第37卷，第8期
2017年8月

光谱学与光谱分析
Spectroscopy and Spectral Analysis

Vol. 37, No. 8, pp2585-2589
August, 2017

基于PLS的水体重金属LIBS特征变量筛选方法研究

胡　丽[1]，赵南京[2*]，李大创[1]，唐　磊[1]，方　丽[2]

1. 合肥师范学院电子信息工程学院，安徽 合肥　230061
2. 中国科学院安徽光学精密机械研究所，安徽 合肥　230031

摘　要　在水体重金属激光诱导等离子体光谱定量分析中，一般提取光谱的多个特征变量进行浓度反演，但变量之间所包含的光谱信息可能存在重叠，回归模型的复杂程度也随之增大。为提取有效特征变量，研究了基于偏最小二乘法(PLS)的变量筛选方法。该方法以待测元素浓度为因变量，多个与待测元素浓度相关的LIBS光谱特征值为自变量，进行PLS建模；依据各原始变量的投影重要性指标值进行变量筛选，提取最优变量子集。结果表明湖库水体中Pb元素的最优变量子集为Pb Ⅰ 405.78 nm峰值及峰值前相邻点光谱值、内标校正值和信背比值，训练集的复相关系数$R_m^2=0.912$。以优化变量组合进行PLS回归分析，测试集预测结果的RSD和RE分别为10.2%和7.9%，显著优于内标法的预测结果。结果还表明，变量筛选结果对于不同元素和不同水样具有一定适用性。研究结果为水体重金属LIBS定量分析提供了优质特征数据，研究方法为其他涉及变量筛选的定量分析提供了参考。

关键词　光谱学；激光诱导击穿光谱；变量筛选；PLS方法
中图分类号：O433.4　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2017)08-2585-05

引　言

激光诱导击穿光谱技术(laser-induced break-down spectroscopy，LIBS)是一种通过探测激光诱导等离子体特征光谱，由分析线的位置及信号强度获取物质成分和浓度的快速分析技术[1]，在化学、医学、生物学及环境监测等领域开展了较多研究工作[2-5]。LIBS定量测量一般采用待测元素的某一条特征谱线进行浓度反演，但单特征峰容易受不确定因素干扰，测量结果的稳定性不高[6]。因此，需要进一步对表征待测元素浓度的LIBS光谱信息进行扩展，提取尽可能多的有用信息，实现多个变量之间信息互补，减小基体效应、系统波动等因素的影响。然而，随着特征向量的进一步开发，彼此之间所包含的光谱信息存在重叠，从而导致严重的共线性问题，影响回归模型质量[7]。如何对大量光谱数据进行筛选，获得有效变量子集，就成为LIBS定量研究中一个关键问题。变量筛选也就是降维，偏最小二乘(partial least squares，PLS)方法能通过描述性方法观察变量间的相关性，对变量信息进行分析和综合，得到对因变量影响较大的自变量[8]。因此，本文以水体中待测重金属浓度为因变量，多个与待测元素浓度相关的LIBS光谱特征值为自变量，进行PLS建模；计算各原始自变量对因变量的影响程度，确定最优变量子集，提高LIBS定量分析的稳定性。

1　实验部分

1.1　仪器与样品

实验采用Nd∶YAG脉冲激光器(Quantel公司，Ultra型，波长1 064 nm，能量100 mJ，脉宽6 ns，频率1 Hz)，采用Avantes三通道光纤光谱仪(AVS-DESKTOP-USB2型，探测范围为200～500 nm，分辨率为0.08 nm)。实验的延时和门宽分别设置为2 μs和1.05 ms。具体实验装置如图1所示。

选取合肥市西北近郊某水库进行采样(16个采样点，采取水面下约0.1 m处的水样)，首先对水样中可能存在的金属元素进行LIBS测量，结果表明水样中含有一定量的Zn和Cu元素，未测出Ni和Pb元素。

实验室通过添加$Pb(NO_3)_2$和$Ni(NiO_3)_2\cdot 6(H_2O)$，将

收稿日期：2015-11-16，**修订日期**：2016-03-02
基金项目：国家自然科学基金项目(41574180)，国家(863)计划项目(2013AA065502，2014AA06A513)，安徽省科技攻关重大项目(1501041119)，中国科学院环境科学与技术重点实验室开放基金项目(2005DP173065-2017-01)资助
作者简介：胡　丽，女，1981年生，合肥师范学院电子信息工程学院讲师　　e-mail：lhu@aiofm.ac.cn
*通讯联系人　　e-mail：njzhao@aiofm.ac.cn

水样配制成 Pb 和 Ni 的混合溶液，16 个水样浓度不等，最大浓度为 10 mg·L^{-1}，最小浓度为 0.015 6 mg·L^{-1}，通过自动制样装置富集 8 mL 样液，制成待测样本，并进行无重复采样打点。

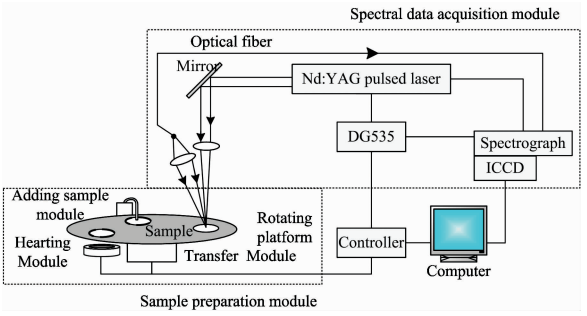

图 1 实验装置示意图

Fig. 1 Schematic diagram of experimental setup

1.2 方法

1.2.1 特征向量拓展与提取

(1) 待测元素浓度相关参数

待测元素特征峰的峰强度直接和浓度相关，且通过多个特征峰提取可以实现信息互补。在实际测量中，由于水体重金属的检测处于痕量级，峰强度较小的特征峰，能否为定量反演提供有效信息，将通过变量筛选来完成。图 2 为样品中 Pb 元素的 LIBS 光谱。考虑到特征谱线积分强度同时包含了谱线的强度与形状信息，可以提取积分强度作为特征变量。受特征谱线的半高全宽和光谱仪分辨率的限制，以特征峰前后共 5 个光谱数据的波长范围作为积分计算的波段宽度。

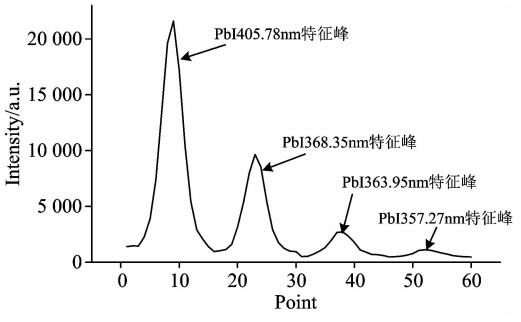

图 2 Pb 元素实测光谱(2.5 mg·L^{-1})

Fig. 2 The measured spectrum of Pb (2.5 mg·L^{-1})

(2)基体效应相关参数

等离子体的密度是表征等离子体特性的重要参数，而等离子体密度无法直接测量，一般可以通过特征谱线展宽进行表征[9]。因此，将特征谱线半高全宽内的光谱数据作为输入向量，有助于提供等离子体激发和发射的信息。除此之外，实验采用石墨基底进行样液的富集，基底含有 C(质量分数为 99.99%)，以及微量的 Ca，Mn 和 Fe 等元素，基底元素含量和组成的差异直接以基体效应的形式影响待测元素光谱的强度和稳定性，因此，可以将基底主要组成元素的特征峰值引入定量分析模型，通过变量筛选决定相应的权重。

(3)Pb 元素特征向量提取

为了更加全面的反映待测元素的激发和发射信息，以 Pb 元素为例，引入 Pb Ⅰ 405.78 nm，Pb Ⅰ 368.35 nm 和 Pb Ⅰ 363.95.30 nm 峰值强度及对应的积分强度；Pb Ⅰ 405.78 nm 和 Pb Ⅰ 368.35 nm 特征峰半高全宽波段内光谱数据；基体组成元素 C Ⅰ 247.86 nm 峰值、Mg Ⅰ 383.829 nm 峰值、Ca Ⅰ 396.84 nm 峰值、Ca Ⅰ 393.366 nm 峰值、Fe Ⅰ 404.58 nm 峰值；同时，对于水体中微量重金属元素，分别利用背景、Fe Ⅰ 404.58 nm 峰值对共振线 Pb Ⅰ 405.78 nm 特征峰值进行校正，得到校正后的特征向量。综上，共得到 Pb 元素的 25 个特征变量。

1.2.2 PLS 变量筛选方法

基于 PLS 算法建立待测元素浓度与 LIBS 光谱多个特征变量的定量回归模型，利用交叉校验确定合适的主成份数(或潜变量数)A。在 PLS 回归分析中，变量投影重要性指标 VIP_j(variable importance projection，VIP)可以表征自变量对因变量的解释能力[10]。第 x_j 个变量的 VIP_j 计算公式如下[11]

$$\mathrm{VIP}_j = sqr\left[m\sum_{a=1}^{A}\omega_{ja}^2 r^2(y,t_a)\Big/\sum_{a=1}^{A}r^2(y,t_a)\right] \tag{1}$$

式(1)中 m 为原始变量数，ω 为合适潜变量数 A 对应的权矢量，$r^2(y,t_a)$是因变量 Y 矩阵(对于单因变量实际上是一矢量)与第 a 个潜变量的得分矢量 t_a 之间的相关系数[11]见式(2)

$$r^2(y,t_a) = 1-\sum_{i=1}^{n}(y_i-t_{ia})^2\Big/\sum_{i-1}^{n}(y_i-\bar{y})^2 \tag{2}$$

构建 VIP_j 值的理论依据如下：原始变量 x_j 对因变量 Y 的影响是通过潜变量 t_a 来实现的，假设 t_a 对 Y 的影响较大，而 x_j 与 t_a 相关性较强，则 x_j 对 Y 的影响就视为很大。换句话说，若是在 $r^2(y,t_a)$值较大的 t_a 上，w_{ja} 取较大的值，则 x_j 对 Y 的影响就很大。所以，原始变量的投影重要性指标可以表征原始变量对因变量的影响程度。

分析合适潜变量数 A 时所有原始变量对模型的贡献，某变量 VIP 值越大，该变量越重要，提取 VIP 值较大的变量，即为主要考察的原始变量。结合 PLS 建模过程，水体重金属 IIBS 特征向量筛选过程如下：首先，将经过预处理后 LIBS 光谱特征向量作为自变量集 X，待测元素浓度为因变量集 Y，建立 PLS 回归模型，以留一法(leave-one-out，LOO)交互检验统计量为目标函数，得到最佳潜变量数 A；第二步，计算各个自变量的 VIP 值，从模型中删除 VIP 最小的自变量，重复上述建模和 VIP 计算过程。直到训练集的复相关系数 R^2 出现相对极大值(或均方根误差出现相对极小值)，即停止变量筛选，对应的特征向量进入最终的 LIBS 定量回归模型。具体流程如图 3 所示。

2 结果与讨论

2.1 变量筛选结果

实验对湖库水体的16个样本进行测量，重复测量30

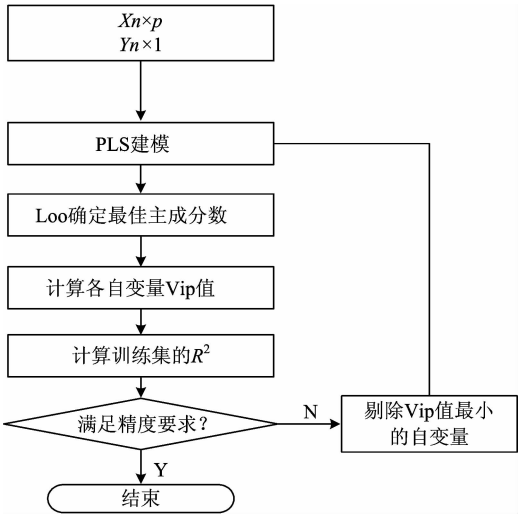

图 3　PLS 变量筛选流程图

Fig. 3　Flow chart of variable selection

次，得到数据矩阵。随机抽取每种浓度中的 20 组数据作为训练集，根据 1.2.1 节，提取 Pb Ⅰ 405.78 nm 峰值强度等 25 个特征值作为自变量，待测元素浓度为因变量 Y，进行 PLS 建模，以 LOO 交叉校验统计量为目标函数，得到模型的最佳潜变量数 $A=4$，以及对应的训练集修正后的复相关系数和均方根误差；计算各个自变量的 VIP 值，筛除 VIP 最小值对应的自变量，循环利用上述过程进行变量筛选。

结果表明，训练集的复相关系数($R_m^2=0.912$)在变量数 $m=4$ 时是一个相对峰值。因此，选择变量数 $m=4$，即变量：Pb Ⅰ 405.78 nm 峰值及峰值前相邻点光谱值、Fe Ⅰ 404.58 nm 内标校正的 Pb Ⅰ 405.78 nm 峰值强度和背景校正的 Pb Ⅰ 405.78 nm 峰值强度进行最后建模。Pb 元素的变量筛选结果表明，对于湖库水体中微量重金属 LIBS 检测，强度较大的特征峰 Pb Ⅰ 405.7 nm 及与其密切相关的峰值前相邻点光谱值能够提供更为准确、灵敏的信息；通过背景和内标元素校正的特征峰值可以降低光谱的随机波动，定量分析的稳定性显著提高。另外，变量筛选的过程还表明，C Ⅰ 247.86 nm，Ca Ⅰ 396.84 nm，Ca Ⅰ 393.366 nm 等基底主要组成元素的特征峰对 Pb 元素定量分析的影响较小，可能由于这些元素含量较高，谱线强度较大，与痕量浓度的重金属元素相关性较小。

2.2　变量筛选结果有效性分析

为验证筛选结果能否为测试集提供优质的输入变量，采用湖库水样剩余的 160 组数据作为测试集，选取最优变量组合(Pb Ⅰ 405.78 峰值、Pb Ⅰ 405.78 峰值前相邻点光谱值、Pb Ⅰ 405.78 内标校正值和 Pb Ⅰ 405.78 背景校正值)作为 PLS 回归模型的输入向量，并与光谱分析中最常用的内标法进行对比。使用相对标准偏差(relative standard deviation，RSD)和相对误差(relative error，RE)用于衡量回归预测的精密度和准确度[12]。

对于测试集的 16 个样本，内标法回归和最佳变量组合回归的 RSD 平均值分别为 18.0%和 10.2%，RE 平均值分别为 21.2%和 7.9%。结果表明优化后的特征输入向量提取了更多与待测元素浓度相关的信息，LIBS 定量分析的稳定性和准确性得到了一定程度提高。图 4 给出了两种变量组合下预测集的 RSD 和 RE 值变化情况。对于内标单变量输入模型，在低浓度区 RSD 值较大，模型预测稳定性较差，说明对于基体差异较大的湖库水体，一元定标往往失效，而优化后的多元输入向量从多个角度对光谱的波动进行校正，提高了回归模型预测的稳定性。最佳变量组合预测的 RE 值整体较小，这里值得注意的是，内标法预测高浓度样本时出现了较大偏差，可能的原因是选取的内标峰值偏小，当待测元素浓度较高时，无法起到内标校正的效果。

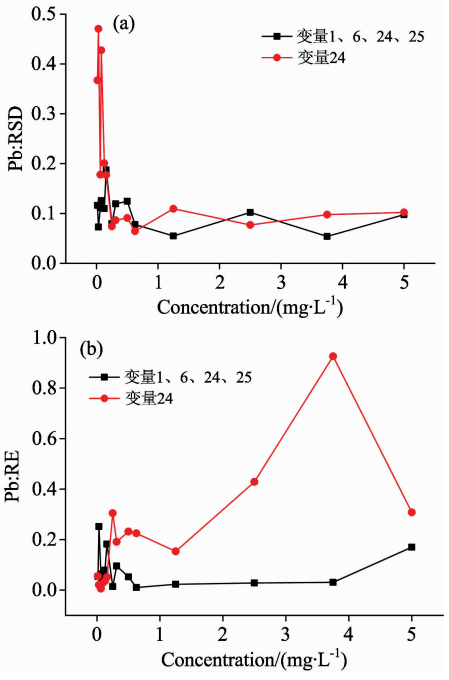

图 4　Pb 元素 RSD 和 RE 随浓度变化趋势

变量 1：Pb Ⅰ 405.78 nm 峰值；变量 6：Pb Ⅰ 405.78 nm 峰前相邻点光谱值；变量 24：Fe Ⅰ 404.58 nm 校正的 Pb Ⅰ 405.78 nm 峰值强度；变量 25：背景校正的 Pb Ⅰ 405.78 nm 峰值强度

Fig. 4　Relative standard deviation and relative error comparison chart of test set

补英文

2.3　变量筛选结果适用性分析

2.3.1　不同元素对比分析

为验证筛选结果对于不同元素的适用性程度，进一步分析 1.2.2 节混合溶液中的 Ni 元素，提取 Ni Ⅰ 341.48 nm 峰值强度等 18 个特征值作为自变量，待测元素浓度为因变量 Y，采用 1.2.2 节方法进行变量筛选。结果表明，训练集的复相关系数($R_m^2=0.972$)在变量数 $m=3$ 时出现一个相对峰值。因此，选择变量数 $m=3$，即变量：Fe Ⅰ 344.06 nm 内标校正的 Ni Ⅰ 361.94 nm 峰值、Ni Ⅰ 349.30 nm 峰值和 Ni Ⅰ 351.50 nm 峰值进行最后建模。

对比 Pb 元素的变量筛选结果，两种元素峰值强度相对较大的特征峰(如 Pb Ⅰ 405.78 nm，Ni Ⅰ 349.30 nm 和 Ni

Ⅰ 351.50 nm 峰值)，内标元素校正的峰值(如 Pb Ⅰ 405.78 nm 内标值和 Ni Ⅰ 361.94 nm 内标值)可以提供有效的浓度信息，消除系统参数波动的影响，定量分析的准确性和稳定性得到提高；同时，石墨基底的主要组成元素(C，Ca 和 Mg 等元素)对两种元素定量分析的影响都较小，因此，对于同一批次的石墨基片，可以忽略基底元素对微量的重金属元素定量检测的影响。但不同元素变量筛选结果存在一定的差异，如背景对 Ni 元素无法起到校正作用，可能与此处的背景相对较弱有关。

2.3.2 不同水样对比分析

为了验证变量筛选结果对于不同水样的适用性程度，选择铜陵市某矿石厂附近的水体为研究对象，取水面下约 0.1 m 处的水样，得到 7 个样品。以分析 Pb 和 Ni 元素为例，由于 Pb 和 Ni 元素含量较低(部分样品低于 LIBS 检出限)，采用添加标准溶液的方式进行样液配制，通过 ICP-AES 方法检测样液的浓度，如表 1 所示。

表 1 样液中 Pb 和 Ni 元素含量

Table 1 Content for Pb and Ni in the sample solution

编号	采样点	检测浓度(ICP-AES)/(mg·L^{-1})			
		原样液		配制后样液	
		Pb	Ni	Pb	Ni
1	矿山上游水	0.006 4	0	4.003 84	4
2	矿山中游水	0.005 4	0	7.001 62	7
3	矿山下游水	0.006 5	0	5.003 25	5
4	矿山下游农田水	0.002 6	0.001 4	6.001 04	6.000 56
5	炼石厂区水	0.005 2	0	3.003 64	3
6	炼石厂污水	0.420 8	1.195	8.084 16	8.239
7	炼石厂污水处理后	0.003 3	0	2.002 64	2

利用 PLS 变量筛选方法确定待测元素的最佳特征向量，Pb 元素的最佳输入变量为 Pb Ⅰ 405.78 nm 特征峰的信背比值，Ni 元素的最佳输入变量为 Ni Ⅰ 349.3 nm 和 Ni Ⅰ 351.5 nm 峰值强度组合向量。对比湖库水样，结果表明不同水样变量筛选结果存在差异。但强度较大的 Pb Ⅰ 405.78 nm(背景校正)、Ni Ⅰ 349.3 nm 和 Ni Ⅰ 351.5 nm 峰值可以提供较为准确的特征信息，对于两种水样仍然是最佳输入变量，研究结论为其他水体中 LIBS 检测的特征向量选取提供了参考。

3 结 论

为解决 LIBS 定量分析中多个特征变量之间的信息重叠和共线性问题，研究了基于 PLS 的变量筛选方法。该方法以待测元素浓度为因变量，多个与待测元素浓度相关的 LIBS 光谱特征变量为自变量，进行 PLS 建模；依据各原始变量的投影重要性指标 VIP 值进行变量筛选，提取最优变量子集。针对湖库水体，Pb 元素最优变量子集为 Pb Ⅰ 405.78 nm 峰值及峰值前相邻点光谱值、Fe Ⅰ 404.58 nm 内标校正的 Pb Ⅰ 405.78 nm 峰值强度和背景校正的 Pb Ⅰ 405.78 nm 峰值强度，训练集的复相关系数。以优化变量组合进行 PLS 回归分析，测试集预测结果的 RSD 和 RE 分别为 10.2% 和 7.9%，显著优于内标法的单变量预测结果。进一步研究了变量筛选结果对于不同元素和不同水样的适用性，实验表明变量筛选结果存在一定适用性，对于湖库水体，两种元素(Pb 和 Ni)相对较大的特征峰和内标校正峰值可以提供有效信息；对于不同水样，强度较大的特征峰值具有一定适用性，但由于基体差异较大，内标元素的校正作用失效。

References

[1] Brech F, Cross L. Appl. Spectrosc., 1962, 16(3): 59.

[2] Sadegh Cheri M, Tavassoli S H. Applied Optics, 2011, 50(9): 1227.

[3] XU Yuan, YAO Ming-yin, LIU Mu-hua, et al(徐 媛，姚明印，刘木华，等). Acta Optica Sinica(光学学报), 2011, 31(12): 290.

[4] LI Wen-hong, WU Zhi-xiang, WANG Rui-wen, et al(李文宏，武志翔，王芮雯，等). Chinese J. Lasers(中国激光), 2014, 41(6): 282.

[5] CHEN Xing-long, DONG Feng-zhong, TAO Guo-qiang, et al(陈兴龙，董凤忠，陶国强，等). Chinese J. Lasers(中国激光), 2013, 40(12): 243.

[6] WANG Chun-long, LIU Jian-guo, ZHAO Nan-jing, et al(王春龙，刘建国，赵南京，等). Acta Optica Sinica(光学学报), 2013, 33(3): 314.

[7] WANG Chun-long, LIU Jian-guo, ZHAO Nan-jing, et al(王春龙，刘建国，赵南京，等). Acta Phys. Sin.(物理学报), 2013, 62(12): 125201.

[8] WANG Hui-wen(王惠文). Partial Least Squares Regression Method and Its Application(偏最小二乘回归方法及其应用). Beijing: National Defence Industry Press(北京：国防工业出版社), 1999. 14.

[9] David A C, RadziemskiLeon J. Handbook of Laser-Induced Breakdown Spectroscopy. Cambridge University Press, 2006. 24.

[10] Zhang Y H, Xia Z N, Qin L T, et al. Mol. Graph. Model., 2010, 29: 214.

[11] ZHANG Heng-xi, GUO Ji-lian(张恒喜，郭基联). Multivariate Date Analysis and Application of Small Sample(小样本多元数据分析及应用). Xi'an: Northwestern Polytechnical University Press(西安：西北工业大学出版社), 2002. 99.

[12] Environmental Monitoring Analysis Method Standard System Revision Techical Quidance(环境监测分析方法标准制修订技术导则). Ministry of Environmental Protection(环境保护部), 2010.

Study of Variable Selection Method Based on PLS for Quantitatively Measuring Heavy Metal in Water with LIBS

HU Li[1], ZHAO Nan-jing[2*], LI Da-chuang[1], TANG Lei[1], FANG Li[2]

1. School of Electronic and Information Engineering, Hefei Normal University, Hefei 230061, China

2. Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China

Abstract The spectral characteristics of multiple variables were regularly extracted for concentration inversion in the quantitative analysis of heavy metal with LIBS. However, overlapped spectral information might be contained among variables and the complexity of the regression model also would be increased. To extract effective feature variables, the variable selection method based on PLS was studied. PLS model established with the concentration of the element under test used as the dependent variable and multiple LIBS spectrum characteristic variables used as independent variables. The optimal variable subset was extracted on the basis of the original variable importance projection index for variable selection. The results showed that the optimal variable subset for Pb was composed of the Pb Ⅰ 405.78 nm peak, the spectral values before Pb Ⅰ 405.78 nm peak, the intensity corrected by the internal standard element and the signal to background in lake water, and the correlation coefficient square of the training set was 0.912. The optimal variable subset was used for PLS regression analysis, and RSD and RE of the test set were 10.2% and 7.9%, respectively, significantly better than the predictive results of the internal standard. It also showed that variable screening results for different elements and different water samples was applicative to some extent, but the internal standard element failed in correction for different water samples. The results provided high quality characteristic data for the LIBS quantitative analysis, and the methods also provided a reference for other quantitative analysis involving variable selection.

Keywords Spectroscopy; Laser induced breakdown spectroscopy; Variable screening; PLS model

(Received Nov. 16, 2015; accepted Mar. 2, 2016)

* Corresponding author