Soil oil content hyperspectral model in Gudong Oilfield

FAN Yanguo, ZHANG Lei

School of Geosciences in China University of petroleum, Qingdao 266555, *China*

Abstract: The detection of oil content in soil has an important practical significance in oil pollution prevention and control. We measure the hyperspectral reflectivity and the oil content for soil samples in Gudong Oilfield. Using variable forecast model and stepwise regression method, we analyze the linear and nonlinear relationships between soil spectral characteristic parameters and oil content. The experiment shows that there is a significant correlation between the third broken line segment slope of envelope line analysis and the oil content. The cubic function of this section slope is the best single variate estimation model. The standard normal variate transformation has the best effect on spectrum pretreatment. When the transformed spectral are used to build multivariate regression model, the adjusted coefficient of determination $\overline{R}^2$ is 0.826, and the total RMSE is 0.531, which is the best forecast model. The method of using hyperspectral data to detect the oil content will provide an effective new way for detecting the oil pollution in soil.

Key words: Gudong Oilfield, soil, oil, hyperspectral

CLC number: TP79 **Document code:** A

Citation format: Fan Y G and Zhang L. 2012. Soil oil content hyperspectral model in Gudong Oilfield. Journal of Remote Sensing, 16(2): 378–389

1 INTRODUCTION

Shengli Oilfield is the second largest oilfield in China, which located in the Yellow River Delta. With the exploitation of oil, the area of the oil-contaminated soil is of 24 percent of the total area there. Ground crude oil is the most serious pollutant. In this region, the mass fraction of oil is 9.2—180.9 mg/kg (He, 2006) and the mass fraction of PAH in sediments is 0.0108—0.252 mg/kg (Hui, et al., 2009). After the oil released into the soil, it will damage the soil structure and permeability. At the same time, it can also change the composition and structure of soil organic matters and reduce soil quality (Zhang, et al., 2008). When the oil content reaches a high level, growth of the plant will be inhibited (Xu & Wang, 2006). Furthermore, certain components of oil hydrocarbons even accumulate in the food, which will endanger human health through the food chain (Chen, et al., 2003). We should detect and assess the oil pollution in soil, so as to minimize and reduce the oil pollution, which will prevent the ecological environment damage and human health hazards.

As the result of the oil pollution, hydrocarbons in the soil become abnormal. In other words, the content of oil hydrocarbons increases. Zhu (1994) points out in his study that the four absorption bands of oil hydrocarbon spectrum are: 1.69—1.97 μm, 2.27—2.46 μm, 3.33—3.53 μm and 6.80—7.48 μm. The double absorption peaks near 1748 nm and 2330 nm are the characteristics to estimate whether the soil contains oil hydrocarbons or not (Feng, et al., 2002). Traditional evaluation for oil pollution needs collecting soil samples and testing the oil content. After that, evaluate and chart for the oil pollution in the soil, according to the test result.

This traditional method is time-consuming, laborious and expensive. In this paper, Gudong Oilfield is chosen as the study area. After spectrometry and analysis on the samples in this area, we set the segment slopes of soil reflective spectrum, standard normal variable transformation, and low differential as independent variables and set the content of oil as induced variable. We then use the mathematical statistics to determine the corresponding relationship between oil content and spectral parameters, evaluating the feasibility of estimating the oil content from the hyperspectral data directly.

2 OVERVIEW OF THE STUDY AREA AND DATA PERPARA-TION

2.1 Study area

The study area, Gudong Oilfield, is located in Kenli County, City of Dongying, Shandong Province, in the north mudflat of the estuary of The Yellow River in China. Gudong Oilfield is an oil picking factory whose annual output of crude oil is 2.6 million tons. According to an analysis through field survey method in the year of 2001, the value of the oil content in the soil is 61.67 mg/kg and the average is 236.40 mg/kg (Liu, et al., 2003). Compared with the general status in Yellow River delta, soil here has higher oil content and is serious polluted. In some samples, the oil content is even more than 500 mg/kg, which illustrates the extent of pollution.

Received: 2011-01-20; **Accepted:** 2011-05-20

First author biography: FAN Yanguo(1965—), male, Ph.D. , associate professor. He is engaged in 3S technology in the teaching and research work of digital homeland, cities and coastal zones direction. E-mail: fanyanguo2003@163.com

2.2 Collecting soil samples

In order to collect soil samples evenly in the study area, firstly we referred to the map from Google Earth and the schematic diagram of Gudong Oilfield to make our route. Then our research group surveyed and sampled in Gudong Oilfield in 2010-05-03. During the time of sampling, the sampling route was adjusted properly due to the oil well arrangement, leakage of oil, traffic problems and some other reasons, so the measured soil samples were slightly different from the predicted ones. Finally, we acquired 42 samples in total from the soil surface in the study area. These samples were air-dried indoors and the dried clods were broken into pieces, then we sifted them through 2 mm screen. After that, we sent these samples to Dongying Supervision and Examination Center for Agricultural Products in May 10, and measured the oil content in May 15. There were two outliers (GDQ-9 and GDQ-20) in these 42 samples, whose concentrations were much higher than other soil samples'. Through analysis we found that GDQ-9 and GDQ-20 contained black solid oil. The remaining 40 sets of data were considered as experimental data. Its original form and descriptive statistics were shown in Table 1 and Table 2.

Table 1 Original record descriptive statistics for oil content in the soil

Number	Oil content /(mg/kg)	Number	Oil content /(mg/kg)	Number	Oil content /(mg/kg)
GDQ-1	694.3	GDQ-15	363.1	GDQ-29	659.5
GDQ-2	295.2	GDQ-16	276.7	GDQ-30	2894.7
GDQ-3	166.6	GDQ-17	704.1	GDQ-31	1154.6
GDQ-4	2882.5	GDQ-18	359.4	GDQ-32	769.5
GDQ-5	225.2	GDQ-19	489.9	GDQ-33	232.0
GDQ-6	472.9	GDQ-20	17434.8	GDQ-34	226.7
GDQ-7	214.8	GDQ-21	191.6	GDQ-35	2854.8
GDQ-8	990.0	GDQ-22	12.7	GDQ-36	2233.8
GDQ-9	20622.0	GDQ-23	78.9	GDQ-37	127.7
GDQ-10	273.7	GDQ-24	24.5	GDQ-38	226.6
GDQ-11	279.6	GDQ-25	247.3	GDQ-39	3882.4
GDQ-12	618.1	GDQ-26	5129.2	GDQ-40	785.7
GDQ-13	94.2	GDQ-27	364.4	GDQ-41	4370.1
GDQ-14	139.0	GDQ-28	1437.0	GDQ-42	824.2

Table 2 Descriptive statistics for oil content in the soil

Quantity of samples	MIN /(g/kg)	MAX /(g/kg)	Average /(g/kg)	Standard deviation
40	0.0127	5.1292	0.9567	1.2747

2.3 Acquisition of spectral data

We carried through spectral test with the 40 samples. In the experiment, we used AvaField3 spectroradiometer (Avantes, Netherlands) to collect the spectrum of the soil samples. The spectral range of the spectrometer was from 1000 nm to 2500 nm with 10 nm resolution and 6 nm spectral sampling interval. The light source we used during the experiment was 1000 W halogen light. Based on previous studies, the incident angle of the light, the distance between light and the soil sample, the angle of the probe, the distance between the probe and the soil surface and some other factors can affect the experimental data in the hyperspectral experiment in laboratory. According to the study result of the indoor geometric test conditions' influence on hyperspectral quality and the physical parameters of the spectrometer (Zhou, 2004), we set the indoor test geometric test conditions: the incident angle of the light was 15°, the distance between the light and the soil sample was 30 cm and the vertical distance between the probe and the soil surface was 10 cm (Fig. 1).

The experiment was conducted in laboratory at the night of May 12. In the experiment, the spectral range of the spectrometer was from 1000 nm to 2500 nm. However, the spectral data at each end was influenced seriously by the outside noise, so we removed the edge band, and reserved the data spectral range from 1350 nm to 2400 nm. Then we used spline interpolation to re-sample with a 5 nm interval. We took the average of three tests of each sample as the spec-

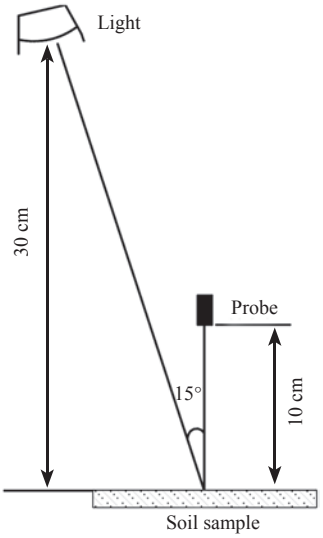

Fig. 1 Geometric conditions in laboratory test

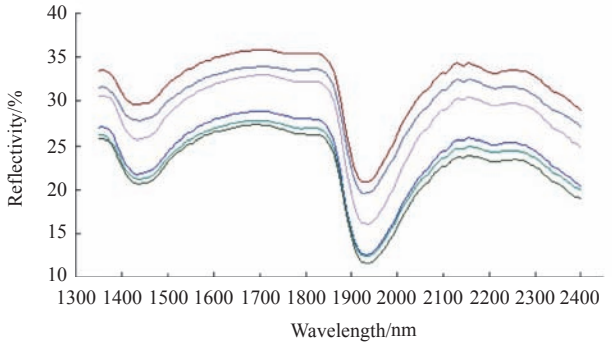

Fig. 2 The spectral reflection data of part of soil samples

tral reflection data for the follow-up analysis. Part of the spectral reflection data was shown in Fig. 2.

3 DATA PREPROCESSING METHODS

3.1 Envelope analysis

Spectral envelope curve is a broken line, which refers to connecting similar extreme value points using straight line point by point, whose peak point's exterior angle is larger than 180°. In this analysis, we acquired the envelope by way of connecting linear maximum point by point (Fig. 3). From the graph, it can be seen directly that the envelope was equivalent to the "shell" of the spectrum curve, and the 4-fragment characteristic of soil samples' spectral reflection curve was obvious.

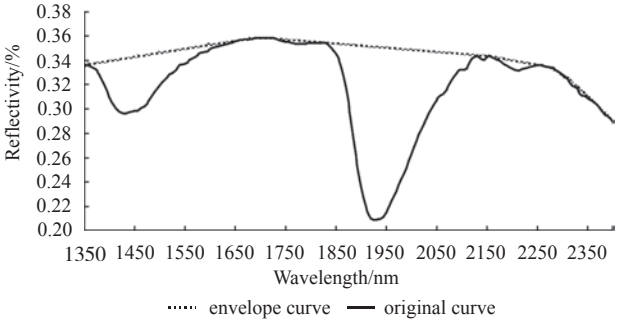

Fig. 3 Original spectrum curve and the envelope

Through the envelope analysis, we can find that the general trend of soil spectra curve can be composed by four line segments. 1350—1695 nm: This segment of curve is gently rising; 1695—2155 nm: This segment is gently falling; 2155—2280 nm: In this segment, the curve slope downward trend increases; 2280—2400 nm: This segment curve shows a sharp decline. The trend slope of the spectral segment is chosen to be the characteristic parameter to represent the soil curve shape feature. Its calculation can be determined as:

$$S = (R_i - R_j)/(\lambda_i - \lambda_j) \tag{1}$$

where S is slope, λ is wavelength, R is the corresponding reflective value and the subscript i and j respectively represents the position of wavelength. The segment curve slope in 1350—1695 nm, 1695—2155 nm, 2155—2280 nm and 2280—2400 nm, respectively corresponds to S_1, S_2, S_3 and S_4.

3.2 Standard normal variable transformation

Standard Normal Variable Transformation (SNV) is mainly used to eliminate the influence of the solid particle size, the surface scattering, the optical path's change to NIR diffuse reflective spectrum (Barnes, et al., 1989). SNV algorithm is aimed at processing a spectral reflection curve, which is based on spectral array of line. Its transformation and calculation can be determined as:

$$R_{i,\ \mathrm{SNV}} = \frac{R_{i,k} - \overline{R}_i}{\sqrt{\dfrac{\sum_{k=1}^{m}(R_{i,k} - \overline{R}_i)}{(m-1)}}} \tag{2}$$

where $\overline{R}_i$ is the average spectral reflection coefficient, k =1, 2, ⋯ , m, m is the band number and m=211, i =1, 2, ⋯ , n, n is the sample number of correction set and n=40, $R_{i,k}$ is the spectral reflection coefficient of the k^{th} band of i^{th} sample.

3.3 Spectrum differential treatment

Researches show that spectrum low-order (first and second) differential treatment has less sensitivity to the noise effect (Goodin, et al., 1993), so it can be used to eliminate parts of background and noise effects in the measurement environment and is helpful to the objective analysis of spectral reflection curve's shape variation. The positive and negative value of the one-order differential spectrum curve respectively represents the reflectance spectra curve rise or fall trend and the maximum and minimum values represented the position where the reflectance spectra curve changed most prominently. When calculate the one-order differential of reflectance spectra, we carry the logarithmic operations firstly on the spectral reflectance data in order to guarantee the normality of the first-order differential spectral data's regression residuals. Its calculation can be determined as:

$$\begin{aligned} R'(\lambda_i) &= \{\log[R(\lambda_{i+1})] - \log[R(\lambda_{i-1})]\}/(2\Delta\lambda) \\ R''(\lambda_i) &= [R'(\lambda_{i+1}) - R'(\lambda_{i-1})]/(2\Delta\lambda) \end{aligned} \tag{3}$$

In Eq. (3), $R'(\lambda_i)$ is the reflectivity logarithm's one-order differential

where the wavelength is λ_i, $R(\lambda_i)$ is the spectral reflectance where the wavelength is λ_i, $\Delta\lambda=\lambda_{i+1}-\lambda_{i-1}=5$ nm, λ_i=1350, 1355, 1360, … , 2400 nm.

4 ANALYSIS AND DISCUSSION

4.1 Correlation analysis

We conducted the correlation analysis of oil content in soil, respectively with line segment slope, standard normal variable transform data, logarithmic first-order differential data and logarithmic second-order differential data respectively (Table 3).

Table 3 Maximum correlation coefficient between spectrum characteristic parameters and oil content

Variables	Correlation coefficient
S_3	–0.74
$R_{SNV}(1735)$	–0.60
$R'(\lambda_{1775})$	0.62
$R''(\lambda_{2280})$	0.59

The result shows that the correlation coefficient is largest between the oil content and slope S_3 where the wavelength is 2155—2280 nm. The largest value is –0.74. In contrast, the correlation is weak between other segment slopes and the oil content; Standard normal variable transform data and oil content mainly presents a negative correlation and the largest correlation coefficient is –0.60; The correlation between the logarithmic first-order differential spectrum and the oil content dramatically changes (Fig. 4). The main reason is that the differential treatment is very sensitive to the high frequency noise and it will reduce the correlation between the differential spectrum and the oil content, though it can purify the influence on the target spectrum caused by part linear or close linear background and noise (Hu, et al., 2009).

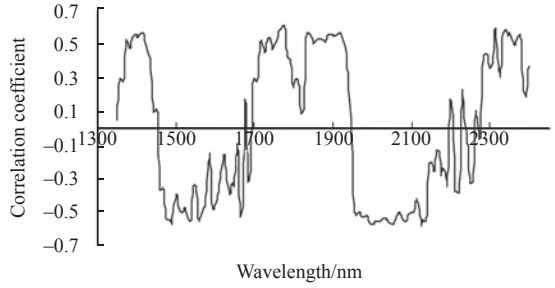

Fig. 4 The correlation figure between logarithmic first-order differential spectrum and oil content

4.2 Simple regression analysis

We conducted regression analysis of the 4 characteristic parameters which had the highest correlation with the content of the oil, using linear, conic and cubic curve models (Table 4). The fitting effect of model was examined by coefficient of determination R^2.

Table 4 Single variable regression analysis result table of 4 characteristic parameters with maximum correlation coefficient and oil content

Variable	Linear fitting R^2	Best nonlinear fitting R^2
S_3	0.548	0.632
$R_{SNV}(1735)$	0.357	0.357
$R'(\lambda_{1775})$	0.379	0.416
$R''(\lambda_{2280})$	0.352	0.446

It turns out that R^2 of characteristic parameter S_3 cubic curve function model whose slope is between 2155 nm and 2280 nm, has a maximum of 0.632, which can predict the content of the oil reasonably (Fig. 5). The regressions of other parameters are undesirable.

Quadratic curve model:

$$y=A+bx+cx^2 \tag{4}$$

Cubic curve model:

$$y=A+bx+cx^2+dx^3 \tag{5}$$

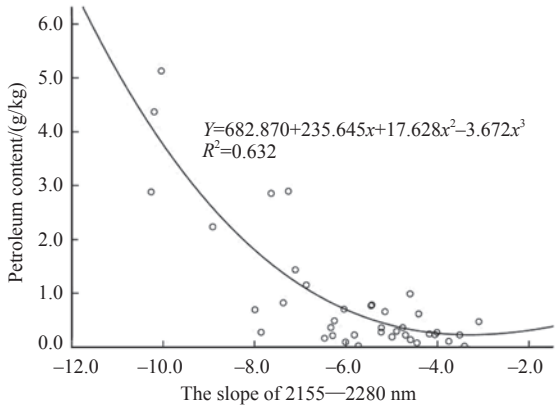

Fig. 5 The correlation figure between the slope S_3 of 2155—2280 nm and oil content

4.3 Multivariate regression analysis

Using the oil content in the soil as the dependent variable, and using various spectral absorption parameters as independent variable, we carried out multiple linear stepwise regression analysis. In the analysis, we took the F significant level as the criteria of stepwise regression analysis and set the probable values of entry and removal values of independent variable respectively as (0.05 and 0.10). We divided the 40 samples into two groups randomly: A group of 28 soil samples were used to establish regression forecasting model, and another group of 12 soil samples were used as inspection samples to verify the established model. The logistic model can be tested in two aspects, goodness of fit and predictive

ability. The model of goodness-of-fit was tested by the size of the adjusted coefficient of determination $\overline{R}^2$ (from modeling samples). When $\overline{R}^2$ is closer to 1, the model's degree of fitting is higher. The model prediction ability can be tested by the size of total Root Mean Square Error (RMSE, from modelling and inspection samples). The selected variables and evaluation of above four feature parameters through stepwise regression analysis were shown in Table 5.

Table 5 Results of oil content's multiple stepwise regression analysis

Model	Selected feature variables	Model ($\overline{R}^2$)	RMSE
S	S_3, S_1	0.639	0.766
R_{SNV}	$R_{SNV}(1735)$, $R_{SNV}(2335)$, $R_{SNV}(2230)$, $R_{SNV}(1615)$, $R_{SNV}(1400)$	0.826	0.531
R'	$R'(\lambda_{1775})$, $R'(\lambda_{2250})$, $R'(\lambda_{2270})$, $R'(\lambda_{2065})$, $R'(\lambda_{1640})$, $R'(\lambda_{1665})$, $R'(\lambda_{2370})$	0.819	0.525
R''	$R''(\lambda_{2280})$, $R''(\lambda_{1420})$, $R''(\lambda_{2140})$, $R''(\lambda_{1485})$	0.530	0.874

From Table 5, the accuracy of multiple regression model is superior to the single factor of regression analysis model. The slope of the spectral segment due to the characteristic variables involved in modelling section is less, and except for the slope of third segment which has a good correlation with oil content. The correlations of other three variables are not significant, indicates the multiple regression model is not ideal. After the second-order differential treatment of logarithm, though it deducts the original spectrum's background spectrum and raises the relativity of some spectral bands and oil content, the noise level increases and the signal to noise ratio decreases, which causes the worst regression effect. The regression effect about Standard Normal Variable Transformation and one-order differential regression of logarithm is the best. The first selected bands of model we chose are 1735 nm and 1775 nm respectively, and they have maximum correlation with oil content, respectively (Table 3). These two bands are located in the first oil hydrocarbon's characteristics absorption interval range of 1690—1790 nm, which confirms that the interval is diagnostic characteristic bands of oil hydrocarbons (Hu, et al., 2009). Although regression effects of these two models are quite close, logarithm's first-order differential needs to introduce 7 variables under the same precision basis while the Standard Normal Variable Transformation only needs 5 variables. Based on the principle of high stability, good prediction ability and less selected band, we take the multiple regression model of Standard Normal Variable Transformation as the optimal model. Model formula is as follows.

$$\begin{aligned} Y_{\text{petroleum}} = &-8.627 - 41.046 \times R_{SNV}(1735) - \\ &27.962 \times R_{SNV}(2335) + 26.677 \times \\ &R_{SNV}(2230) + 38.540 \times R_{SNV}(1615) - \\ &9.520 \times R_{SNV}(1400) \end{aligned} \quad (6)$$

The modelling samples and inspection samples of this model are concentrated near the line $y = x$. The adjusted coefficient of determination $\overline{R}^2$ of modeling samples is 0.826, and the adjusted coefficient of determination $\overline{R}^2$ of inspection samples is 0.848 (Fig. 6).

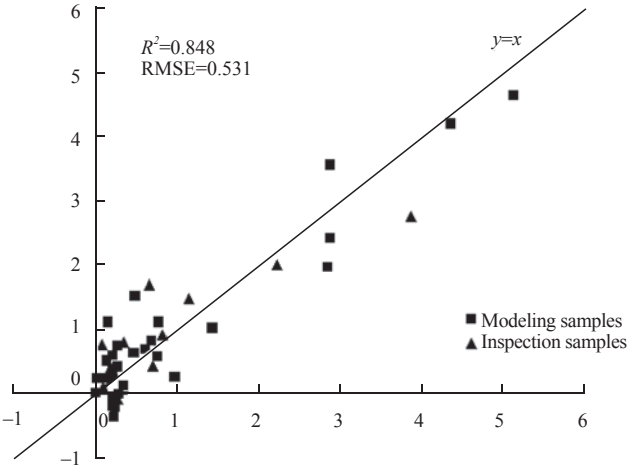

Fig. 6 The scatter diagram of the predictive value and the actual value of the Standard Normal Variable transformed multiple regression models

5 CONCLUSIONS

We analyze the relation between oil content in soil and spectral characteristic parameters quantitatively through indoor spectral experiments. The results show the potential to relate high spectral prediction model with soil oil content. Major points are:

(1) The band range of 1690—1790 nm is main spectral response area to oil content in soil. The wave band with maximum correlation of oil content and spectral logarithm's first-order differential is 1775 nm, and the largest correlation coefficient is –0.60;

(2) Through simple regressing analysis by linear model, quadratic curve model and cubic curve model, it turns out that characteristic parameter S_3 which slope is between 2155 nm and 2280 nm, has a maximum R^2 of 0.632;

(3) The coefficient of determination $\overline{R}^2$ is adjusted by standard normal variable transform modelling samples, the coefficient of determination $\overline{R}^2$ adjusted by inspection samples is 0.848 (RMSE=0.531), with fewer independent variables, which is indi-

cated as the best model.

We provide a new effective idea about the detection of the soil polluted by oil using hyperspectral data.

REFERENCES

Barnes R J, Dhanoa M S and Lister S J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy, 43(5): 772–777

Chen J Z, Chen J, Xie X J and Zhang X L. 2003. Soil pollution and its environmental impact. Soils, 35(4): 298–303

Feng X L and Shi Y G. 2002. Near infrared spectroscopy and its application in the analysis of petroleum. Beijing: China Petrochemical Press, 21–25

Goodin D G, Han L, Fraser R N, Rundquist D C, Stebbins W A and Schalles J F. 1993. Analysis of suspended solids in water using remotely sensed high resolution derivative spectra. Photogrammetric Engineering and Remote Sensing, 59(4): 505–510

He Q C. 2006. Geological Environment and Sustainable Development of the Yellow River Delta. Beijing: Geological Publishing House, 10–15

Hu P, Tian Q J and Yan B K. 2009. Estimating acidolysis hydrocarbon content in surface of Qaidam basin through hyperspectral remote sensing. Natural Gas Industry, 29(2): 32–36

Hui Y M, Zheng M H, Liu Z T and Gao L R. 2009. Distribution of polycyclic aromatic hydrocarbons in sediments from Yellow River Estuary and Yangtze River Estuary, China. Journal of Environmental Sciences, 21(12): 1625–1631

Liu Q S, Liu G H and Li H G. 2003. Comparison of the statistical distribution of oil content in soil in the Shengtuo oil field with that in the Gudong oil field. Chinese Journal of Soil Science, 34(6): 592–593

Sheng Z, Xie S Q and Pan C Y. 2005. Probability Theory and Mathematical Statistics. Beijing: Higher Education Press, 41–45

Xu D P and Wang B. 2006. Effect of petroleum hydrocarbon in soil on growth of sorghum and corn. Express Information of Mining Industry, 25(12): 28–30

Zhang X J, Ji W, Kang Z J, Sun D Y, Shan W and Wang J. 2008. Hazard analysis of petroleum pollutants in soil. Science and Technology in Chemical Industry, 16(6): 60–65

Zhou Q. 2004. Study on the Hyperspectral Prediction Models of Soil Organic Matter and Its Differences among Soil Types. Hangzhou: Zhejiang University, 28–29

Zhu Z H. 1994. Hydrocarbon microseepage theory and oil-gas reservoir detecting by remote sensing. Remote Sensing Technology and Application, 9(1): 1–10

孤东油田土壤石油类含量的高光谱反演模型

樊彦国，张磊

中国石油大学(华东)地球科学与技术学院，山东 青岛 266555

摘 要：土壤中石油类含量的检测对石油污染的预防与治理具有重要的实际意义。本文首先进行孤东油田土壤样品高光谱反射率的室内测定及石油含量的检测，然后利用单变量预测模型和逐步回归方法分析了土壤光谱特征参数与石油类含量之间的线性和非线性关系，结果表明：包络线分析的第三折线段斜率与石油类含量相关性最好，该段斜率的三次曲线函数为石油类含量的最佳单变量估算模型。标准正态变量变换对光谱的预处理效果最好，利用变换后光谱建立多元模型，其调整的判定系数$\overline{R}^2$是0.826，总均方根RMSE是0.531，且自变量个数较少，为最优预测模型。本文提出的利用高光谱数据检测土壤中石油类含量的方法，为土壤石油类污染检测提供了一种有效的新思路。

关键词：孤东油田，土壤，石油类，高光谱

中图分类号：TP79 **文献标志码**：A

引用格式：樊彦国，张磊. 2012. 孤东油田土壤石油类含量的高光谱反演模型. 遥感学报，16(2): 378–389
Fan Y G and Zhang L. 2012. Soil oil content hyperspectral model in Gudong Oilfield. Journal of Remote Sensing, 16(2): 378–389

1 引 言

坐落于黄河三角洲的胜利油田是中国第二大油田，随着油田的开发开采致使黄河三角洲受石油污染土壤面积占总面积的24%，落地原油是最大的污染物，区域内石油类质量分数在9.2—180.9 mg/kg之间(何庆成，2006)；沉积物中多环芳烃的质量分数为0.0108—0.252 mg/kg(Hui 等，2009)。石油排入土壤后，能破坏土壤结构，影响土壤的通透性，改变土壤有机质的组成和结构，降低土壤质量(张学佳 等，2008)。当土壤中的石油类含量较高时，会对植物的生长产生抑制作用(许端平 等，2006)。土壤的严重污染会导致石油烃的某些成分在粮食中积累，影响粮食的品质，并通过食物链，危害人类健康(陈晶中 等，2003)。为了尽量减小和降低土壤石油类污染对生态环境的破坏和对人类健康的危害，需要及时准确地对土壤石油类污染进行检测和评估。

土壤被石油污染后会造成土壤烃组分异常，即增加了石油类烃的含量。朱振海(1994)在油气藏烃类微渗漏理论的介绍中指出石油类烃光谱的4个特征吸收带：1.69—1.97 μm、2.27—2.46 μm、3.33—3.53 μm和6.80—7.48 μm。在1748 nm附近的双吸收峰特征以及2330 nm附近的双吸收峰特征是判断土壤中是否含有石油烃类物质的诊断特征(冯新泸 等，2002)。传统的石油污染评价，需要大量土样的野外采集与样本中石油类含量的化验，然后根据化验结果进行土壤石油类污染的评价和制图，这样做即费时、费力，又要较高的成本。

本文以孤东油田为实验区，通过对实验区土壤样本的室内光谱测定和分析，利用数理统计的方法，以土壤反射光谱数据的分段斜率、标准正态变量变换和低阶微分为自变量，以石油类含量为因变量，确定石油类含量与各光谱特征参数的对应关系，进行从高光谱数据直接获取土壤石油类含量的方法探索和可行性分析。

收稿日期：2011-01-20；**修订日期**：2011-05-20

基金项目：中央高校基本科研业务费专项资金；山东省科技计划项目(编号: J08LD55)

第一作者简介：樊彦国(1965—)，男，博士，副教授，从事3S技术在数字国土、城市及海岸带方向的教学与研究工作。E-mail: fanyanguo2003@163.com。

2 研究区概况与数据准备

2.1 研究区概况

研究区孤东油田位于山东省东营市垦利县境内，地处黄河入海口北侧的滩涂地带，是一个集油气勘探、开发、处理和集输为一体，年产原油260 万t的采油厂。根据刘庆生等人(2003)2001年的实地考察分析，孤东油田土壤石油类含量背景值为61.67 mg/kg，土壤石油类含量平均值为236.40 mg/kg，与黄河三角洲区域总体情况相比，其土壤中石油类物质含量相对较高，土壤石油类污染相对严重。有些样点的石油类含量大于500 mg/kg，从而证明孤东油田的部分区域已经被严重污染。

2.2 土壤样本采集

一般样本的采集应均匀地分布于整个研究区域。课题组参考Google Earth及孤东油田示意图制定采样路线，并于2010-05-03对孤东油田进行实地考察采样，在采样过程中，考虑到油井布置、石油泄漏及交通等客观因素，对拟定的采样路线进行了修改，最终在研究区共采集42个土壤表层样本。将土壤样本于室内自然风干，敲碎干结土块过2 mm筛后，东营市农产品质量监督检测中心进行土壤中石油类含量化验。在42个土样中有两个样品点(GDQ-9和GDQ-20)出现异常值，浓度远远高于其他土壤样点，通过分析发现，GDQ-9和GDQ-20号土样中本身含有黑色固体油污。因此，除去这两个异常样本点，将剩余的40组土壤样本数据作为实验数据，其原始记录表以及描述性统计量见表1和表2。

2.3 光谱数据采集

对40个土壤样本进行光谱测试。本次实验采用荷兰Avantes公司生产的AvaField3型地物光谱仪进行土壤样本的光谱采集工作。该仪器光谱范围1000—2500 nm，光谱分辨率10 nm，光谱采用间隔6 nm。试验用光源为1000 W卤光灯。研究表明，在进行室内高光谱实验时，光源的入射角度、光源距被测土样表面的距离、探头的角度以及到被测土样表面的距离等都会对实验数据产生影响。根据周清(2004)就室内几何测试条件对高光谱质量影响的研究结果，及采集仪器的物理参数，本次室内试验几何测试条件设计为：光源入射角15°、光源距离土样表面30 cm、探头垂直于土壤表面距离10 cm(图1)。

实验选择在2010-05-12的夜晚在室内完成，尽量缩短与石油类化验的时间差。试验用光谱仪的光谱范围是1000—2500 nm，但是两端光谱数据受外界噪声影响很大，所以去除边缘波段保留1350—2400 nm数据后，将数据采用样条插值的方法，重采样至5 nm等间隔。同一样本的3次试验数据取平均后作为后续分析的光谱反射数据。部分光谱反射数据见图2。

表1 土壤样本石油类含量检测原始记录表以及描述性统计量

编号	石油类含量/(mg/kg)	编号	石油类含量/(mg/kg)	编号	石油类含量/(mg/kg)
GDQ–1	694.3	GDQ–15	363.1	GDQ–29	659.5
GDQ–2	295.2	GDQ–16	276.7	GDQ–30	2894.7
GDQ–3	166.6	GDQ–17	704.1	GDQ–31	1154.6
GDQ–4	2882.5	GDQ–18	359.4	GDQ–32	769.5
GDQ–5	225.2	GDQ–19	489.9	GDQ–33	232.0
GDQ–6	472.9	GDQ–20	17434.8	GDQ–34	226.7
GDQ–7	214.8	GDQ–21	191.6	GDQ–35	2854.8
GDQ–8	990.0	GDQ–22	12.7	GDQ–36	2233.8
GDQ–9	20622.0	GDQ–23	78.9	GDQ–37	127.7
GDQ–10	273.7	GDQ–24	24.5	GDQ–38	226.6
GDQ–11	279.6	GDQ–25	247.3	GDQ–39	3882.4
GDQ–12	618.1	GDQ–26	5129.2	GDQ–40	785.7
GDQ–13	94.2	GDQ–27	364.4	GDQ–41	4370.1
GDQ–14	139.0	GDQ–28	1437.0	GDQ–42	824.2

表2 土壤样本石油类含量描述性统计量表

样本个数	最小值/(g/kg)	最大值/(g/kg)	平均值/(g/kg)	标准差
40	0.0127	5.1292	0.9567	1.2747

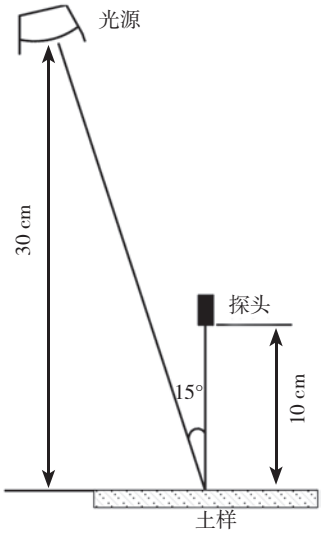

图1 室内几何测试条件

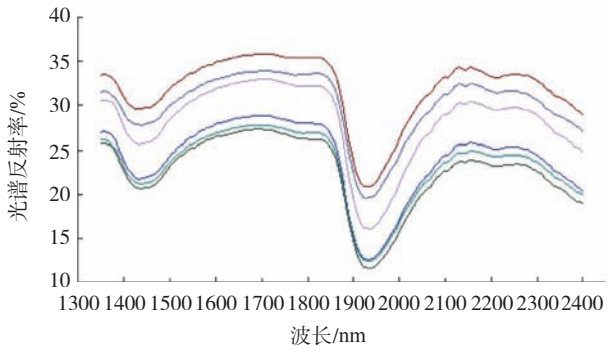

图2 部分土样光谱反射数据

3 数据的预处理方法

3.1 包络线分析

光谱曲线的包络线是指逐点直线连接同类极值点，并使折线在峰值点上的外角大于180° 所构成的折线，本次分析采用的是逐点直线连接极大值点所构成包络线(图3)。

由图3可示，土壤光谱曲线的变化趋势可分为4段来表示：1350—1695 nm，此段曲线斜率平缓上升；1695—2155 nm，此段曲线斜率平缓下降；2155—2280 nm，此段曲线斜率下降趋势有所增加；2280—2400 nm，此段曲线急剧下降。本次分析选择光谱段的趋势斜率作为特征参数来表征土壤曲线的形状特征，计算方法如下：

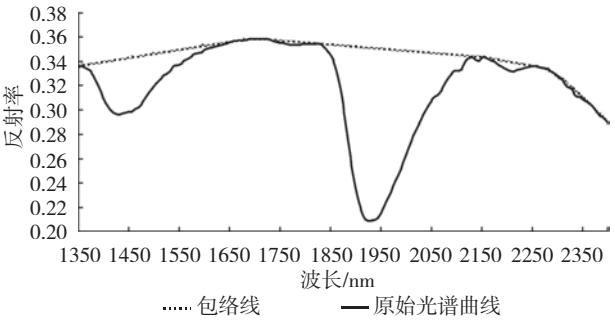

图3 原始光谱曲线与包络线

$$S=(R_i-R_j)/(\lambda_i-\lambda_j) \tag{1}$$

式中，S为斜率，λ为波长，R为相应的反射率值，下标i、j分别表示起始波长位置。S_1、S_2、S_3、S_4分别表示1350—1695 nm、1695—2155 nm、2155—2280 nm和2280—2400 nm段斜率。

3.2 标准正态变量变换

标准正态变量变换SNV(Standard Normal Variate Transformation)主要是用来消除固体颗粒大小、表面散射以及光程变化对NIR漫反射光谱的影响(Barnes 等，1989)。SNV算法是针对一条光谱反射曲线进行处理，即基于光谱阵的行，其变换计算方法为：

$$R_{i,\ \mathrm{SNV}}=\frac{R_{i,k}-\overline{R}_i}{\sqrt{\dfrac{\sum_{k=1}^{m}(R_{i,k}-\overline{R}_i)}{(m-1)}}} \tag{2}$$

式中，$\overline{R}_i$为第i个样本的平均光谱反射系数；k=1，2，…，m，m为波段个数，本实验m=211；i=1，2，…，n，n为校正集样本数，本实验n=40；$R_{i,k}$为第i个样品的第k个波段的光谱反射系数。

3.3 光谱微分处理

研究表明，光谱的低阶(一阶、二阶)微分处理对噪声影响的敏感性较低(Goodin 等，1993)，可以消除部分测量环境中的背景及噪声影响，有助于客观分析光谱反射曲线的形态变化。光谱曲线一阶微分的正负值分别表示反射率光谱曲线的上升或下降趋势，极大值与极小值代表反射率光谱曲线变化最显著的位置。计算反射率光谱一阶微分时，为了保证一阶微分光谱数据回归残差的正态性，先对反射率光谱数据进行对数运算，计算方法如下：

$$R'(\lambda_i)=\{\log[R(\lambda_{i+1})]-\log[R(\lambda_{i-1})]\}/(2\Delta\lambda)$$
$$R''(\lambda_i)=[R'(\lambda_{i+1})-R'(\lambda_{i-1})]/(2\Delta\lambda) \quad (3)$$

式中，$R'(\lambda_i)$为反射率对数在波长λ_i处的一阶微分，λ_i为波长，$R(\lambda_i)$为波长λ_i处的光谱反射率；$\Delta\lambda=\lambda_{i+1}-\lambda_{i-1}=5$ nm，λ_i=1350，1355，1360，…，2400 nm。

4 分析与讨论

4.1 相关性分析

将折线段斜率、标准正态变量变换、对数的一阶微分数据、对数的二阶微分数据分别与土壤中石油类含量做相关性分析，其中最大相关系数见表3。

表3 光谱特征参数与石油类含量的最大相关系数表

变量	相关系数
S_3	–0.74
$R_{SNV}(1735)$	–0.60
$R'(\lambda_{1775})$	0.62
$R''(\lambda_{2280})$	0.59

由数据结果表所示，2155—2280 nm段斜率S_3较其他变量与石油类含量相关系数最大，达到–0.74；标准正态变量变换数据与石油类含量主要成负相关，其中最大相关系数是–0.60；对数的一阶微分光谱与石油类含量的相关性呈现出剧烈的变化(图4)，这种相关关系急剧变化的主要原因是，微分处理对高频噪声十分敏感，这会降低微分光谱与石油类含量的相关性(胡畔 等，2009)。

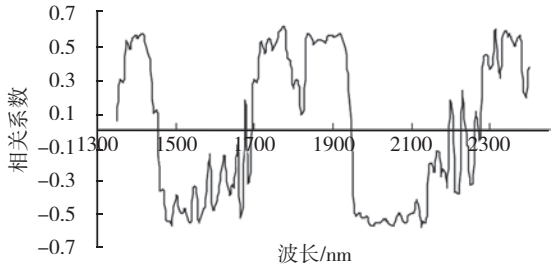

图4 对数的一阶微分光谱与石油类含量的相关关系图

4.2 单变量回归分析

将4个与石油类含量相关性最高的特征参数，逐个用线性、二次曲线、三次曲线等模型进行回归分析(表4)。用判定系数R^2检验模型的拟合效果。判定系数等于回归平方和ESS在总体平方和TSS所占的比率，即回归方程所能解释的因变量变异性的百分比，其取值范围是[0，1]，越接近1说明实际观测点离样本线越近，拟合优度越高。

表4 4个相关系数最大的特征参数与石油类含量的单变量回归分析结果

变量	线性拟合R^2	最佳非线性拟合R^2
S_3	0.548	0.632
$R_{SNV}(1735)$	0.357	0.357
$R'(\lambda_{1775})$	0.379	0.416
$R''(\lambda_{2280})$	0.352	0.446

回归分析结果发现，2155—2280 nm段斜率特征参数S_3的三次曲线函数模型的R^2最大为0.632，可以较合理地预测出石油类含量(图5)。其他参数回归效果并不理想。

二次曲线回归基本模型：

$$y=A+bx+cx^2 \quad (4)$$

三次曲线回归基本模型：

$$y=A+bx+cx^2+dx^3 \quad (5)$$

具体的计算方法请参考《概率论与数理统计》(盛骤 等，2005)，在本文中回归分析采用SPSS软件辅助实现。

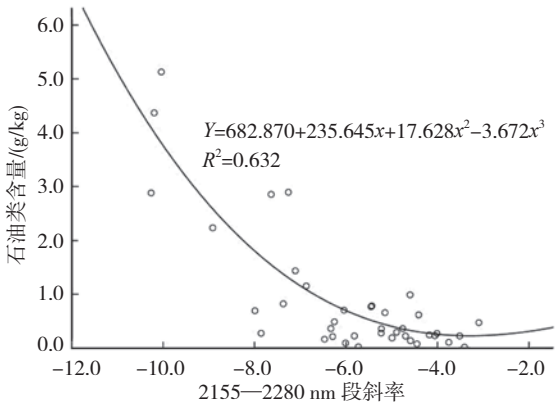

图5 2155—2280 nm 段斜率S_3与石油类含量的关系图

4.3 多元线性回归分析

对各种光谱特征参数—土壤石油类含量进行多元线性逐步回归分析。以F显著性水平作为逐步回归分析的准则，选入和剔除自变量的概率值分别设置为0.05和0.10。将所有40个样本随机分成两组：一组28个土壤样本用于建立回归预测模型，另一组12

个土壤样本作为检验样本用于验证建立的模型。从模型的拟合优度和预测能力两方面对模型进行验证，模型的拟合优度用调整的判定系数$\overline{R}^2$的大小检验(来自建模样本)，$\overline{R}^2$越接近1，模型的拟合度越高；模型的预报能力用总均方根差RMSE(来自建模与检验样本)的大小来检验，RMSE越小，模型的预测能力与精度越高。上述4种特征参数逐步回归分析入选变量和评定见表5。

由表5可知，多元回归模型较单因素回归分析模型精度要高。光谱段斜率由于参与建模的特征变量较少，并且除第三段斜率与石油类含量相关性较好外，其他3个变量相关性并不明显，所以多元回归模型并不理想；对数的二阶微分处理后，虽然扣除了原始光谱的背景光谱，提高了某些光谱波段与石油类含量的相关性，但是噪音水平增大，信噪比减小，致使回归效果最差；标准正态变量变换与对数的一阶微分的回归效果最好，模型第一个入选波段分别是1735 nm与1775 nm，与石油类含量分别具有最大相关性(表3)，这两个波段均位于首个石油类烃的特征吸收区间1690—1790 nm，说明该区间是石油类烃的诊断性特征波段，这与胡畔等人(2009)的研究结果一致，具体波段略有差异。虽然这两套模型回归效果十分接近，但在同等精度的基础上对数的一阶微分需要引入7个变量，而标准正态变量变换只需5个变量，基于稳定性高、预测能力强和入选波段少的原则，确定标准正态变量变换多元回归模型为最优模型。模型公式：

$$\begin{aligned} Y_{\text{石油类}} = &-8.627 - 41.046 \times R_{\text{SNV}}(1735) - \\ &27.962 \times R_{\text{SNV}}(2335) + 26.677 \times \\ &R_{\text{SNV}}(2230) + 38.540 \times R_{\text{SNV}}(1615) - \\ &9.520 \times R_{\text{SNV}}(1400) \end{aligned} \quad (6)$$

该模型建模样本与检验样本均集中在直线$y=x$附近，建模样本调整的判定系数$\overline{R}^2$为0.826，检验样本调整的判定系数$\overline{R}^2$为0.848 (图6)。

表5 石油类含量多元逐步回归分析结果

模型	入选特征变量	模型 ($\overline{R}^2$)	总均方根差 (RMSE)
S	S_3，S_1	0.639	0.766
R_{SNV}	$R_{\text{SNV}}(1735)$，$R_{\text{SNV}}(2335)$，$R_{\text{SNV}}(2230)$，$R_{\text{SNV}}(1615)$，$R_{\text{SNV}}(1400)$	0.826	0.531
R'	$R'(\lambda_{1775})$，$R'(\lambda_{2250})$，$R'(\lambda_{2270})$，$R'(\lambda_{2065})$，$R'(\lambda_{1640})$，$R'(\lambda_{1665})$，$R'(\lambda_{2370})$	0.819	0.525
R''	$R''(\lambda_{2280})$，$R''(\lambda_{1420})$，$R''(\lambda_{2140})$，$R''(\lambda_{1485})$	0.530	0.874

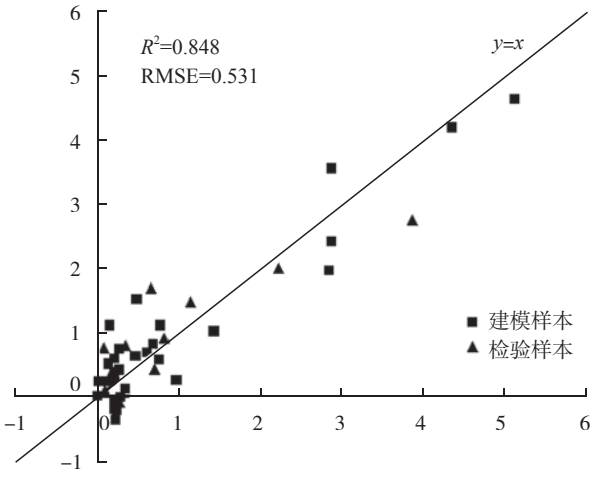

图6 标准正态变量变换多元回归模型预测值与测量值散点图

5 结 论

通过室内光谱实验定量分析了土壤石油类含量与光谱特征参数间的关系，并建立了土壤石油类含量高光谱预测模型。结果表明：

(1)1690—1790 nm波段范围是土壤中石油类的主要光谱响应区域，石油类含量与光谱对数的一阶微分最大相关性波段是1775 nm，相关系数为0.62，与光谱标准正态变量变换最大相关性波段是1735 nm，相关系数为－0.60；

(2)通过单变量的线性、二次曲线、三次曲线等模型回归分析，2155—2280 nm段斜率特征参数S_3的三次曲线函数模型的判定系数R^2最大为0.632，可以较合理地预测出石油类含量；

(3)标准正态变量变换模型建模样本调整的判定系数$\overline{R}^2$为0.826，检验样本调整的判定系数$\overline{R}^2$为0.848，RMSE为0.531，且自变量较少，确定为最优模型。

本文提出的利用高光谱数据检测土壤中石油类含量的方法，为土壤石油类污染检测提供了一种有效的新思路。

参考文献(References)

Barnes R J, Dhanoa M S and Lister S J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy, 43(5): 772–777

陈晶中, 陈杰, 谢学俭, 张学雷. 2003. 土壤污染及其环境效应. 土壤, 35(4): 298–303

冯新泸, 史永刚. 2002. 近红外光谱及其在石油产品分析中的应用. 北京: 中国石化出版社, 21–25

Goodin D G, Han L, Fraser R N, Rundquist D C, Stebbins W A and Schalles J F. 1993. Analysis of suspended solids in water using remotely sensed high resolution derivative spectra. Photogrammetric Engineering and Remote Sensing, 59(4): 505–510

何庆成. 2006. 黄河三角洲地质环境与可持续发展. 北京: 地质出版社, 10–15

胡畔, 田庆久, 闫柏琨. 2009. 柴达木盆地地表酸解烃含量的高光谱遥感估算. 天然气工业, 29(2): 32–36

Hui Y M, Zheng M H, Liu Z T and Gao L R. 2009. Distribution of polycyclic aromatic hydrocarbons in sediments from Yellow River Estuary and Yangtze River Estuary, China. Journal of Environmental Sciences, 21(12): 1625–1631

刘庆生, 刘高焕, 励惠国. 2003. 胜坨、孤东油田土壤石油类物质含量及其变化. 土壤通报, 34(6): 592–593

盛骤, 谢式千, 潘承毅. 2005. 概率论与数理统计. 北京: 高教出版社, 41–45

许端平, 王波. 2006. 土壤中石油烃类污染物对高粱玉米生长的影响研究. 矿业快报, 25(12): 28–30

张学佳, 纪巍, 康志军, 孙大勇, 单伟, 王建. 2008. 石油类污染物对土壤生态环境的危害. 化工科技, 16(6): 60–65

周清. 2004. 土壤有机质含量高光谱预测模型及其差异性研究. 杭州: 浙江大学, 28–29

朱振海. 1994. 油气藏烃类微渗漏理论及遥感油气资源探测的机理研究. 遥感技术与应用, 9(1): 1–10