceptions, and new calibrations were created at that point. Despite HCs predicted from regression models for the training set (field samples), the hydrocarbons predicted from regression models for the training set (field samples) showed some accuracy parameters in the validation mode. The Total validation parameters shown in Table (3) were utilized to detect hydrocarbon pollution concentration in two sets of soils.

Table 3: validation statistics parameters related to PLS regression models were utilized to anticipate HCs concentration in two sets of soils; the certainty level of $p < 0.005$ was used in all prediction models.

PLSR with log10 treatment	*No. of samples*	*PLS components*	*Calibration*		*Validation*	
			R^2	**RMSE**	**R^2**	**RMSE**
TEST SET	5	4	0.986	1.408	0.949	2.855
TRAINING SET	22	7	0.582	51.925	0.323	67.892

Generally, a few parts per million (ppm) of oil concentration is sufficient to obtain a clear spectral signature. Some studies indicate that concentrations of hydrocarbons in soil ranging from 10 to 100 ppm could be sufficient to obtain a distinctive spectral fingerprint using advanced spectroscopy techniques [24]. However, to obtain accurate and reliable results, field

and laboratory experiments must be performed to determine the optimal concentration based on the type of soil, hydrocarbons, and technology used [25].
For example, research comparing reflectance spectroscopy and solvent extraction methods for analyzing total petroleum hydrocarbons (TPH) in soil indicated that spectral methods could effectively detect and differentiate hydrocarbon concentrations as low as 50 ppm. Furthermore, various studies have confirmed that these concentrations are adequate for identifying and quantifying hydrocarbons in contaminated soil samples using infrared (IR) spectroscopy techniques. IR techniques provide distinct advantages in sample preparation, resolution, and data acquisition, making them highly relevant in different applications such as chemical identification, material characterization, and environment quality control [26]. These findings underscore the potential of spectral methods for effective environmental monitoring of hydrocarbon contamination.

4. Conclusion

This study demonstrated the possibility of using remote sensing techniques to determine petroleum pollution in soil. Two soil sets were used, and various preprocessing transformations enhanced the spectral data. The data were inspected utilizing Principal Component Analysis (PCA) and Partial Least Squares (PLS) Regression. Interesting and characteristic absorption features were recognized within the mixture of soil with the crude oils in the range between 1700-1800 2200-2400 nm. All samples with different concentrations of hydrocarbons showed a distinct reflectivity and sensitivity to hydrocarbons in soil samples within the range 1700-1800 2200-2400 nm. They are recognized primarily by the spectral geometry between 1700 and 2400 nm. The characteristic absorption peaks at 1720, 1750, 2220, 2300, and 2350 nm were indicative of crude oil content; PCA models connected to the spectral signature demonstrated the ability to distinguish the density of the hydrocarbons. The calibration models produced by PLS are vigorous, of high quality, and can be utilized to anticipate the concentration of crude oils in mixtures with soil. Such data and models are employable as a reference for classifying obscure samples of contaminated substrates. Overall, this research provides a foundational approach for advancing methodologies in the environmental monitoring of petroleum contamination.

5. Acknowledgements

References

[1] J. S.Muhammad,K. A.Jasim and A. H.Shaban, "The Concentration of the Toxic Elements (Cd, Hg, As) in Diyala Governorate Soil Utilizing GIS Techniques," *Journal of Physics: Conference Series,* p. 032063, 2021.

[2] M.M. Arce ,S.Sanllorente, S. Ruiz , M.S. Sánchez , L.A. Sarabia , M.C. Ortiz, "Method operable design region obtained with a partial least squares model inversion in the determination of ten polycyclic aromatic hydrocarbons by liquid chromatography with fluorescence detection," *Journal of Chromatography A,* vol. 1657, p. 462577, 2021.

[3] R.Pelta, E.Ben-Dor, "Assessing the detection limit of petroleum hydrocarbon in soils using hyperspectral remote-sensing," *Remote Sensing of Environment,* vol. 224, pp. 145-153, 2019.

[4] Z. A. Khaleel , A. H. Shaban , A. A. Al Maliki, "Investigations for Soil Contamination with Hydrocarbon Compounds near," *Journal of Physics: Conference,* p. 2754 012025, 2024.

[5] J. S. Muhammad,K. A. Jasim ,A. H. Shaban, "Evaluation of (Ni, Cr, Cu) Concentration in the Soil of Diyala Utilizing GIS Techniques," *IOP Conf. Series: Materials Science and Engineering,* p. 072111, 2020.

[6] H. Liang, G. Liu, "Research on quantitative analysis method of PLS hydrocarbon gas infrared spectroscopy based on net signal analysis and density peak clustering," *Measurement,* vol. 188,

p. 110392, 2022.

[7] T. Miao, N. Sihota, F. Pfeifer, C. McDaniel, M. D. Neves, and H. W. Siesler, "Rapid Determination of the Total Petroleum Hydrocarbon Content of Soils by Handheld Fourier Transform Near-Infrared Spectroscopy," *Analytical Chemistry,* vol. 95, no. 17, pp. 6888-6893, 2023.

[8] T. Lammoglia , C. R. de Souza Filho, "Spectroscopic characterization of oils yielded from Brazilian offshore basins: Potential applications of remote sensing," *Remote Sensing of Environment,* vol. 115, pp. 2525-2535, 2011.

[9] G. Schwartz, G. Eshel, E. Ben Dor., "Reflectance spectroscopy as a tool for monitoring contaminated soils," *Soil Contam,* p. 6790, 2011.

[10] O.O. Olatunde ,S. L. Della Tan ,K.A. Shiekh, S. Benjakul, N.P. Nirmal, "Oladipupo Odunayo Olatunde , Steffi Louisa Della Tan , Khursheed Ahmad Shiekh , Soottawat Benjakul , Nilesh Prakash Nirmal," *Food Chemistry,* vol. 341, p. 128251, 2021.

[11] K. A. Olatunde, "Determination of petroleum hydrocarbon contamination in soil using VNIR DRS and PLSR modeling," *Heliyon,* vol. 7, pp. 2405-8440, 2021.

[12] R.N. Okparanma ,A. M. Mouazen, "Visible and Near-Infrared Spectroscopy Analysis of a Polycyclic Aromatic Hydrocarbon in Soils," *The Scientific World Journal,* vol. 2013, p. 9, 2013.

[13] K.Kareem, "Crude Oil Spillage and the Impact of Drilling Processes on the Soil at Rumaila Oil Field- Southern Iraq," *Iraqi Journal of Science,* vol. 57, pp. 918-929, 2016.

[14] M. J. Adams , F. Awaja , S. Bhargava , S. Grocott , M. Romeo, "Prediction of oil yield from oil shale minerals using diffuse reflectance infrared Fourier transform spectroscopy," *Fuel,* vol. 84, no. 14–15, pp. 1986-1991, 2005.

[15] N. A. Al-Ridha ,G. H. AL-Sharaa ,R. A. Altimimi, *Iraqi Journal of Science,* vol. 57, pp. 7272-2732, 2016.

[16] R. E. C. Pabón, C. R. de Souza Filho, "Spectroscopic characterization of red latosols contaminated by petroleum-hydrocarbon and empirical model to estimate pollutant content and type," *Remote Sensing of Environment,* vol. 175, pp. 323-336, 2016.

[17] V. Achard ,P. Foucher, D. Dubucq, "Hydrocarbon Pollution Detection and Mapping Based on the Combination of Various Hyperspectral Imaging Processing Tools," *Remote Sensing,* vol. 13, no. 5, p. 1020, 2021.

[18] W. Ng, B.P. Malone, B. Minasny, "Rapid assessment of petroleum-contaminated soils with infrared spectroscopy," *Geoderma,* vol. 289, pp. 150-160, 2017.

[19] R.D.P.M. Scafutto, C. R. De Souza Filho, "Quantitative characterization of crude oils and fuels in mineral substrates using reflectance spectroscopy: Implications for remote sensing," *International Journal of Applied Earth Observation and Geoinformation,* vol. 50, pp. 221-242, 2016.

[20] R. Laref, D. Ahmadou, E. Losson, M.Siadat, "Orthogonal Signal Correction to Improve Stability Regression," *Journal of Sensors,* vol. 2017, p. 8, 2017.

[21] S. Wold, J. Trygg, A. Berglund, H. Antti, "Some recent developments in PLS modeling," *Chemometrics and Intelligent Laboratory Systems,* vol. 58, no. 2, pp. 131-150, 2001.

[22] H. S. Bingari, A. Gibson, E. Butcher, R. Teeuw , F. Couceiro, "Application of near infrared spectroscopy in sub-surface monitoring of petroleum contaminants in laboratory-prepared soils," *Soil and Sediment Contamination: An International Journal,* vol. 32, no. 4, pp. 400-416, 2023.

[23] P. Shi, Q. Jiang ,Z. Li, "Hyperspectral Characteristic Band Selection and Estimation Content of Soil Petroleum Hydrocarbon Based on GARF-PLSR," *Journal of imaging,* vol. 9, no. 4, p. 87, 2023.

[24] G.Shin, R.Yang, N. Zhao,G. Yin ,J. Yang, Y. Jiang, W.Liu, "Rapid Detection of Total Petroleum Hydrocarbons in Soil Using Advanced Fluorescence Imaging Techniques," *American Chemical Society,* vol. 9, no. 27, p. 29350–29359, 2024.

[25] A. M. Kadim, W. R. Saleh, "Morphological and Optical Properties of CdS Quantum Dots Synthesized with Different pH values," *Iraqi Journal of Science,* vol. 58, no. 3, pp. 1207-1213, 2017.

[26] G. Schwartz ,E. Ben-Dor, G. Eshel, "Quantitative Analysis of Total Petroleum Hydrocarbons in Soils: Comparison between Reflectance Spectroscopy and Solvent Extraction by 3 Certified Laboratories," *Applied and Environmental Soil Science,* vol. 2012, p. 751956, 2012.