

## 基于偏最小二乘法的巢湖悬浮物浓度反演\*

刘忠华, 李云梅\*\*, 吕 恒, 徐祎凡, 徐 昕, 黄家柱, 檀 静, 郭宇龙

(南京师范大学虚拟地理环境教育部重点实验室, 南京 210046)

**摘 要:** 悬浮物浓度是水质评价的重要参数. 对 2009 年 6 月巢湖实测的高光谱数据进行小波变换以去除数据冗余, 减少建模时间. 考虑到不同的小波基函数和分解尺度对数据压缩的影响, 采用 3 个指标作为评价标准, 最终选择小波基函数为 Db4, 分解尺度为 4 的小波变换, 经小波变换后, 原来的 451 个波段的光谱数据压缩为 34 个特征变量. 利用小波变换后的光谱数据建立了悬浮物浓度反演的偏最小二乘法 (PLS) 模型, 其中 20 个样本用于建模, 9 个样本用于验证, 结果表明: 当主成分个数为 3 时, PLS 模型具有最佳的效果, 此时模型的  $R^2$  为 0.93,  $R^2$  (pred) 为 0.89,  $PRESS$  为 3.29, 3 个主成分累积解释了 98.60% 的自变量信息和 92.37% 的因变量信息. 此外, PLS 模型能够充分利用高光谱数据信息, 具有较高的精度和稳定性 ( $R^2 = 0.93$ ,  $RMSE = 4.77 \text{ mg/L}$ ,  $MAPE = 9.02\%$ ). 通过与单波段模型、光谱一阶微分模型及波段比值模型的对比如分析得出, PLS 模型无论是从建模样本精度还是验证样本的误差方面均高于传统的经验模型, 适合于利用高光谱数据进行悬浮物浓度的反演.

**关键词:** 小波变换; 偏最小二乘法; 高光谱数据; 悬浮物; 巢湖

## Inversion of suspended matter concentration in Lake Chaohu based on Partial Least-squares Regression

LIU Zhonghua, LI Yunmei, LÜ Heng, XU Yifan, XU Xin, HUANG Jiazhu, TAN Jing & GUO Yulong

(Key Laboratory of Virtual Geographic Environment of Education Ministry, Nanjing Normal University, Nanjing 210046, P. R. China)

**Abstract:** Suspended matter concentration is an important parameter of water quality evaluation. Hyperspectral data measured in Lake Chaohu in June, 2009 were processed by wavelet transform in order to remove data redundancy and reduce modeling time. Three evaluation indexes were selected considering the effect of different wavelet functions and decomposed scales on the data compression, and the wavelet function Db4 and decomposed scale 4 were determined finally. The original hyperspectral data of 451 bands were compressed to 34 feature variables by the wavelet transform. Then 20 samples were used to construct Partial Least-squares Regression (PLS) inversion model of suspended matter concentration, and other 9 samples were used for model verification. The results show that PLS model is suited when the number of principal components is 3 and its  $R^2$  is 0.93,  $R^2$  (pred) is 0.89 and  $PRESS$  is 3.29. These three principal components explain 98.60% of independent variables information and 92.37% of dependent variables information. PLS model with  $R^2$  of 0.93,  $RMSE$  of 4.77 mg/L, and  $MAPE$  of 9.02% can make full use of the information of hyperspectral data, and hence have higher accuracy and stability. In addition, single band model, spectral one-order differential model and band ratio model were used to compare with PLS model. The results show that PLS model is better than traditional empirical models no matter on the accuracy of modeling samples or the error of validation samples, indicating that it is suitable for the inversion of suspended matter concentration by using hyperspectral data.

**Keywords:** Wavelet transform; Partial Least-squares Regression; hyperspectral data; suspended matter; Lake Chaohu

\* 高分辨率对地观测系统国家科技重大专项项目 (E0203/1112/JC01)、国家自然科学基金项目 (40971215)、江苏省 2008 年度普通高校研究生科研创新计划 (CX09B-301Z) 和南京师范大学优秀博士论文培养计划项目 (12432116011036) 联合资助. 2010-05-21 收稿; 2010-10-08 收修改稿. 刘忠华, 男, 1987 年生, 硕士研究生; E-mail: liuzhonghua1987@yeah.net.

\*\* 通讯作者; E-mail: liyunmei@njjnu.edu.cn.

常规的水质监测调查速度慢、监测周期长,难以满足对大面积水质监测的要求. 遥感技术作为一种区域性水环境调查和监测手段,可以克服常规水质监测方法的不足. 然而,要利用遥感信息精确反演水质参数,其关键是建立水质参数和光谱特征之间的响应关系<sup>[1]</sup>. 近年来,随着高光谱技术的迅速发展,其光谱分辨率不断提高,使得利用水质参数和光谱特征间的响应关系进行水质参数反演的精度大大提高<sup>[2]</sup>. 水体中悬浮物的含量是重要的水质参数,尤其是悬浮泥沙含量大小直接影响水体的透明度、浑浊度和水色等光学性质,进而影响到水体的生态条件和初级生产力<sup>[3]</sup>. 目前,国内外许多学者<sup>[4-13]</sup>通过各种方法寻求能反映水质参数的敏感波段对悬浮物浓度进行反演,常用的方法主要有单波段法、波段比值法、光谱一阶微分法等. 随着对水体辐射传输机理认识的深入,结合生物光学模型的半分析方法也得到了广泛的应用. 徐京萍和张兵等<sup>[14-15]</sup>根据生物光学模型在近红外波段的简化条件,建立了悬浮物浓度反演的单波段模型. 然而上述方法仅利用了高光谱数据中的单个波段或几个波段的光谱信息,在选择特征波段的同时丢弃了另外一些对悬浮物浓度反演有用的信息. 偏最小二乘法(Partial Least-squares Regression, PLS)是近年来发展起来的一种新型的数据分析方法,它可以较好地解决普通多元回归无法解决的一些问题,可以实现全谱数据的分析,对样本容量少、自变量多的情况具有独特的优势<sup>[16]</sup>,适合于高光谱遥感数据的处理.

本研究尝试利用巢湖实测高光谱数据结合偏最小二乘法进行巢湖悬浮物浓度的反演. 由于所采集的高光谱数据信息量大,首先利用小波变换对原始数据进行数据压缩、平滑去噪等处理,然后利用小波变换后的特征变量代替原始数据进行悬浮物浓度的反演,并通过与其他模型比较,评价该模型的精度及其适用性,以期为提高巢湖水质监测及水质参数的高光谱反演精度奠定基础.

## 1 材料与方法

### 1.1 采样时间与点位

巢湖位于安徽省中部,东西长 54.5 km,南北宽 21 km,平均水深 4.5 m,水域面积 750 km<sup>2</sup>,是安徽省内最大的湖泊,为我国五大淡水湖之一. 2009 年 6 月 13-16 日对巢湖 32 个样点进行采样(图 1).

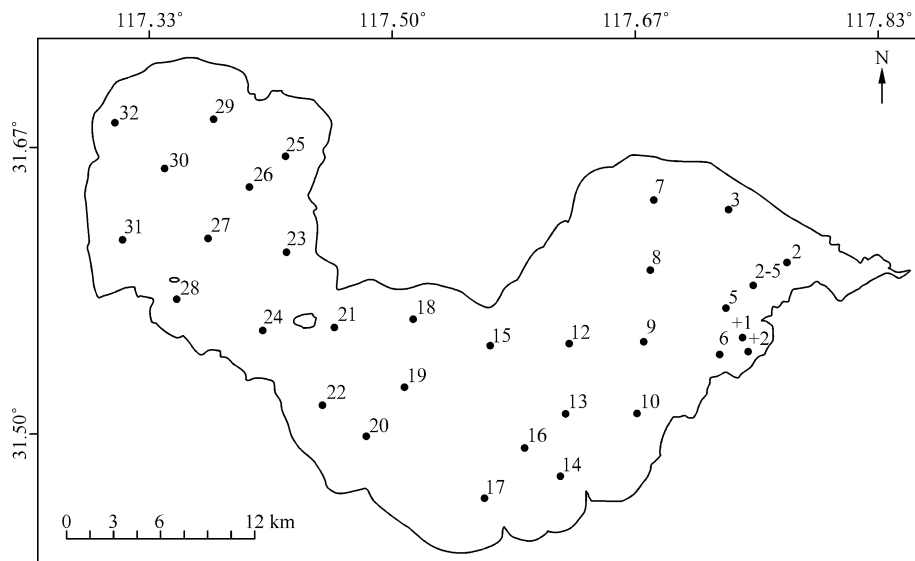


图 1 巢湖样点分布

Fig. 1 Sampling sites in Lake Chaohu

### 1.2 数据的获取

水面的反射光谱采用美国 ASD 公司生产的 FieldSpec Pro 便携式光谱辐射计测量,其波段范围为 350-1050 nm,光谱分辨率为 2 nm. 为减少水体镜面反射和船体阴影的影响,更好地提取出反映水体信息的离水辐

亮度  $L_w$  和遥感反射率  $R_{rs}$ , 采用唐军武等<sup>[17]</sup>提出的关于内陆二类水体水面以上光谱测量的方法. 测量时, 天空晴朗无云, 湖面平静, 待船停稳后, 在甲板开阔处(距水面 1m 左右)分别测量标准灰板、遮挡标准灰板、水体和天空光的光谱信息, 以上 4 个参数各测量 10 条光谱曲线, 在测量水面反射光谱的同时记录各测点的 GPS 坐标和当时的风速、风向以及时间. 遥感反射率通常利用式(1)进行提取<sup>[17]</sup>:

$$R_{rs} = \frac{(L_{sw} - rL_{sky})\rho_p}{\pi L_p^*} \quad (1)$$

式中,  $R_{rs}$  为遥感反射率,  $L_{sw}$  为总的辐亮度,  $L_{sky}$  为天空漫散射光, 不带有任何水体信息,  $r$  参照唐军武等<sup>[18]</sup>的研究取值为 0.025,  $L_p^*$  为标准灰板实测值,  $\rho_p$  为参考板的反射率.

在光谱测量的同时采集表层水样, 低温冷藏带回实验室测量悬浮物的浓度. 悬浮物浓度采用常规的干燥、烘烧、称重法(GB 11901-89 标准)测定.

### 1.3 数据的处理

利用 MATLAB R2009b 软件对原始高光谱数据进行小波变换处理, 利用光谱专用分析软件 Unscrambler 9.8 对压缩后的新变量进行分析, 建立偏最小二乘(PLS)校正模型, 对其它数据的处理(包括计算最大值、最小值、均值以及数据的拟合等)利用 SPSS 11.5 完成.

研究中用于评价模型的指标均方根误差(RMSE)和平均相对误差(MAPE)分别利用式(2)、(3)计算得到:

$$RMSE = \sqrt{\sum_{i=1}^n (y_i - y'_i)^2 / n} \quad (2)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - y'_i}{y_i} \right| \quad (3)$$

式中,  $y_i$ 、 $y'_i$  分别为实测悬浮物浓度和预测悬浮物浓度,  $n$  为样本数目.

## 2 偏最小二乘回归分析方法

### 2.1 偏最小二乘建模原理

偏最小二乘法是一种新型的数据分析方法, 它将多元线性回归分析、主成分分析与典型相关分析有机结合起来, 其建模原理也是建立在这 3 种分析方法之上的, 通过从自变量集合中提取若干相互独立的主成分来建立与因变量之间的关系. 与主成分分析不同的是, 偏最小二乘法所提取的成分既能很好地概括自变量系统的信息, 又能够最好地解释因变量. 具体建模原理<sup>[19]</sup>为: 设有  $p$  个自变量  $X = \{x_1, x_2, \dots, x_p\}$ 、 $q$  个因变量  $Y = \{y_1, y_2, \dots, y_q\}$  和  $n$  个样本点, 分别在  $X$  与  $Y$  中提取出主成分  $t_1$  和  $u_1$ , 要求  $t_1$  和  $u_1$  应尽可能大的携带它们各自数据表中的变异信息, 同时  $t_1$  和  $u_1$  的相关程度能够达到最大, 使得  $t_1$  和  $u_1$  应尽可能好的代表数据表  $X$  与  $Y$ , 同时自变量成分  $t_1$  对因变量成分  $u_1$  又具有最强的解释能力. 在第一个成分  $t_1$  和  $u_1$  被提取后, 分别实施  $X$  对  $t_1$  的回归及  $Y$  对  $t_1$  的回归, 如果回归方程已经达到满意的程度, 则算法终止, 否则, 利用  $X$  被  $t_1$  解释后的残余信息以及  $Y$  被  $t_1$  解释后的残余信息进行第二轮的成分提取, 如此往复, 直到达到一个较为满意的精度为止. 若对  $X$  提取了  $m$  个成分  $t_1, t_2, \dots, t_m$ , 偏最小二乘法将通过实施  $y_k$  对  $t_1, t_2, \dots, t_m$  的回归, 然后再表达成  $y_k$  对原变量  $x_1, x_2, \dots, x_m$  的回归方程( $k = 1, 2, \dots, q$ ).

### 2.2 数据的标准化处理

在利用偏最小二乘方法进行建模前首先要对数据进行标准化处理, 标准化处理的目的是使样本点的分布结构有利于计算, 并尽量避免数据的舍入误差. 将自变量矩阵  $X$  和因变量矩阵  $Y$  标准化后分别得到矩阵  $E_0$  和  $F_0$ , 其中,  $E_0 = (x_{ij}^*)_{n \times p}$ ,  $F_0 = [(y_i - \bar{y})/s_y]_{n \times 1}$ ,  $x_{ij}^* = (x_{ij} - \bar{x}_j)/s_j$  ( $i = 1, 2, \dots, n; j = 1, 2, \dots, p$ ),  $\bar{x}$  和  $s_j$  分别是  $x_j$  的均值和标准差,  $\bar{y}$  和  $s_y$  分别是  $Y$  的均值和标准差.

### 2.3 交叉有效性检验

在 PLS 方法中, 主成分个数的提取是通过分析交叉有效性来确定的, 假设  $y_i$  为样本  $i$  的观测值,  $\hat{y}_{hi}$  为使用全部样本并提取  $h$  个成分回归建模后第  $i$  个样本的拟合值,  $\hat{y}_{h-i}$  为在建模时剔除样本  $i$  并提取  $h$  个成分回归建模后第  $i$  个样本的拟合值, 则:

$$SS_h = \sum_{i=1}^n (y_i - \hat{y}_{hi})^2, PRESS_h = \sum_{i=1}^n (y_i - \hat{y}_{h-i})^2 \quad (4)$$

式中,  $SS_h$  和  $PRESS_h$  分别为  $y$  在两种模型下的预测误差平方和, 若  $PRESS_h/SS_{h-1} \leq 0.95^2$  则表明加入主成分能够改善模型精度, 否则不能. 最终模型的评价可用  $R^2(\text{pred})$  和一般建模方法中的决定系数  $R^2$  来实现:

$$R^2 = 1 - \frac{SS_h}{\sum_{i=1}^n (y_i - \bar{y})^2}, R^2(\text{pred}) = 1 - \frac{PRESS_h}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

其中,  $R^2(\text{pred})$  可避免模型出现过拟合现象, 即当最佳模型确定后, 增加主成分个数反而会使  $R^2(\text{pred})$  降低.

## 2.4 成分解释能力分析

在偏最小二乘回归计算的过程中, 所提取的自变量成分一方面要尽可能多地代表自变量的变异信息, 另一方面又尽可能与因变量相联系, 解释因变量中的信息, 因此主成分累积解释能力是确定最佳主成分个数和评价模型精度的指标之一.

## 3 结果与分析

### 3.1 巢湖水体反射光谱特征

对计算得到的遥感反射率光谱曲线进行筛选, 最终选取遥感反射率和悬浮物浓度对应较好的 29 个样点的数据进行模型的建立及验证. 由于实测的水体反射光谱在 850nm 后迅速下降, 数据的信噪比降低, 并且含有较大的噪声, 因此本研究只对 400–850nm 范围内的光谱数据进行分析. 巢湖 29 个样点在 400–850nm

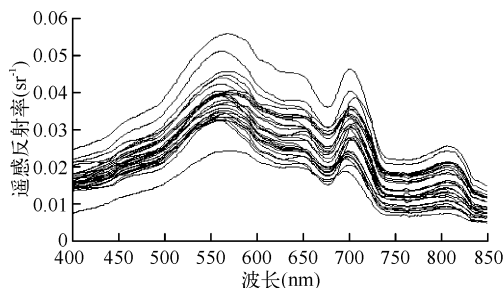


图2 巢湖水体反射光谱曲线

Fig. 2 Remote sensing reflectance in Lake Chaohu

间的反射光谱曲线具有典型的内陆二类水体的光谱特征: 在蓝光波段 (400–500nm), 由于叶绿素 *a* 和黄色物质的强吸收作用, 使得遥感反射率在该波段范围内较低; 550–580nm 范围内出现一反射峰, 主要是由于叶绿素和胡萝卜素在该波段范围内的弱吸收及细胞的散射作用形成的; 620nm 左右由于藻青蛋白的吸收形成一小的反射谷; 675nm 左右由于叶绿素的强烈吸收形成吸收谷; 700nm 左右存在一明显的叶绿素荧光峰, 其存在与否被认为是判定水体是否含藻类叶绿素的依据<sup>[20]</sup>; 700nm 以后由于纯水的强吸收占据主导作用使得反射率迅速下降, 直到 808nm 左右又出现一反射峰, 该反射峰可能是由于悬浮物的散射形成的 (图 2).

### 3.2 原始数据的小波变换

水体属于弱反射体, 传感器所接收到的水体的信号较弱, 易受噪声和背景的干扰. 此外, 本研究中所采用的水体反射光谱的波长范围为 450–850nm, 共 451 个波段. 有研究<sup>[21-23]</sup>认为, PLS 建模时波长数目如果远远多于样本的数目, 会使建立的模型具有随机性, 与直接采用全谱进行 PLS 建模相比, 利用经过小波变换后的数据进行 PLS 建模不仅可以有效地压缩原始光谱数据和消除数据噪声, 而且能够降低所建模型的随机性, 减少变量数目, 从而大大提高运算速度, 并能获得更高的预测精度. 因此, 本研究首先对原始光谱数据进行小波变换, 然后采用小波变换后的数据进行 PLS 模型的建立.

小波变换是从傅立叶变换发展起来的一种新的信号处理方法, 信号经小波分解之后, 噪音信号会在高频部分出现, 而对于有用的信号分量大部分在低频部分出现, 据此可以将高频部分小波系数中低于某一阈值的系数去除, 而对其余系数重新编码, 只保留编码后的小波系数, 这样可大大减少数据贮存量, 达到信号压缩的目的. 本研究中采用正交小波基函数, 对原始数据进行多尺度的小波分解来对原始光谱数据进行压缩, 考虑到不同的正交小波基函数 (Db1、Db2、Db3、Db4) 与分解尺度 (3、4、5、6) 对数据压缩的影响, 以信息保留量、压缩后的特征变量数目和所有样点总的均方根误差 (*TRMSE*) 三个指标来确定小波变换所使用的正交小波基函数与分解尺度. *TRMSE* 计算公式为:

$$TRMSE = \sum_{i=1}^m \sqrt{(\sum_{j=1}^N (y_m(j) - y_w(j))^2)/N}$$

(6)

式中,  $m$  为总的样点数目,  $N$  为总的波段数目,  $y_m(j)$ 、 $y_w(j)$  分别表示第  $j$  波段处实测遥感反射率和经小波变换再重构后的遥感反射率。

在 MATLAB 软件中分别采用 4 种不同的正交小波基函数和 4 种分解尺度对原始光谱数据进行小波变换处理,得到原始数据的小波系数,并删掉小波系数中足够小而被认为不代表有用信息的系数,然后再用保留的系数对光谱数据进行重构,对重构后的光谱数据和原始光谱数据的三个指标进行统计。

不同的正交小波基函数及不同的分解尺度对数据的压缩影响不同,总体上来看,分解尺度越大,其信息的保留量越少,也就是其损失的光谱信息越多,  $TRMSE$  越大,但能用来代替原始光谱数据的特征变量数目越少,即原始数据的压缩程度越大;在 4 种正交小波基函数中,采用 Db4 小波基函数进行小波变换后,同一分解尺度下所保留的信息量最多,  $TRMSE$  也小于其它 3 种小波基函数,但用于代替原始光谱数据的特征变量数目大于其它 3 种小波基函数(表 1)。水体本身属于弱反射体,其遥感反射率相对较小,因此在对原始光谱数据进行小波变换时应在保留原始光谱信息的基础上对数据进行最大程度的压缩,以提高最终模型的反演精度。通过对表 1 中不同小波基函数及不同的分解尺度所对应的信息保留量、特征变量数目和  $TRMSE$  三个指标进行综合分析,最终选取小波基函数为 Db4,分解尺度为 4,经过该小波变换后的数据保留了 99.99% 的原始光谱信息,几乎在没有损失光谱信息的基础上对原始光谱数据进行了最大程度的压缩,使原来的 451 个波段数据减少到用 34 个特征变量进行替代。对原始光谱数据进行小波变换(小波基函数为 Db4,分解尺度为 4)后,再进行小波重构后的光谱数据见图 3,通过与图 2 的对比可以看出,重构后的光谱数据曲线比原始光谱数据更为平滑,有效地去除了个别波段的噪声,在光谱形状上,重构后的光谱数据与原始光谱数据一致,光谱数据的特征峰、特征谷波段并没有因数据压缩而减弱或消失,仍保持与原始数据一致,因此可以利用小波变换后的特征变量代替原有的波段进行建模而不影响最终模型的反演结果及精度。

3.3 PLS 模型的建立及评价

3.3.1 最佳主成分个数的确定和成分解释能力分析  
从 29 个样点中随机选取 20 个样点进行建模,剩余的 9 个样点进行模型的验证。利用原始光谱数据经小波

表 1 原始数据与重构数据指标统计

Tab. 1 Index statistics between raw data and reconstructive data

小波基函数	分解(尺度)	信息保留量(%)	特征变量数目	TRMSE
Db1	3	99.96	57	0.0156
	4	99.84	29	0.0309
	5	99.56	15	0.0518
	6	99.07	8	0.0760
Db2	3	100.00	59	0.0048
	4	99.97	31	0.0137
	5	99.80	17	0.0349
	6	99.20	10	0.0696
Db3	3	100.00	60	0.0029
	4	99.98	32	0.0106
	5	99.68	18	0.0443
	6	99.02	11	0.0799
Db4	3	100.00	62	0.0025
	4	99.99	34	0.0082
	5	99.87	20	0.0275
	6	99.40	13	0.0658

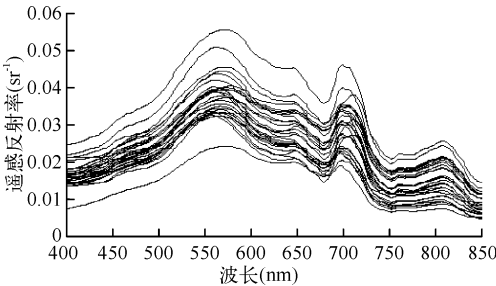


图 3 小波重构后的光谱数据

Fig. 3 Spectral data after wavelet reconstruction

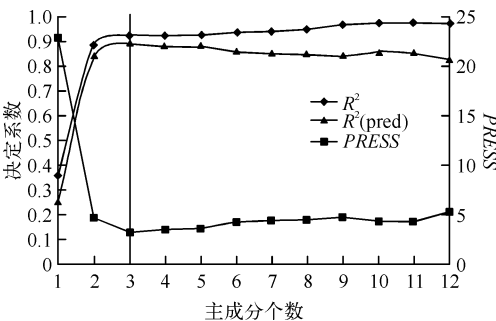


图 4 模型最佳主成分个数的确定

Fig. 4 Determination of the number of principal component of the model

变换后的特征变量和悬浮物浓度首先经标准化处理然后再依照上述原理建立 PLS 模型, 经交互验证法判断, 确定最佳主成分个数. PLS 模型中最佳主成分的确定既要保证成分对系统解释能力最强, 又要克服变量之间的多重共线性关系. 本研究中采用模型的  $R^2$ 、 $R^2(\text{pred})$  以及  $PRESS$  来确定模型的最佳主成分个数. 随着主成分个数的增加, 模型的  $R^2$  不断增大, 而  $R^2(\text{pred})$  先增大, 当主成分个数为 3 时达到最大值, 而后逐渐减小, 同时  $PRESS$  也在主成分个数为 3 时达到最小值. 因此确定模型的最佳主成分个数为 3, 此时模型的  $R^2$  为 0.93,  $R^2(\text{pred})$  为 0.89,  $PRESS$  为 3.29 (图 4).

当模型的主成分个数取 3 时, 对自变量  $X$  来说, 第一主成分的解释能力最强, 反映了其 80.87% 的信息, 其次为第二主成分和第三主成分, 而对因变量  $Y$  来说, 第二主成分对其解释能力最强, 为 52.62%, 第一主成分也有部分解释能力, 第三主成分的解释能力则相对较小; 从累积解释能力来看, 3 个主成分累积解释了 98.60% 的自变量信息和 92.37% 的因变量信息, 均达到了较高的解释水平, 说明利用偏最小二乘法拟合出的回归模型能概括原始数据所携带的绝大部分信息 (表 2).

表 2 各主成分对  $X$ 、 $Y$  的解释能力

Tab. 2 The explanation ability of each principal component

主成分	对 $X$ 解释能力 (%)	对 $Y$ 解释能力 (%)	对 $X$ 累积解释能力 (%)	对 $Y$ 累积解释能力 (%)
PC1	80.87	36.10	80.87	36.10
PC2	14.13	52.62	95.00	88.72
PC3	3.60	3.65	98.60	92.37

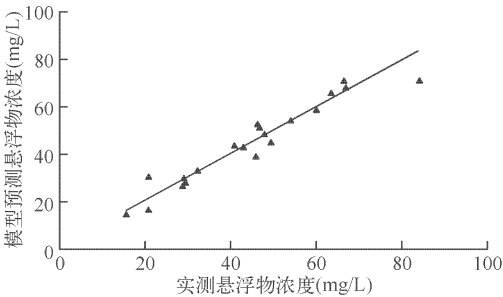


图 5 实测悬浮物浓度与模型预测悬浮物浓度关系  
Fig. 5 The relationship of TSM between measured data and model predicted data

表 3 9 个验证样本悬浮物实测值与预测值比较  
Tab. 3 Comparison of measured and predicted suspended matter concentration of 9 validation samples

样点号	实测浓度 (mg/L)	预测浓度 (mg/L)	相对误差 (%)
20	17.80	14.66	17.67
27	28.55	28.62	0.23
31	29.80	32.32	8.44
7	35.00	41.29	17.97
8	45.13	48.96	8.48
9	46.87	52.76	12.57
17	50.60	56.24	11.15
13	67.53	59.17	12.39
32	63.55	60.20	5.28

3.3.2 PLS 模型的建立 根据确定的最佳主成分的个数进行 PLS 模型的建立, 结果表明所建立的模型具有较高的精度, 其  $R^2$  为 0.93, 20 个建模样本实测悬浮物浓度与模型预测悬浮物浓度的最大相对误差为 45.62%, 最小相对误差 0.19%, 90% 的样点的相对误差在 15% 以内, 20 个样点的  $MAPE$  为 9.02%,  $RMSE$  为 4.77mg/L (表 4). 实测值与预测值具有很好的对应关系, 基本上均匀分布在 1:1 线的两侧 (图 5), 这也说明所建立的模型具有相对较高的精度.

模型精度较高并不意味着模型具有较好的预测能力. 为了验证模型的预测能力及评价模型的稳定性, 将未参与建模的 9 个样点的数据代入上述建立的 PLS 模型中, 结果显示, 9 个验证样本中, 实测悬浮物浓度与预测悬浮物浓度的相关系数达 0.95, 最大相对误差为 17.97%, 最小相对误差为 0.23% (表 3),  $MAPE$  为 10.46%,  $RMSE$  为 4.92mg/L, 略高于建模样本的  $MAPE$  和  $RMSE$  (表 4). 总体上看, 所建立的 PLS 模型可以较好的反演悬浮物浓度, 具有较好的预测能力和稳定性, 适用于巢湖水体悬浮物浓度的高光谱反演.

3.3.3 各波段回归系数分析 PLS 模型中各波段的回归系数能够反映出各变量对于模型的重要性. 不同的波段对模型的影响程度不同, 400–610nm 范围内各波段的回归系数为负, 而 610nm 之后各波段的回归系数为正. 从各波段回归系数的数值大小来看, 400–580nm 范围内各波段的回归系数相差不大, 基本上在 -0.05 左右变动, 这表明该波长范围内各波段对最终模型的贡献比

较均衡;580-680nm 范围内,各波段的回归系数相对较小,其原因可能是由于太湖水体中叶绿素浓度相对较高,且该波段范围内的遥感反射率对叶绿素浓度的变化比较敏感,从而对悬浮物浓度的估算产生了一定的影响;680nm 以后各波段的回归系数逐渐增大,并出现了两个相对高值区:705-734nm 和 789-824nm,这两个波长范围内各波段的回归系数都超过了 0.1. 总体来看,可见光波段范围内各波段的回归系数相对较小,且大部分波段的回归系数相差不大,对最终模型的贡献相对较小并且较为均衡,而近红外波段范围内各波段的回归系数相对较大且变化也相对较大,对最终模型的贡献相对较大但不均衡(图 6). 该研究结果与徐京萍等<sup>[24]</sup>对石头口门水库的研究结果有所不同,其原因主要是由于不同研究区域水体组分及其浓度存在较大的差异,进而对遥感反射率产生不同的影响.

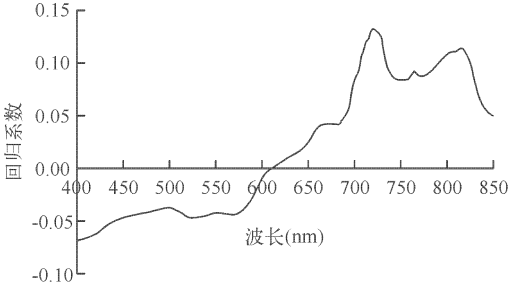


图 6 各波段回归系数

Fig. 6 Regression coefficient of each wavelength

3. 4 模型的对比和评价

本研究中分别建立了常用的用于悬浮物浓度反演的单波段模型、波段比值模型及光谱一阶微分模型. 通过对各种模型的对比分析来评价 PLS 模型的反演精度及模型的稳定性和适用性(表 4).

表 4 不同建模方法比较\*  
Tab. 4 Comparison of different models

建模方法	自变量	拟合方程	建模样本			验证样本		
			$R^2$	MAPE (%)	RMSE (mg/L)	$R^2$	MAPE (%)	RMSE (mg/L)
PLS			0.93	9.02	4.77	0.90	10.46	4.92
单波段	$R_{817}$	$Y = -195087X^2 + 8811.8X - 37.329$	0.66	19.70	10.11	0.81	17.17	7.53
一阶微分	$R_{(832) '}$	$Y = 16.944e^{-3262X}$	0.83	15.90	8.81	0.63	17.89	11.60
波段比值	$R_{693}/R_{579}$	$Y = 218.07X - 119.54$	0.84	14.49	6.92	0.92	12.71	4.97

\* X 为单波段的遥感反射率或两个波段遥感反射率的比值,Y 为悬浮物浓度.

前人研究结果表明,悬浮物浓度的反演一般采用绿光、红光波段或近红外波段<sup>[25-28]</sup>. 李素菊等<sup>[29]</sup>通过分析巢湖各波段反射率与悬浮物浓度的相关关系,发现 819nm 处的相关系数最大,并以此波段的遥感反射率为自变量,悬浮物浓度为因变量建立了线性模型;Ahn 等<sup>[6]</sup>成功地利用 625nm 波长估算了我国黄海外海的悬浮物浓度. 据此,本研究中对各波段的遥感反射率与悬浮物浓度求相关系数,发现 817nm 处的相关系数最高,为 0.85,与李素菊等<sup>[29]</sup>的研究结果相近. 以 817nm 波段处的遥感反射率为自变量,悬浮物浓度为因变量,分别建立线性、指数、乘幂、二次函数模型,发现二次函数模型精度最高,因此以二次函数模型作为悬浮物浓度反演的单波段模型来拟合方程.

光谱一阶微分是处理高光谱数据的一种重要的方法,对光谱进行一阶微分可以去除部分线性或接近线性的背景噪声对目标光谱的影响<sup>[30]</sup>. 已有的研究结果表明利用光谱一阶微分方法可以有效地估算悬浮物浓度<sup>[29,31-32]</sup>. ASD 野外光谱辐射计采集的是离散型数据,因此光谱数据一阶微分可用以下公式近似计算:

$$R(\lambda_i)' = \frac{R(\lambda_{i+1}) - R(\lambda_{i-1})}{\lambda_{i+1} - \lambda_{i-1}} \tag{7}$$

式中, $\lambda_{i+1}$ 、 $\lambda_i$ 、 $\lambda_{i-1}$ 为相邻波长, $R(\lambda_i)$ 为波长  $\lambda_i$  的一阶微分反射光谱. 通过计算各波段的光谱一阶微分值与悬浮物浓度的相关系数,发现 832nm 处的相关系数最大,最大负相关系数为 -0.87,并建立了以 832nm 处的光谱一阶微分值为自变量,悬浮物浓度为因变量的线性、指数、二次函数模型. 通过比较发现指数模型精度最高,并将指数模型作为悬浮物浓度反演的一阶微分模型.

Pulliainen 等<sup>[33]</sup>研究认为对水体的不同波段做比值处理可以部分的消除大气影响,也可以消除在空间上和时间上水面粗糙度变化的干扰. 本研究中分别计算两个不同波段遥感反射率的比值与悬浮物浓度的相关关系,发现 693nm 与 579nm 处的遥感反射率的比值 ( $R_{693}/R_{579}$ ) 与悬浮物浓度相关性最大,相关系数为 0.93,由此分别建立了以  $R_{693}/R_{579}$  为自变量,以悬浮物浓度为因变量的线性、指数、乘幂、二次函数模型,结果发现四种模型中线性模型的精度最高,因此,将线性模型作为悬浮物浓度反演的波段比值模型.

4 种模型中除单波段模型的建模精度稍低外,其余 3 种模型的精度均较好,相比之下,PLS 模型具有更高的精度,其建模样本的  $R^2$  为 0.93(表 4),高于一阶微分模型和波段比值模型,模型的 *MAPE* 和 *RMSE* 也远低于另外两种模型. 通过利用 9 个未知样本对所建立的 4 种模型进行验证,发现虽然一阶微分模型具有相对较高的建模精度,但其稳定性最差,其模型验证的精度甚至低于建模精度较低的单波段模型. 这是因为一阶微分值受数据的影响较大,对高频信息敏感,模型稳定性较差. PLS 模型和波段比值模型都具有相对较高的预测精度,波段比值模型验证样本的  $R^2$  稍高于 PLS 模型,其验证样本的 *MAPE* 和 *RMSE* 也高于 PLS 模型. 综合分析上述 4 种模型得出,单波段模型精度较差,一阶微分模型虽然建模精度相对较高,但其稳定性较差,PLS 模型和波段比值模型都具有相对较高的精度,但相比之下,PLS 模型比波段比值模型具有更高的建模精度和模型的稳定性,因此,无论从建模样本的精度还是从验证样本的误差和稳定性分析,PLS 模型均优于传统的经验模型,能够充分利用光谱的有用信息,适合于巢湖悬浮物浓度的高光谱反演.

#### 4 参考文献

- [1] 周 艺,周伟奇,王世新等. 遥感技术在内陆水体水质监测中的应用. 水科学进展,2004,15(3):312-317.
- [2] 疏小舟,尹 球,匡定波等. 内陆水体藻类叶绿素浓度与反射光谱特征的关系. 遥感学报,2000,4(1):41-45.
- [3] 汪小钦,王钦敏,郭群勇等. 遥感在悬浮物浓度提取中的应用——以福建闽江口为例. 遥感学报,2003,7(1):54-57.
- [4] Gin KY, Koh ST. Spectral irradiance profiles of suspended marine clay for the estimation of suspended sediment concentration in tropical waters. *International Journal of Remote Sensing*, 2003,24(16):3235-3245.
- [5] Härmä P, Vepsäläinen J, Hannonen T *et al.* Detection of water quality using simulated satellite data and semi-empirical algorithms in Finland. *Science of the Total Environment*, 2001,268:107-121.
- [6] Ahn YH, Moon JE, Gallegos S. Development of suspended particulate matter algorithms for ocean color remote sensing. *Korean Journal of Remote Sensing*, 2001,17(4):285-295.
- [7] 刘堂友,匡定波,尹 球. 湖泊藻类叶绿素 a 和悬浮物浓度高光谱定量遥感模型研究. 红外与毫米波学报,2004,23(1):11-15.
- [8] 张 芸,张 鹰,王晶晶. 悬浮物浓度遥感反演模型研究. 海洋科学,2008,32(5):32-37.
- [9] 孙德勇,李云梅,王 桥等. 基于实测高光谱的太湖水体悬浮物浓度遥感估算研究. 红外与毫米波学报,2009,28(2):124-128.
- [10] Mishra AK. Retrieval of suspended sediment in the estuarine waters using IRS-1C WiFS data. *International Journal of Applied Earth Observation and Geoinformation*, 2004,6:83-95.
- [11] 乐成峰,李云梅,查 勇等. 太湖悬浮物对水体生态环境的影响及其高光谱反演. 环境科学学报,2008,28(10):2148-2155.
- [12] Doxaran D, Froidefond JM, Castaing P. Remote-sensing reflectance of turbid sediment-dominated waters. Reduction of sediment type variations and changing illumination conditions effects by use of reflectance ratios. *Applied Optics*, 2003,42(15):2623-2634.
- [13] Teodoro AC, Marcal ARS, Veloso-Gomes F. Correlation analysis of water wave reflectance and local TSM concentrations in the breaking zone with remote sensing techniques. *Journal of Coastal Research*, 2007,23(6):1491-1497.
- [14] 徐京萍,张 柏,宋开山等. 近红外波段二类水体悬浮物生物光学反演模型研究. 光谱学与光谱分析,2008,28(10):2273-2277.
- [15] 张 兵,申 茜,李俊生等. 太湖水体 3 种典型水质参数的高光谱遥感反演. 湖泊科学,2009,21(2):182-192.
- [16] 张恒喜,郭基联,朱家元等. 小样本多元数据分析方法及应用. 西安:西北工业大学出版社,2002.
- [17] 唐军武. 海洋光学特性模拟与遥感模型[学位论文]. 北京:中国科学院遥感应用研究所,1999:107-110.



- [18] 唐军武, 田国良, 汪小勇等. 水体光谱测量与分析 I: 水面以上测量法. 遥感学报, 2004, **8**(1): 37-44.
- [19] 张正健, 刘志红, 郭艳芬等. 偏最小二乘在遥感监测西藏草地生物量上的应用. 草地学报, 2009, **17**(6): 735-739.
- [20] 焦洪波, 查勇, 李云梅等. 基于高光谱遥感反射比的太湖水体叶绿素 a 含量估算模型. 遥感学报, 2006, **10**(2): 242-248.
- [21] 楮小立, 田高友, 袁洪福等. 小波变换结合多维偏最小二乘方法用于近红外光谱定量分析. 分析化学, 2006, **34**: 175-178.
- [22] 司圣柱, 司 娟. 小波变换—偏最小二乘法用于三种食用色素的可见分光光度法的同时测定. 分析仪器, 2007, **2**: 48-51.
- [23] 王芳, 陈 达, 邵学广. 小波变换和偏最小二乘法在烟草常规成分预测中的应用. 烟草科技, 2004, **3**: 31-34.
- [24] 徐京萍, 张 柏, 宋开山等. 用偏最小二乘法提取石头口门水库水色信息. 中国科学院研究生院学报, 2007, **24**(6): 814-819.
- [25] Dekker AG. Analytical algorithms for lake water TSM estimation for retrospective analyses of TM and SPOT sensor data. *International Journal of Remote Sensing*, 2002, **23**(1): 15-35.
- [26] Doxaran D, Froidefond J, Lavender S *et al.* Spectral signature of highly turbid waters: application with SPOT data to quantify suspended particulate matter concentrations. *Remote Sensing of Environment*, 2002, **81**(1): 149-161.
- [27] Binding CE, Bowers DG, Mitchelson-Jacob EG. Estimating suspended sediment concentrations from ocean color measurements in moderately turbid waters: The impact of variable particle scattering properties. *Remote Sensing of Environment*, 2005, **94**(3): 373-383.
- [28] Gitelson A, Garbuzov G, Szilagyi F *et al.* Quantitative remote sensing methods for real-time monitoring of inland waters quality. *International Journal of Remote Sensing*, 1993, **14**(7): 1269-1295.
- [29] 李素菊, 王学军. 巢湖水体悬浮物含量与光谱反射率的关系. 城市环境与城市生态, 2003, **16**(6): 66-68.
- [30] 浦瑞良, 宫 鹏. 高光谱遥感及其应用. 北京: 高等教育出版社, 2000: 53.
- [31] 徐京萍, 张 柏, 蔺 钰等. 结合高光谱数据反演吉林石头口门水库悬浮物含量和透明度. 湖泊科学, 2007, **19**(3): 269-274.
- [32] 吕 恒, 李新国, 江 南. 基于反射率光谱和模拟 MERIS 数据的太湖悬浮物遥感定量模型. 湖泊科学, 2005, **17**(2): 104-109.
- [33] Pulliainen J, Kallio K, Eloheimo K. A semi-operative approach to lake water quality retrieval from remote sensing data. *Science of the Total Environment*, 2001, **268**: 79-93.