SPA-PLS的高含水原油近红外光谱含水率分析

韩 建，李雨昭，曹志民*，刘 强，牟海维

东北石油大学电子科学学院，黑龙江 大庆 163318

摘 要 准确及时的检测原油含水率对注水策略调整、原油开采能力评估、油井开发寿命预测等均具有重要意义。然而，当前我国大多数油田均已进入高含水的开发中晚期，含水率测量难度大且准确率不高。在此背景下，开展了高含水情况下利用近红外光谱进行原油含水率测量的研究。首先介绍了目前原油含水率检测的常用方法，分析了它们的优劣。理论上，由于水的近红外光吸收带与原油中C—H键的吸收带有明显区别，根据Lambert-Beer吸收定律和吸光度线性叠加定律可知，不同含水率高含水原油近红外光谱会存在较强响应差异。为此，对高含水原油进行近红外光谱检测，建立原油含水率与近红外光谱响应间的非线性映射模型，可实现高含水原油含水率的精确测量。为了验证该方法的有效性，搭建了近红外光谱数据采集实验装置：采用白炽灯作为光源，经过光路调节成平行光后垂直射入样品池，用近红外光谱仪(海洋光学 NIR512)采集光谱用于分析。其中，接收光谱仪带宽为900～1 700 nm，平均分成512个波段。光谱数据利用光谱仪配套软件储存在电脑中。样本采用相同厚度不同比例的油水混合物，样本含水率范围为70%～99%，共采集数据60组，每组重复3次取平均值。得到原始数据后，先进行原始数据预处理，以减少数据采集时来自高频随机噪音及温度不稳定、样本不均匀、基线漂移、光散射等不利因素的影响。分别选用了S-G滤波、一阶导数和S-G滤波＋一阶导数作为数据预处理的方法，利用连续投影算法(SPA)对光谱数据进行降维，并利用偏最小二乘法(PLS)和多元线性回归(MLR)进行建模，模型精度通过计算均方根误差值(RMSE)和相关系数(r)来验证。对比发现，使用S-G滤波＋一阶导数建立的模型RMSE值最小(RMSE＝0.007 0，r＝0.998 3)。使用SPA降维后的模型要优于全波段PLS模型(RMSE＝0.083 3，r＝0.920 6)与MLR模型(RMSE＝0.099 9，r＝0.967 1)。利用SPA提取出的31个特征波长建立的模型仅占全波段的6.05%，并获得了较好的精度。证明了利用光谱检测高含水原油含水率可行性，并且得到了满意的精度，为高含水原油的含水率检测提供了新的方法，为进一步利用近红外光进行高含水原油的快速检测与在线监测提供参考。

关键词 近红外光谱；高含水率原油；连续投影算法；偏最小二乘法

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)11-3452-07

引 言

油井含水率是管理采油进度、调整相应作业模式所需的重要指标，准确实时的预测含水率也是降低水平井钻井成本的重要因素之一。目前，世界上许多主要油田都处于高含水阶段，综合含水量几乎都超过了90%甚至更高。然而，准确、及时的预测高含水原油含水率仍是一个具有挑战性的课题。传统的方法是用化学分析方法在实验室中测定含水率，这需要消耗大量的人力和时间。除了手动操作外，常用的利用传感器检测含水率的方法有电容、伽马射线、微波等[1]，但目前使用的很多传感器在高含水情况下都有一定局限性。如电容法由于极板的边缘效应，测量误差在3%左右[2]，伽马射线法测量含水率范围窄，且会产生辐射，微波法仪器复杂，尤其是在即时的传感约束下。

可见-近红外光谱(visible and near infrared spectroscopy，Vis-NIRS)可以充分利用全谱段或多波长的光谱数据对物质的品质、种类、化学成分等进行定性和定量分析，已广泛应用于农业、石油化工、食品、制药等领域，取得了可喜的成果[3-5]。近红外光谱区与有机分子(如有机碳氮源)中的含氢基团(C—H)振动的合频与各级倍频的吸收一致，可以得到样品中有机分子含氢基团的特征振动信息。与上述传感器相

收稿日期：2018-10-18，**修订日期**：2019-02-15

基金项目：国家自然科学基金项目(51574087)资助

作者简介：韩 建，1976年生，东北石油大学电子科学学院教授 e-mail：han-jian@126.com

*通讯联系人 e-mail：dahai0464@sina.com

比，近红外(NIR)光谱法因检测设备相对简单，能够即时反映出结果，且能得到较高的分辨率，在高含水原油检测中被证明是一种理想方法[6-8]。

理论上，通过扫描样品的可见近红外光谱，利用 H_2O 和 C—H 键对近红外光波的吸收差异，可实现测量原油含水率。但是利用近红外光谱检测高含水原油含水率的研究很少，有一些关于利用近红外光谱法检测高含水的油水混合物含水率的研究。如检测润滑油、汽轮机油的含水率和对污水中的油污染物的检测，这些研究为利用近红外光谱检测高含水原油含水率提供了参考[9-11]。然而，现有的基于近红外光谱的含水率预测方法仍然比较复杂，因为使用了完整的测量带，没有进行简化。每次计算都会耗费大量时间。因此，为了解决这一问题，我们提出了一种计算效率高的近红外光谱分析方法，利用连续投影算法(SPA)和偏最小二乘法(PLS)来预测原油含水率，简称 SPA-PLS。

1　原理与方法

1.1　Lambert-Beer 吸收定律

近红外光谱法的理论基础是基于 Lambert-Beer 吸收定律和吸光度线性叠加定律，如图 1 所示。当单色光通过油和水均匀溶液时，其透射光强可表示为

$$I = I_0 \mathrm{e}^{-(\alpha_1 L_1 + \alpha_2 L_2)} \tag{1}$$

其中：I 指透射光强，I_0 指入射光强，α_1 和 α_2 是油和水的吸收系数，L_1 和 L_2 是油和水的厚度。在式(1)中，α_1 和 α_2 为已知量，I 和 I_0 为可测量，可求得关于油和水厚度 L_1 和 L_2 的表达式，显然仅利用一束单色光是无法得到的，需利用第二束不同频率的单色光同时进行测试，并要求油和水对两种频率单色光的吸收系数不同，进而联立方程组求得 L_1 和 L_2，实现原油含水率测量。

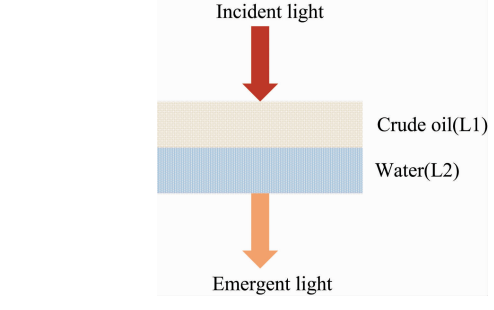


图 1　介质对光的吸收

Fig. 1　Absorption of light by medium

由于水和原油在在近红外段的吸收带不同(如表 1 所示)，因此可以利用近红外光谱对油水混合物进行检测。

1.2　偏最小二乘连续投影算法(PLS-SPA)提取光谱特征

全频带光谱中包含大量无关、冗杂的信息。为了提取特征波长，采用了连续投影算法(SPA)进行降维。连续投影算法是一种使矢量空间共线性最小化的前向变量选择算法，Bregman 于 1965 年首先提出[12]。它可以通过提取全波段的几个特征波长来够消除原始光谱矩阵中冗余的信息，可用于

表 1　C—H 键和水在近红外段的吸收带

Table 1　The near infrared absorption bands of C—H bond and water

C—H 吸收带/nm	H_2O 吸收带/nm
1 690～1 755	1 762～1 977
1 127～1 270	1 319～1 548
845～878	1 095～1 165
	926～978

光谱特征波长的筛选。近年来，国内外学者在利用光谱分析检测作物和食品中某些重要成分的含量时利用了连续投影算法作有效波长的选取[13-15]。具体算法如下：

设有光谱矩阵 $\boldsymbol{X}^{n\times p}$ 及样本性质参数矢量 $\boldsymbol{Y}$，其中，设样本容量为 n，光谱总波长为 p。利用 SPA 进行波长选择的算法步骤分为两个阶段：

阶段一：对光谱矩阵 $\boldsymbol{X}^{n\times p}$ 进行分组。共分成 p 组，集合设为 $\boldsymbol{sl}=[s_1, s_2, \cdots, s_m]\in \boldsymbol{R}^{p\times m}$。每组选择 m 个波长[$m\leqslant M=\min(n, p)$]。各波长矢量是通过下列步骤计算得出的：

第一步，令 $i=1$，$k=1, 2, \cdots, p$，$z_i=x_k$；$s_1^k=x_k$；$sl(k, 1)=k$；$u=1, 2, \cdots, m$；

第二步，基于 z_i 构造正交投影算子。其中 $\boldsymbol{I}$ 为 $n\times n$ 的单位矩阵；

$$\boldsymbol{P}_i = \boldsymbol{I} - \frac{z_i (z_i)^{\mathrm{T}}}{(z_i)^{\mathrm{T}} z_i} \tag{2}$$

将还未被选入的各波长矢量的位置集合记为 v，即 $v\in[1, p]\&\&v\notin sl$；$s_v^k = p_i x_v$。

第三步，计算各 s_v^k 的正交投影矢量，并从中选出波长位置，即

$$sl(k, u) = \underset{v\in[1, p]\&\&v\notin sl}{\operatorname{argmax}} \| s_v^k \|；z_i = x_{sl(u)} \tag{3}$$

第四步，令 $i=i+1$，若 $i<m$，返回至第 2 步开始选择下一波长矢量。重复上述步骤得到降后的维光谱矩阵 $\boldsymbol{sl}=[s_1, s_2, \cdots, s_m]\in \boldsymbol{R}^{p\times m}$。

阶段二：利用多元定量校正模型完成最优波长的选定。在此，选用偏最小二乘法(PLS)建立 NIR 光强度与含水量之间的相应显式关系如式(4)所示

$$\boldsymbol{Y} = a_1^j \boldsymbol{X}_{sl(j,1)} + a_2^j \boldsymbol{X}_{sl(j,2)} + \cdots + a_m^j \boldsymbol{X}_{sl(j,m)} + \varepsilon_j \tag{4}$$

偏最小二乘法(PLS)广泛应用于近红外光谱分析。根据以上步骤所得出的选择结果，建立原油含水率光谱数据预测模型。选用均方根误差(RMSE)和相关系数(r)作为模型精度的评价指标

$$\mathrm{RMSE}_j = \sqrt{\frac{\sum \boldsymbol{Y}_{\mathrm{test}}^{\mathrm{truth}} - \boldsymbol{Y}_{\mathrm{est}}^{N}}{N_{\mathrm{test}}}} \tag{5}$$

$$r = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\delta_x}\ \sqrt{\delta_y}} \tag{6}$$

$$\mathrm{Cov}(X, Y) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y}) \tag{7}$$

模型的预测均方根误差越小，相关系数越接近 1，模型的精度则越高。本实验以其均方根误差和相关系数为目标优化所建模型，则最小均方根误差且相关系数最大所对应的变量位置和个数就是最优波长。

2 实验部分

实验装置示意图和实物图如图 2 所示。

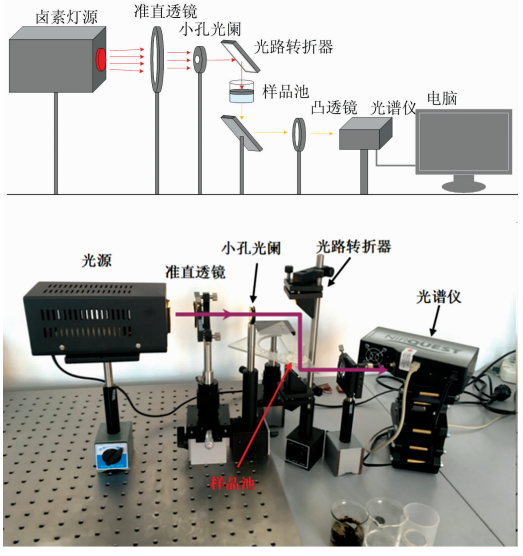


图 2 实验装置

Fig. 2 Experimental device

光源发出的光通过准直透镜和小孔光栅形成平行光，利用光路转折器将光透过样品池，再对透射光进行聚焦后由光谱仪采集。调配含水率从 99%～70%的含水原油作为测试样本，样本含水率以 0.5%递减，共获取 60 组，实验重复 3 次，取平均值作为该含水率样本的光谱值。

光谱采集使用海洋光学 NIR512 近红外光谱分析仪，波长范围为 850～1 700 nm，光学分辨率为 3.1 nm w/25 mslit (共 512 个波段)。14.5 V 卤素灯作为外部光源。随机选取 48 组油样为实验数据，随机选取 12 组作为验证数据。MATLAB R2016a 软件处理光谱数据。

3 结果与讨论

3.1 光谱预处理

由于原始光谱采集时会受到温度及来自高频随机噪音、样本不均匀、基线漂移、光散射等不利因素的影响，实验设计中，为尽量减少温度和液体表面张力对样本带来的影响，采用了 25 mm 大面积样品池。油品膨胀系数为0.000 528，水的膨胀系数为0.000 208。根据液体膨胀体积[式(8)]可以求得在室温 20～35 ℃范围，水和油的体积变化率分别为 0.312%和 0.792%，实验温度恒定情况下影响很小。

$$V_2 = V_1[1+\alpha(t_2-t_1)] \tag{8}$$

式(8)中，t_1 和 t_2 为温度，V_1 为在 t_1℃时的体积，V_2 为在 t_2℃时的体积。α 为膨胀系数。

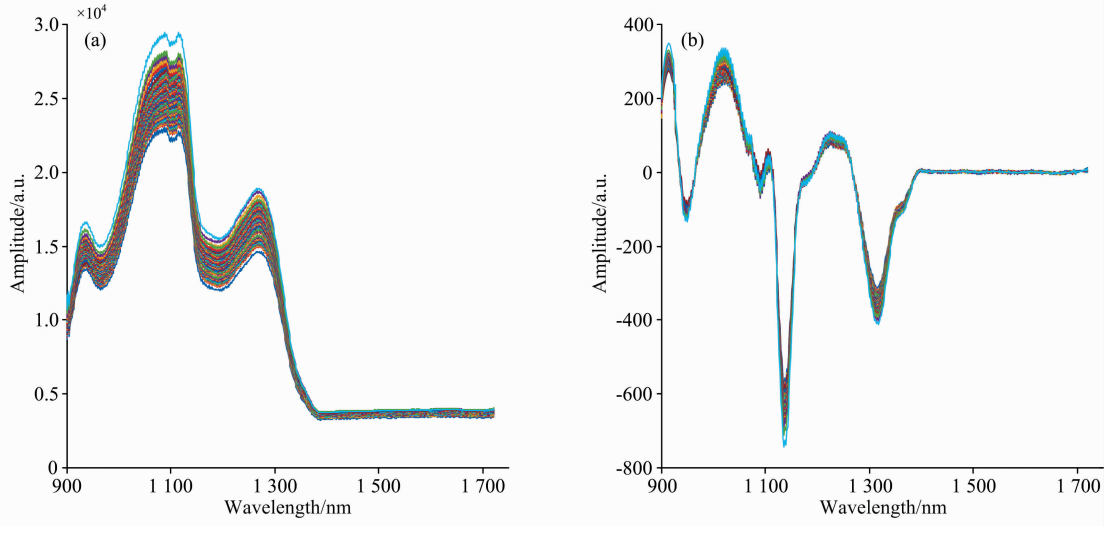
需采取一定的预处理方法以消除来自高频随机噪音、样本不均匀、基线漂移、光散射等不利因素的影响，实现最优的建模效果。

48 组含水原油样本的近红外吸收光谱如图 3(a)所示。可以看出，在 980，1 100 和 1 200 nm 附近有明显的吸收。由于样本在样品池中会受到凹液面折射等影响，不利于特征波长的选取，因此，对原始数据进行了一阶导数处理。

图 3(b)为原始光谱的一阶导数光谱，可以看到，一阶导数光谱突出了光谱的吸收特征，并且消除了原始光谱中的基线漂移和部分背景噪声，1 050，1 150 和 1 330 nm 等吸收峰特征较为明显，部分波长处油中含水量光谱显示出较为明显的差别。

一阶导数处理后的光谱依然有噪声存在，因此选用了 Savitzky-Golay 滤波器(S-G 滤波器)对光谱数据进行进一步处理。

图 3(c)为经过 S-G 滤波器处理后的吸收光谱，图 3(d)为经过 S-G 滤波器处理后的一阶导数吸收光谱。可以看到，光谱曲线具有了较高的平滑度，能提高矫正模型的精度。在后续的数据处理中，都是以一阶导数加上 S-G 滤波进行预处理后进行的。



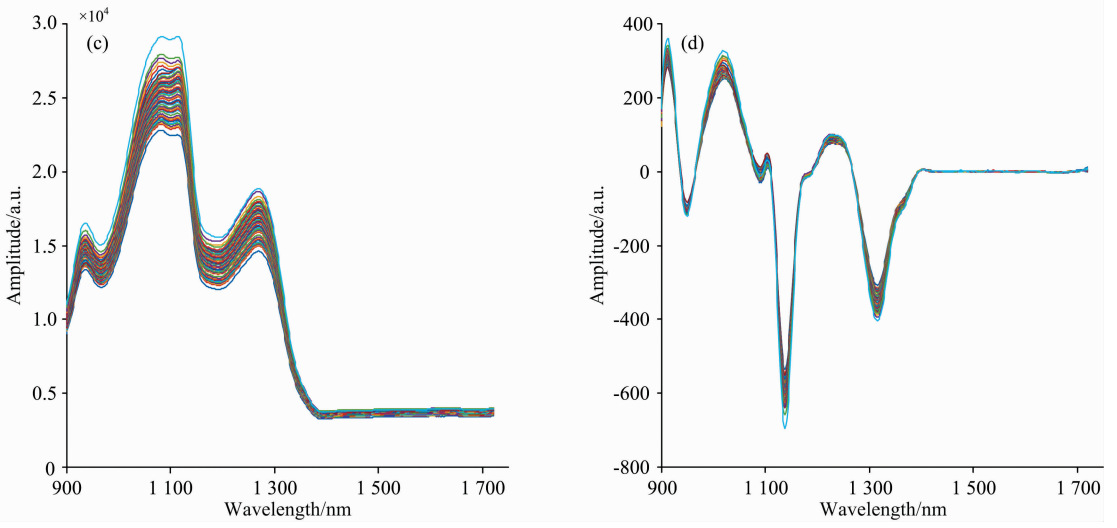


图 3　利用不同方式处理的近红外光谱吸收图

(a)：原始谱；(b)：原始谱一阶导数；(c)：原始谱 S-G 滤波；(d)：原始谱一阶导数 S-G 滤波

Fig. 3　NIR absorption spectra treated by different ways

(a)：Original；(b)：First derivative；

(c)：Original spectra treated by S-G filtering；(d)：First derivative+S-G filtering

3.2　PLS-SPA 光谱特征提取

由于全波段所含有信息量过大，每次计算需要大量的时间。因此采用 SPA 筛选特征波长，不仅有利于减少运算量，加快运算速度，还对今后实际应用时，例如近红外含水率测量传感器的制作有参考价值。利用四种不同的预处理方法进行 SPA 选择的变量数及对应的变量如表 2 所示。

表 2　不同预处理方法使用 SPA 进行选择的变量数和变量

Table 2　Number and value of selected variables by SPA for different preprocessing methods

Method	Preprocessing	Number of selected	Selected variable
SPA	Original spectrum	33	1 200.73, 1 062.36, 1 122.77, 1 059.08, 1 258.9, 1 068.9, 1 266.97, 1 065.63, 1 268.58, 1 072.18, 1 263.74, 1 055.81, 1 252.45, 1 075.45, 1 265.35, 1 052.53, 1 271.8, 1 060.72, 1 260.52, 1 063.99, 1 257.29, 1 057.45, 1 273.41, 1 073.81, 1 254.06, 1 070.54, 1 262.13, 1 067.27, 1 255.68, 1 078.72, 1 278.25, 1 054.17, 1 270.19
	First derivative spectrum	40	1 337.73, 1 139.05, 1 090.15, 1 135.8, 1 093.42, 1 137.42, 1 086.89, 1 134.17, 1 096.68, 1 132.54, 1 091.79, 1 140.68, 1 083.62, 1 142.3, 1 088.52, 1 130.91, 1 095.05, 1 143.93, 1 099.95, 1 129.29, 1 101.58, 1 145.56, 1 039.42, 1 147.18, 1 036.14, 1 148.81, 1 042.7, 1 127.66, 1 098.32, 1 126.03, 1 103.21, 1 150.43, 1 032.86, 1 305.6, 1 029.57, 1 307.21, 922.37, 1 310.43, 925.68, 1 313.64
	Original spectrum+ S-G filtering	41	1 184.52, 1 063.99, 1 263.74, 1 067.27, 1 262.13, 1 065.63, 1 265.35, 1 062.36, 1 260.52, 1 060.72, 1 266.97, 1 068.9, 1 258.9, 1 070.54, 1 268.58, 1 059.08, 1 257.29, 1 072.18, 1 270.19, 1 057.45, 1 255.68, 1 073.81, 1 254.06, 1 055.81, 1 271.8, 1 054.17, 1 252.45, 1 075.45, 1 273.41, 1 052.53, 1 250.84, 1 077.08, 1 275.02, 1 050.89, 1 249.22, 1 078.72, 1 276.63, 1 049.26, 1 247.61, 1 080.35, 1 278.25
	First derivative spectrum+ S-G filtering	31	1 174.79, 1 137.42, 1 091.79, 1 139.05, 1 093.42, 1 135.8, 1 090.15, 1 140.68, 1 088.52, 1 134.17, 1 095.05, 1 142.3, 1 096.68, 1 132.54, 1 086.89, 1 143.93, 1 085.25, 1 130.91, 1 098.32, 1 145.56, 1 099.95, 1 129.29, 1 083.62, 1 147.18, 1 101.58, 1 127.66, 1 081.98, 1 148.81, 1 037.78, 1 150.43, 1 036.14,

如表 2 所示，四种预处理方法用 SPA 选择的变量数目不同，表 2 中的具体变量是按照投影其均方根误差的大小进行排列，即第一个为最优的变量(波长)，第二个为剩余子集中最优，依次排列。例如一阶导数与 S-G 滤波处理后的经过

SPA 选择的第一个变量为 1 174.79 nm，就是在所选择变量集的 31 个变量中最优的波长。虽然预处理方法不同导致选择的变量个数和变量都不全相同，但是有很多相同或相近的波长在所有处理方法中都有出现。因为一阶导数光谱与 S-G 滤波预处理方法消除了基线漂移和随机噪声等不利因素且保证了数据的完整性，所以一阶导数光谱与 S-G 滤波预处理方法选择的变量是四种方法中最少的。SPA 选择的不同变量数的均方根误差值如图 4(a)表示，可以看到，在 15 个变量之前曲线下降很快，说明数据过拟合；15 到 30 个变量时，均方根误差呈缓慢下降的趋势，波动不大，直至第 31 个波段为 RMSE 最小值，此处相对系数 $r=0.992\ 3$，为最大值。因而油中含水量光谱波长选择 31 个。图 4(b)表示的是使用 SPA 对高含水原油光谱数据进行降维所得到的含水量波长在全谱中的分布情况。在图中用空心圆标出。

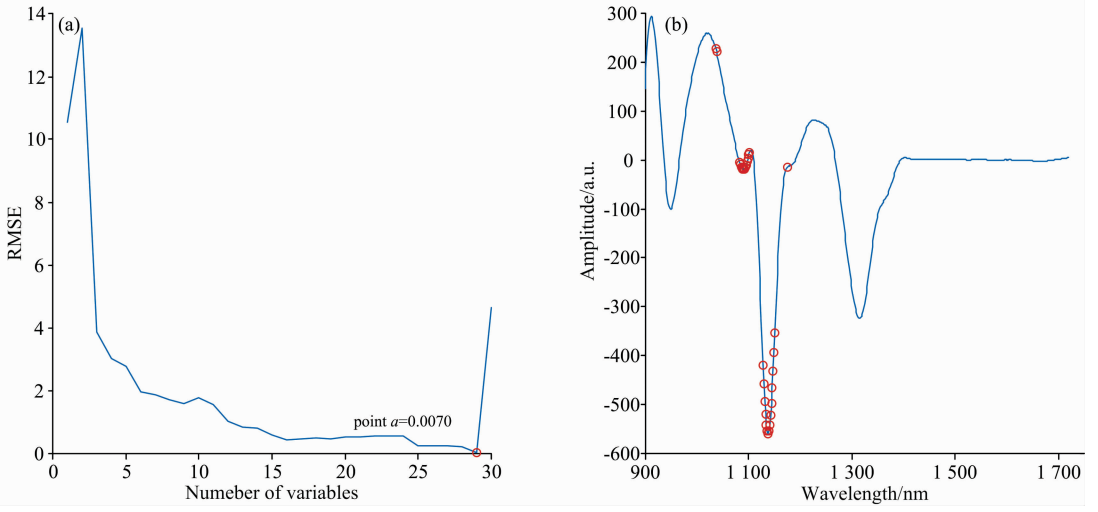


图 4 SPA 选择过程中 RMSE 的变化和特征波长的选取

(a)：RMSE 值；(b)：特征波长分布

Fig. 4 Change of RMSE and selection of characteristic wavelength in SPA selection process

(a)：RMSE；(b)：Characteristic wavelength

表 3 SPA-PLS 与 MLR 模型的预测结果

Table 3 Prediction results by SPA-PLS models and MLR models

Preprocessing	Selected/latent variables number	RMSE of SPA/PLS	RMSE of MLR	r of SPA/PLS	r of MLR
Original spectrum	33/27	0.085 9	0.283 3	0.932 4	0.943 1
First derivative spectrum	40/29	0.048 4	0.222 2	0.979 1	0.944 4
Original spectrum +S-G filtering	41/27	0.015 8	0.384 5	0.972 2	0.925 0
First derivative spectrum+S-G filtering	31/30	0.007 0	0.099 9	0.998 3	0.967 1

3.3 近红外模型的建立及预测

利用 PLS 模型和多元线性回归(MLR)模型对不同方法处理后的光谱分别建模，光谱数据和油中含水率值作为输入值，建立预测模型。利用一阶导数光谱和 S-G 滤波处理后的 RMSE 值为 0.083 3，相关系数 r 为 0.920 6，虽然精度高于 MLR(RMSE=0.283 3，$r=0.943\ 1$)，但该模型使用了全光谱 512 个波段的信息，计算量大，处理时间长，因此需要有效的波长选择方法，提取有效波长，进一步优化模型。

利用 SPA 对使用不同预处理方法后的光谱数据进行有效波长提取，分别建立相应的 SPA-PLS 模型，并使用 MLR 进行对比，预测结果如表 3 所示。

从图 4(b)中可以看到，利用一阶导数 S-G 滤波所选取的波长数量为 31 个，被选取的波长点集中在 1 100 和 1 150 nm 处，与图 2 油中含水量在 1 100～1 200 nm 附近出现的吸收峰差别一致。这说明，可以利用这类吸收峰作为特征波长来建立原油含水率的预测模型。

通过对不同预处理建模和预测效果进行比较，不同预处理方法后 SPA-PLS 模型比原始光谱的 PLS 模型和 MLR 模型在校正集和验证集的预测相关系数都有所提高，但仅使用一阶导数的 SPA+PLS 和 MLR 方法的 RMSE 值都相比原始数据略有增大，说明仅使用一阶导数作为数据处理会出现消除有用信息的问题。此外其他几种方法都要优于原始光谱，说明对数据进行预处理是必要的。在采用一阶导数光谱与 S-G 滤波处理后 SPA-PLS 的效果最佳，其对验证集样本进行预测结果，RMSE=0.007 0，$r=0.998\ 3$，高于 MLR 及其他处理方法，获得了满意的预测精度。

4　结　论

通过搭建的实验装置对相同厚度不同含水率的原油进行近红外光谱检测，利用一阶导数处理加上 S-G 滤波来提高校正模型的精度。运用 SPA 选择油中含水量的近红外光谱波长，通过比较不同的预处理方法运用 SPA 选择的变量个数和对应变量，有很多相同或相近的波长被选出。使用一阶导数光谱和 S-G 滤波预处理方法经过 SPA 选择波长数量为 31 个，它们集中在一阶导数与 S-G 滤波谱中 1 100 和 1 150 nm 附近的波段，无信息的平缓区域几乎没有波长被选取，与油中含水量在 1 150 nm 附近出现明显的吸收峰差别一致。

一阶导数光谱与 S-G 滤波处理后的全谱 PLS 预测模型的均方根误差为 0.083 3，相关系数为 0.920 6，误差较小。但作为后续量信息检测的应用，该模型计算量大，处理时间长。在采用一阶导数光谱与 S-G 滤波处理后 SPA-PLS 的效果最佳，结果表明，利用 SPA 提取出的 31 个特征波长建立的模型仅占全波段的 6.05%，RMSE=0.007 0，r=0.998 3，并且优于 MLR 模型(RMSE=0.099 9，r=0.967 1)获得了较好的精度。为进一步研究高含水原油的快速检测与在线监测奠定了基础。

References

[1] HE Qian-qian, YUE Lai-shen(贺倩倩，跃来深). Journal of Xi'an Technological University(西安工业大学学报), 2016, 36(10): 792.
[2] CHEN Hong, YUE Lai-shen, TONG Yi-jie, et al(陈　鸿，跃来深，仝毅杰，等). Journal of Xian Technological University(西安工业大学学报), 2017, 37(12): 870.
[3] SUN Tong, WU Yi-qing, LI Xiao-zhen, et al(孙　通，吴宜青，李晓珍，等). Acta Optica Sinica(光学学报), 2015, 36(6): 342.
[4] Mazurek S, Szostak R, Kita A. Journal of Molecular Structure, 2016, 1126: 213.
[5] Elradi Abass, Satti Merghany. J. Sc. Tech., 2011, 13(3): 137.
[6] Zaitcev E V, Grigoriev B V, Mikhailov P Y, et al. The Infrared Method of Determinin the Water-Cut of a Nonhomogeneous Water-Gaz-Oil Stream. SPE Russian Petroleum Technology Conference and Exhibition, SPE-182105-MS, 2016.
[7] Lv H, Su X, Wang Y, et al. Chemosphere, 2018, 206: 293.
[8] Douglas R K, Nawar S, Cipullo S, et al. Science of the Total Environment, 2018, 626: 1108.
[9] Zamora D, Blanco M, Bautista M, et al. Talanta, 2012, 89: 478.
[10] Borges G R, Farias G B, Braz T M, et al. Fuel, 2015, 147: 43.
[11] Zude M, Pflanz M, Spinelli L, et al. Journal of Food Engineering, 2011, 103(1): 68.
[12] Costa G B D, Fernandes D D S, Gomes A A, et al. Food Chemistry, 2016, 196: 539.
[13] Ghasemi-Varnamkhasti M, Mohtasebi S S, Rodriguez-Mendez M L, et al. Talanta, 2012, 89(2): 286.
[14] CHEN Bin, LIU Ge, ZHANG Xian-ming(陈　彬，刘　阁，张贤明). Infrared and Laser Engineering(红外与激光工程), 2013, 42(12): 3168.
[15] Krepper G, Romeo F, Fernandes D D D S, et al. Spectrochimica Acta Part A: Molecular & Biomolecular Spectroscopy, 2017, 189: 300.

Water Content Prediction for High Water-Cut Crude Oil Based on SPA-PLS Using Near Infrared Spectroscopy

HAN Jian, LI Yu-zhao, CAO Zhi-min*, LIU Qiang, MOU Hai-wei
School of Electronic Science, Northeast Petroleum University, Daqing　163318, China

Abstract　Accurately and timely measuring water content of the crude oil is of great significance for water injection strategy adjustment, crude oil exploitation capacity assessment, and oil well development lift prediction. However, at present, most of China's oil fields have entered the mid- or late- development stage with high water content. And the corresponding water content is difficult to measure accurately. Under this circumstance, this paper carried out research on the measurement of water content of the crude oil using near-infrared spectroscopy. Specifically, commonly employed methods for measuring water content of the crude oil were introduced, and advantages and disadvantages of these methods were analyzed. Theoretically, since the near-infrared absorption band of water is significantly different from the absorption of C—H bond in crude oil, according to Lambert-Beer's law of absorption and linear law of absorbance, there is a strong response difference in the near-infrared spectrum of high water cut crude oil with different water content. Therefore, we proposed to use near-infrared spectroscopy to accurately measure the crude oil with high water content. And then, by analyzing the measured near-infrared spectrum, non-linear mapping between the water content of the testing crude oil and the near-infrared spectrum can be established. With the obtained non-linear map-

ping model, water content of the crude oil can be accurately calculated. In order to evaluate the performance of this method, we constructed a hardware platform for collecting near-infrared data. In this platform, Incandescent lamp was employed as a light source, and near-infrared spectrometer (Ocean Optics NIR512) was used to collect near-infrared in range 900～1 700 nm with 512 uniformly divided sub bands. The collected data were stored in the computer using the spectrometer supporting software. With the obtained near-infrared data, the raw data preprocessing was performed to reduce the influence of temperature and high frequency random noise, sample unevenness, baseline drift, light scattering, and et al. In this paper, S-G filtering, or first order derivative, or S-G filtering+first order derivative techniques were employed as the preprocessing method; Successive Projection Algorithm (SPA) was used to reduce the dimension of the raw data; Partial Least Square (PLS) and Multiple Linear regression (MLR) were employed to construct the corresponding non-linear mapping model; Root Mean Square Error (RMSE) and Correlation coefficient (R) were used to evaluate the quantitative measuring performance. Experimental results illustrated that: model constructed using S-G filtering+first order derivative as preprocessing method can achieve the best RMSE (RMSE=0.007 0, r=0.998 3); Model constructed with reduced dimensional data using SPA method is better than the one (RMSE=0.083 3, r=0.920 6) constructed by PLS with full band data and the one (RMSE=0.099 9, r=0.967 1) constructed by MLR with full band. Obviously, although the 31 dimensionality-reduced feature bands obtained by SPA method are only 6.05% of the full band data, the corresponding water content measuring accuracy of the crude oil is very promising. In general, we validate the feasibility of using spectroscopy technique to measure water content of the high water content crude oil, and satisfactory accuracy can be achieved. Therefore, it can be said that this paper provides a new method for water content measurement of high water content crude oil, and provides reference for accurately and timely measuring high water content crude oil using near-infrared spectroscopy.

Keywords Near-infrared spectroscopy (NIR); High water content crude oil; Successive projection algorithm (SPA); Partial least square (PLS)

(Received Oct. 18, 2018; accepted Feb. 15, 2019)

* Corresponding author