doi:10.6041/j.issn.1000-1298.2019.11.023

基于高光谱定量反演模型的污水综合水质评价

陈俊英[1,2]　邢　正[1]　张智韬[1,2]　劳聪聪[1]　栗现文[1,2]　王海峰[1]

(1.西北农林科技大学水利与建筑工程学院,陕西杨凌 712100;
2.西北农林科技大学中国旱区节水农业研究院,陕西杨凌 712100)

摘要:为改善高光谱遥感对污水水质信息状况定量反演模型的预测评价效果,以陕西某污水处理厂采集的污水样品为研究对象,采用主成分分析法(Principal component analysis,PCA)对污水水质进行综合评价,获取水质评价的综合评价因子,同时利用 ASD FieldSpec 3 型高光谱仪获取污水的原始光谱,经过数据预处理和不同数学变换后,共获取了 4 种光谱指标:平滑后光谱反射率(SG)、倒数之对数(LR)、标准正态化(SNV)和去包络线(CR)。分别采用偏最小二乘回归法(Partial least squares regression,PLSR)、逐步回归法(Stepwise regression,SR)、极限学习机法(Extreme learning machine,ELM)构建了基于水质综合评价因子的高光谱水质反演模型,并对反演结果进行精度验证与比较。结果表明,本组水样的平滑后光谱数据和经过标准正态化变换的光谱数据建模具有较好的建模效果,其建模的预测 RPD 均在 2.5 以上;在 3 种模型中,PLSR 模型和 ELM 模型均具备很好的建模预测效果;逐步回归法的建模效果较 PLSR 模型和 ELM 模型有所下降,但是其 SG - SR、SNV - SR 模型的 R_c^2 均在 0.8 以上、R_p^2 均在 0.85 以上,RPD 均在 3.0 以上,证明其仍拥有很好的反演预测效果,且进行了特征波段的优选,实现了对模型的优化;SNV - SR - ELM($R_c^2 = 0.956$,$R_p^2 = 0.954$,$RMSE = 0.500$,$RPD = 4.651$)为最佳模型,SNV - SR - ELM 模型的建立为高光谱反演水质模型的优化、污水水质的快速监测和综合评价提供了途径。

关键词:水质综合评价;高光谱;预处理;偏最小二乘回归;极限学习机

中图分类号:X524　　**文献标识码**:A　　**文章编号**:1000-1298(2019)11-0200-10

Comprehensive Evaluation of Waste Water Quality Based on Quantitative Inversion Model Hyperspectral Technology

CHEN Junying[1,2]　XING Zheng[1]　ZHANG Zhitao[1,2]　LAO Congcong[1]　LI Xianwen[1,2]　WANG Haifeng[1]

(1. *College of Water Resources and Architectural Engineering*, *Northwest A&F University*, *Yangling*, *Shaanxi* 712100, *China*
2. *Institute of Water-saving Agriculture in Arid Areas of China*, *Northwest A&F University*, *Yangling*, *Shaanxi* 712100, *China*)

Abstract: A comprehensive inversion of the water quality information of sewage water was realized through the combination of hyperspectral technology and water quality comprehensive evaluation method. Taking the sewage sample collected by a sewage treatment plant in Shaanxi as the research object, principal component analysis (PCA) was used to comprehensively evaluate the sewage water quality to obtain a comprehensive evaluation factor for water quality. At the same time, the original wastewater spectrum was obtained by the ASD FieldSpec 3 hyperspectral instrument. After data preprocessing and different mathematical transformations, four spectral indices were obtained: spectral reflectance (SG), reciprocal logarithm (LR), standard normal variable (SNV) and continuum removed (CR). Based on partial least squares regression (PLSR), stepwise regression (SR) and extreme learning machine (ELM), a hyperspectral model of inversion water quality comprehensive evaluation factor was constructed. The results showed that the original spectral data of this group of water samples and the spectral data modeling by standard normalization transformation had good modeling results, and the prediction effect RPD of the model was above 2.5. Among the three models, the PLSR model and the ELM model had good modeling prediction effects, while stepwise regression modeling results were declined compared with PLSR model and ELM model, the R_c^2 and R_p^2 of the REF - SR and SNV - SR models were all above 0.8 and 0.85, and the RPD was above 3.0, which still had a very good inversion

收稿日期:2019-04-04　修回日期:2019-05-10

基金项目:国家重点研发计划项目(2017YFC0403302)和国家自然科学基金项目(41502225、51979234)

作者简介:陈俊英(1975—),女,副教授,博士,主要从事节水农业和水土资源高效利用研究,E-mail:cjyrose@126.com

prediction effect, and it achieved the optimization of the model and the optimization of the characteristic band, and SNV－SR－ELM ($R_c^2 = 0.956$, $R_p^2 = 0.954$, $RMSE = 0.500$, $RPD = 4.651$) was the best model. The establishment of SNV－SR－ELM model provided a way for the optimization of hyperspectral inversion water quality model and the rapid evaluation of sewage water quality.

Key words: comprehensive evaluation of water quality; hyperspectral; pretreatment; partial least squares regression; extreme learning machine

0　引言

近年来,利用高光谱遥感技术评价和监测水资源水质信息状况方面的研究愈发深入[1]。而应用高光谱技术检测水体水质的关键在于水质的综合评价和光谱数据与水质参数间数学模型的建立。对于光谱数据和水质参数间数学模型的建立,国内外学者对影响水体质量的几个主要参数指标的遥感估算进行了大量的研究,如化学需氧量(Chemical oxygen demand, COD)、浊度、总磷、生物耗氧量(Biological oxygen demand, BOD)、总氮等。YE 等[2]应用 UVE－SPA－LS－SVM 的方法实现了对 COD 的建模预测;吕航等[3]利用 HJ－1A 卫星 HSI 高光谱遥感数据,建立了 9 个水质参数与水体光谱反射率之间的估算模型;曹引等[4]建立了水体浊度的高光谱定量反演模型,为水体浊度大面积遥感监测的业务化管理提供了技术支持;BANSOD 等[5]通过高光谱的图像数据,对恒河的水质参数进行了反演评价。在对于水质单一参数的高光谱数据反演模型建立上,目前已经达到可以定量的效果[6－9]。但影响水体质量的水质因子数目众多,利用单独的某项水质参数来描述水质的信息状况不够全面,因此往往会对水质进行综合评价。对于水质的综合评价,目前的研究也较为成熟。马小雪等[10]利用主成分分析法对温瑞塘河流域多项水质参数进行时空分异特征分析和潜在污染源的识别;徐国宾等[11]利用模糊标识指数对水质达标状况、水质类别和主要污染因子进行综合评价。这些水质综合评价方法通过已有的水质参数资料能够很好地表征水质的信息状况,但在实时性方面存在不足。因此,需建立高光谱结合水质综合评价方法反演水质模型,充分发挥高光谱遥感的实时性、大范围性与水质综合评价方法的全面性、准确性的优势。目前对于高光谱结合水质综合评价方法反演水质模型的研究较少。

本文将一组来自污水处理厂各处理工艺处的水体样品分成两份,一份经由室内理化试验,检测各项水质参数,并利用主成分分析对水体水质进行综合评价,得到水质综合评价因子;同时对另一份水体样品进行高光谱数据的采集,将采集到的光谱数据进行不同的预处理,采用偏最小二乘法、逐步回归法和极限学习机法对光谱数据和水质综合评价因子进行建模预测以及验证。比较各预处理方法及对应的建模方法的验证结果,选出更适合用于水质综合评价高光谱反演的数据预处理及对应建模方法,为建立高光谱结合水质综合评价方法反演水质模型,实现对水质信息状况的大范围实时监测提供可行的路径。

1　试验材料与方法

1.1　样本采集及样本水质指标的测定

试验用水水样取自某生活污水处理厂,取水位置分别为生活污水处理的不同工艺处,即进水口、厌氧池、好氧池、沉淀池、出水口,对照的水样为纯净水。各水样的各项水质参数经由室内理化试验测定,结果见表 1(部分)。

1.2　光谱测定

污水样品采用 ASD Field Spec 3 型地物光谱仪测量高光谱数据。光谱仪波长范围为 350～2 500 nm,采样间隔为 1.4 nm(350～1 000 nm)和 2 nm(1 000～2 500 nm),重采样间隔为 1 nm。光谱测量在暗室中进行,光源为 DH－2000 型氘卤钨灯光源[12]。

1.3　光谱数据预处理

本试验中 87 个样品获得的光谱波段为 350～2 500 nm。由于试验条件以及其他因素的影响,测量的光谱中可能包含了一些冗余信息以及噪声,因此需要对获得的光谱波段进行选择以提高建模的准确度。

由图 1 可看到,在 350～400 nm、2 300～2 500 nm 波段范围由于处于边缘噪声较大,不适用于建模。而大于 2 000 nm 波段,反射率很小,可利用的信息很少,难以找出不同样本光谱图的差别。故本研究选用 400～2 000 nm 的光谱波段。在建模前需要对光谱数据进行一定的预处理以削弱由测试环境及其他干扰因素导致的影响,提高数据信噪比。本文采用预处理方法有 Savitzky－Golay(SG)平滑、标准正态化(SNV)、去包络线(CR)和倒数之对数(LR)预处理等。

1.3.1　Savitzky－Golay 平滑处理

平滑滤波是光谱分析中常用的预处理方法之

表 1 主要水质参数

Tab. 1 Main water quality parameters

参数	进水口	厌氧池	好氧池	沉淀池	出水口
NH_4^+-N 含量/(mg·L^{-1})	34.853	1.723	1.499	0	0
浊度(NTU)	51.40	62.40	70.3	5.11	2.98
碱度($CaCO_3$含量)/(mg·L^{-1})	251.70	147.02	148.20	101.15	103.50
总硬度/(mmol·L^{-1})	1.09	1.13	1.13	1.17	1.11
K^+ 含量/(mg·L^{-1})	16.11	16.07	17.67	17.41	17.52
Na^+ 含量/(mg·L^{-1})	71.52	59.55	61.87	58.55	58.67
Ca^{2+} 含量/(mg·L^{-1})	5.68	6.06	6.02	6.48	6.18
Mg^{2+} 含量/(mg·L^{-1})	1.31	1.73	1.59	1.45	1.41
CO_3^{2-} 含量/(mg·L^{-1})	150.87	88.13	88.83	60.63	62.04
PO_4^- 含量/(mg·L^{-1})	0.70	3.57	4.26	2.96	2.97
总溶解性物质含量/(mg·L^{-1})	351	323	317	344	343
总悬浮物含量/(mg·L^{-1})	62	140	125	49	28
COD/(mg·L^{-1})	425	140	134	23	20
BOD/(mg·L^{-1})	86	8.5	13	0	6.2

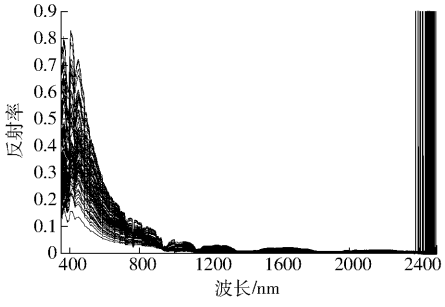

图 1 全波段原始光谱反射率曲线

Fig. 1 Full-band original spectral reflectance curves

一,通常利用 Savitzky - Golay 方法进行平滑滤波。Savitzky - Golay 方法是一种在时域内基于局域多项式最小二乘法拟合的滤波方法。其最大的特点在于在滤除噪声的同时可以确保信号的形状、宽度不变,可以提高光谱的平滑性,并降低噪声的干扰[13]。本次处理中移动窗口宽度为 5 及多项式次数为 3。

利用 Savitzky - Golay 滤波法对所有光谱数据(400 ~ 2 000 nm)进行平滑、去噪,取去噪声后的部分光谱数据曲线如图 2a 所示。

1.3.2 标准正态化处理

标准正态变量变换(SNV)预处理针对一条光谱进行处理,主要是消除光程变化、散射和颗粒大小之间的干扰[13]。计算公式为

$$X_i = \frac{x_{i,k} - x_i}{\sqrt{\dfrac{\sum_{k=1}^{m}(x_{i,k} - x_i)^2}{m-1}}} \quad (i = 1,2,\cdots,n) \quad (1)$$

式中 X_i——i 个样品光谱的平均值(标量)

m——波长点数

n——校正集样品数

采用标准正态变量变换(SNV)对所有光谱数据(400 ~ 2 000 nm)进行处理,取处理后的部分光谱数据曲线如图 2b 所示。

1.3.3 去包络线处理

去包络线(CR)处理将光谱曲线归一化到 0 ~ 1 之间,能够突出光谱曲线的吸收和反射特征,增强光谱曲线各波段之间的对比性[14]。

采用去包络线(CR)对所有光谱数据(400 ~ 2 000 nm)进行处理,取处理后的部分光谱数据曲线如图 2c 所示。

1.3.4 倒数之对数处理

在高光谱研究中,常将反射率进行倒数之对数变换,该变换形式有利于处理非线性问题,增强相似光谱之间的差异,并适当减少随机误差[15]。

采用倒数之对数处理方法对所有光谱数据(400 ~ 2 000 nm)进行处理,取处理后的部分光谱数据曲线如图 2d 所示。

平滑光谱反射率 SG、LR 在 ViewSpec Pro V6.0.11 软件中处理获得,指标 CR 利用 ENVI 5.1 的 Continuum Removed 模块处理得到。其他数据预处理通过 The Unscrambler X 10.4 实现。

1.4 模型建立与验证

1.4.1 样品集的划分

样品集的划分采用 Kennard - Stone 算法(简称 K - S)。K - S 算法是根据已经被选择的样品计算未被选择的样品的最小欧氏距离,然后通过选择经由计算的欧氏距离最大的样品进入校正集,以此反复,直至选出的样品数达到指定要求[16]。K - S 算法在选择具有代表性的样品方面已经被证明有着较好的效果[17]。本研究选取 58 个水质样本作为建模集,29 个水质样本作为验证集,分别用于模型的建立以及精度验证。

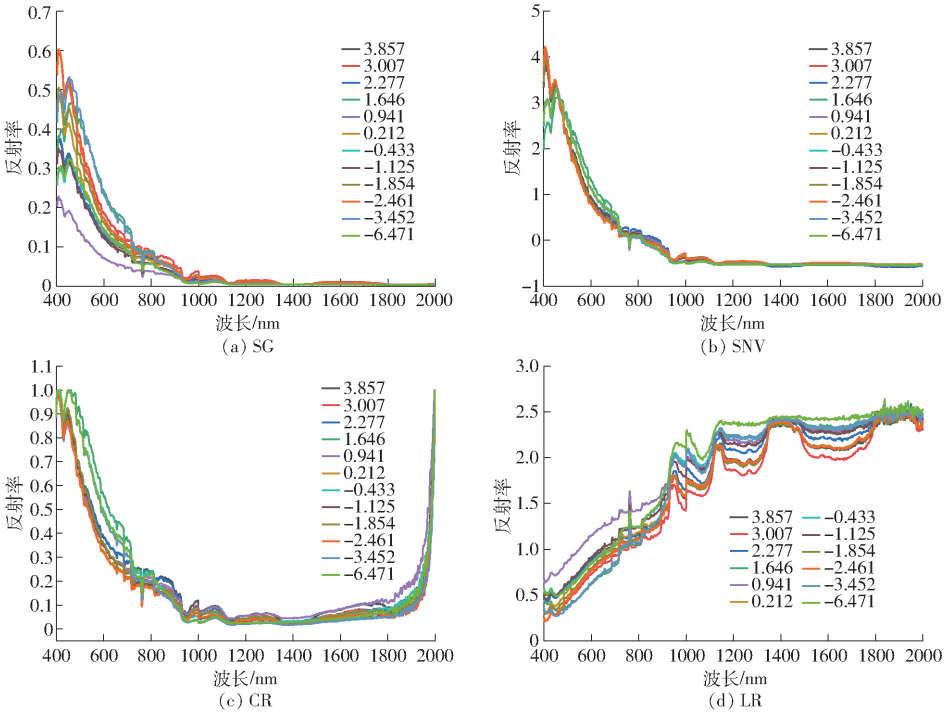

图 2　不同预处理后的光谱反射率曲线

Fig. 2　Spectral reflectance curves after different pretreatments

1.4.2　模型方法和评价指标

采用偏最小二乘回归法(PLSR)、逐步回归法(SR)和极限学习机(ELM)这 3 种回归方法建立高光谱遥感对水质综合评价的反演模型。其中 PLSR 在建模过程中具有降维、信息集成和波段优选等方法极大地提高了系统提取主成分的能力，得到了广泛地应用，可较好地解决自变量之间存在的共线性问题[18-20]。而 SR 是一种便捷高效的模型优化方法，对高光谱数据的“降维”具有较好的作用[21-22]。ELM 是一类基于前馈神经网络的机器学习算法，与传统的前馈神经网络相比较，ELM 有着学习效率高、精度高且参数调整简单等优点[23-25]。

模型的筛选利用 5 种指标：校正均方根误差(Root mean square error of calibration，RMSEC)、建模决定系数 R_c^2(Modeling determination of coefficients)、预测均方根误差(Root mean square error of prediction，RMSEP)、预测决定系数 R_p^2(Predicting determination of coefficients)、相对分析误差(Relative prediction deviation，RPD)。R_c^2、R_p^2 用以判定模型的稳定程度，越接近 1 说明模型的稳定性越好；RMSEC 及 RMSEP 用于表征模型的准确性，其值越小表明模型的精度越高[26]。另外，当 $RPD < 1.5$ 时，模型几乎无法对样本进行预测；当 $1.5 \leqslant RPD < 2$ 时，模型可以对样本进行粗略估计；当 $2 \leqslant RPD < 2.5$ 时，表明模型具有较好的定量预测能力；当 $2.5 \leqslant RPD < 3$ 时，模型具有很好的预测能力；当 $RPD \geqslant 3.0$ 时，表示模型具有极好的预测能力[22]。其中 R^2、RMSE 及 RPD 的计算公式为

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{3}$$

$$RPD = \frac{S_D}{RMSE} \tag{4}$$

式中　y_i、$\hat{y}_i$——验证样本的观测值和预测值

$\bar{y}_i$——样本观测值的平均值

n——验证样本数

S_D——样本观测值方差

$RMSE$——均方根误差

2　结果与分析

2.1　不同污水样本高光谱曲线特征分析

选 12 条较为典型的污水样本经过平滑(SG)、倒数之对数(LR)、去包络线(CR)、标准正态化(SNV)4 种预处理后的光谱曲线，见图 2。

由图 2a 可以发现，12 条光谱曲线波形基本相似。图 2b、2c、2d 分别是水体原始光谱经标准正态化(SNV)、去包络线(CR)、倒数之对数(LR)3 种预处理后的反射率。从图 2a、2b 可以看出，在 410、480 nm 波段处有明显的反射峰值，在 440、760、900、1 000 nm 处有明显的吸收峰值。这与吕航等[3]的研

究较为符合,是由于在 410、470 nm 波段处有较多的水质参数对反射更为敏感,而在 440、760、900、1 000 nm 处有较多的水质参数对光谱的吸收更为敏感。从图 2c、2d 可以看到,经由去包络线 CR 和倒数之对数 LR 处理后,放大了 1 000 ~ 2 000 nm 处的光谱波段,使得光谱吸收带变得更加明显,可以看到在 1 400 nm附近处和 1 900 nm 附近处也有着光谱的吸收敏感波段区,也验证了 CR 预处理能够突出光谱曲线的吸收和反射特征,增强光谱曲线各波段之间的对比性,以及 LR 预处理在增强相似光谱之间差异的优越性。

2.2 水质综合评价

水质系统是一个由各种水质污染指标变量组成的复杂系统,其内里蕴含众多能够影响水体质量的水质因子,每一种水质因子都只从某一方面表征了水体质量,而这些水质因子之间又往往有着不同程度的相关性,因此在对水质进行评价时,将这些水质因子都利用起来有一定的难度[27]。

主成分分析(Principal component analysis,PCA)是一种多元统计分析方法,其主要是利用降维的思想,把最初错综复杂的多个变量归纳总结成为少数几个综合变量,其中每一个综合变量都是原始变量的线性组合,各综合变量之间不存在相关性,从而实现利用少数几个综合变量来反映原始变量的绝大部分信息,且所含的信息互不重叠[28]。

主成分分析法(PCA)分析水质参数的基本思想是将 n 个水体样本的 P 个污染指标变量 $X_j(j=1,2,\cdots,P)$,通过对这 P 个污染指标变量相关性的研究,重新组合构造出 $m(m<P)$ 个综合指标,这 m 个综合指标能够反映原指标所提供的绝大部分信息[29]。

选择影响污水水质的水质参数中的 14 个独立的水质参数进行分析,14 个独立指标分别是 COD(X_1)、BOD(X_2)、NH_4^+-N 含量(X_3)、Na^+ 含量(X_4)、PO_4^- 含量(X_5)、K^+ 含量(X_6)、Ca^{2+} 含量(X_7)、Mg^{2+} 含量(X_8)、总溶解性物质含量(X_9)、总悬浮物含量(X_{10})、CO_3^{2-} 含量(X_{11})、浊度(X_{12})、碱度(X_{13})、总硬度(X_{14})。具体水质参数见表 1。运用统计软件 SPSS 23.0 对各项水质参数进行主成分分析,由于前 2 个特征值大于 1,且累计贡献率达到 95.954%,故前 2 个主成分分量可以代表全部指标,再根据这 2 个特征值的贡献率得到水质的综合评价因子,对 87 种水质进行定量分析评价。各水质参数主成分贡献特征值结果见表 2。水质综合评价因子越低,水质越好,水质分级标准见表 3[30-31]。

2.3 PLSR 建模及预测

偏最小二乘回归 PLSR 具有主成分分析、典型相关分析和多元线性回归等的优点。运用 The Unscrambler X 10.4 软件将全波段(400 ~ 2 000 nm)4 种光谱指标(SG、LR、SNV、CR)作为自变量,以水质综合评价因子作为因变量,通过将均方根误差(RMSE)和决定系数 R^2 对主因子数作图的方法确定最佳主因子数,建立 PLSR 回归模型。建模以及验证结果见表 4。

表 2 特征值贡献率

Tab.2 Eigenvalue contribution rate

编号	特征值	贡献率/%	累计贡献率/%
X_1	9.774	69.817	69.817
X_2	3.659	26.137	95.954
X_3	0.565	4.036	99.990
X_4	0.001	0.010	100

表 3 水质分级标准

Tab.3 Water quality classification standards

水质综合评价因子	水质类别
≤ -3	Ⅰ
(-3,-2]	Ⅱ
(-2,-1]	Ⅲ
(-1,0]	Ⅳ
>0	Ⅴ

表 4 水质指标的偏最小二乘模型

Tab.4 PLSR model of water quality indicators

光谱指标	主因子数	建模集		验证集		
		R_c^2	RMSEC	R_p^2	RMSEP	RPD
SG	9	0.906	0.748	0.869	0.755	2.737
SNV	9	0.943	0.567	0.913	0.674	3.047
CR	9	0.753	1.248	0.762	0.923	1.654
LR	6	0.737	1.226	0.705	1.201	1.706

由表 4 可得,基于 4 种光谱指标(SG、SNV、CR、LR)所建立的全波段水质参数的偏最小二乘(PLSR)模型中,SG、SNV 等具有很好的效果,建模集的决定系数 R_c^2 均在 0.9 以上,验证集的决定系数 R_p^2 均在 0.85 以上,相对分析误差结果均在 2.5 以上,具备很好的定量预测能力。其中 SNV - PLSR 模型拥有最高的 R_c^2、R_p^2、RPD 值和最小的 RMSE 值,其值分别为 0.943、0.913、3.047、0.674,为 PLSR 模型中的最佳模型。CR - PLSR 与 LR - PLSR 模型的效果都不理想,只能对样本的水质状况做一个定性的估计。这表明对原始数据经过不同的预处理以及数学变换后,并不一定能够提高模型的精度,合适的预处理和数学变换对模型精度的提高有重要作用。总体来说,PLSR 模型对于水质综合评价因子的反演效果很好,这进一步验证了 PLSR 模型在处理高光谱数据建模的适用性。

2.4　SR 建模及预测

逐步回归是一种线性回归模型自变量选择方法，其基本思想是将自变量逐个引入，根据自变量对因变量的解释程度或显著性，将对因变量解释程度小或者显著性低的自变量进行剔除，保留显著的解释变量，如此反复，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止，完成对数据的大幅降维，得到了最优的解释变量集。本文运用全波段 4 种光谱指标（SG、LR、SNV、CR）作为自变量，水质综合评价因子为因变量。变量入选和剔除的显著水平分别设为 0.15 和 0.25，由“最优”解释变量集所建立模型的结果见表 5。

表 5　水质指标的 SR 模型

Tab.5　SR model of water quality indicators

光谱指标	波段数目	中心波长/nm	建模集 R_c^2	验证集 R_p^2	验证集 RMSEP	验证集 RPD
SG	1	996	0.046	0.516	2.297	0.124
	2	996、400	0.497	0.694	1.309	1.593
	3	996、400、671	0.583	0.707	1.288	1.689
	4	996、400、671、970	0.645	0.780	1.143	2.101
	5	996、400、671、970、430	0.728	0.878	0.874	2.908
	6	996、400、671、970、430、708	0.781	0.896	0.856	3.115
	7	996、400、671、970、430、708、1 118	0.791	0.891	0.847	3.069
	8	996、400、671、970、430、708、1 118、1 910	0.815	0.852	0.997	2.637
	9	996、400、671、970、430、708、1 118、1 910、1 098	0.842	0.873	0.929	2.845
	10	996、400、671、970、430、708、1 118、1 910、1 098、1 106	0.85	0.868	0.940	2.794
	11	996、400、671、970、430、708、1 118、1 910、1 098、1 106、1 924	0.854	0.869	0.933	2.805
	12	996、400、671、970、430、708、1 118、1 910、1 098、1 106、1 924、1 102	0.857	0.869	0.928	2.808
SNV	1	846	0.154	0.267	1.987	0.377
	2	846、836	0.276	0.457	1.621	0.765
	3	846、836、766	0.497	0.751	1.112	1.530
	4	846、836、766、400	0.513	0.765	1.081	1.596
	5	846、836、766、400、672	0.572	0.808	0.951	2.043
	6	846、836、766、400、672、636	0.775	0.910	0.697	3.378
	7	846、836、766、400、672、636、1 008	0.820	0.924	0.660	3.596
	8	846、836、766、400、672、636、1 008、812	0.831	0.910	0.682	3.371
	7	846、766、400、672、636、1 008、812	0.827	0.894	0.725	3.084
	8	846、766、400、672、636、1 008、812、1 012	0.834	0.893	0.738	3.084
	9	846、766、400、672、636、1 008、812、1 012、837	0.839	0.912	0.688	3.398
LR	1	992	0.013	0.363	2.191	0.014
	2	992、970	0.328	0.498	1.551	0.871
	3	992、970、400	0.496	0.545	1.596	0.846
	4	992、970、400、412	0.580	0.576	1.408	0.846
CR	1	996	0.128	0.274	1.869	1.344
	2	996、762	0.311	0.264	1.605	0.489
	3	996、762、2 000	0.341	0.275	1.585	0.658
	4	996、762、2 000、402	0.429	0.311	1.541	0.641
	5	996、762、2 000、402、1 982	0.495	0.327	1.525	0.772
	4	996、2 000、402、1 982	0.494	0.324	1.531	0.786
	5	996、2 000、402、1 982、1 996	0.519	0.325	1.552	0.879
	6	996、2 000、402、1 982、1 996、584	0.552	0.388	1.452	0.826
	7	996、2 000、402、1 982、1 996、584、433	0.602	0.474	1.354	1.014
	8	996、2 000、402、1 982、1 996、584、433、807	0.703	0.617	1.152	1.298
	9	996、2 000、402、1 982、1 996、584、433、807、906	0.724	0.639	1.116	1.344

由表 5 可以看出，逐步回归方法通过对波段的“筛选”，剔除了大量对水质综合评价因子不显著的波段数据，仅保留了原数据约 1% 的显著波段数据，其降维效果非常显著。在 SR 建模对数据的“降维”过程中，原始光谱 SG 保留的波段数目最多，表明在“降维”过程中，原始光谱的波段和水质综合评价因子间具有较好相关性；而 LR 处理后的光谱数据保留的波段数目最少，表明在“降维”过程中，原始光

谱的波段数值和水质综合评价因子间具有较差的相关性。

由表 5 可以看到,从单个波段建模开始,其模型的 R_c^2、R_p^2、RPD 开始逐渐增大,有一个骤增到平缓的过程,这与张智韬等[22]的研究较为符合。在 SR 建模的 4 种模型中,SNV - SR 模型具有最好的建模预测效果,通过逐步回归筛选的波段数最多达 9 个。当波段数 $m = 7$ 时,SNV - SR 模型的 $R_p^2 = 0.924$,$RPD = 3.596$,$RMSE = 0.660$,是 SR 模型中的最佳模型。而 LR - SR 模型在通过"筛选"时,所保留的波段数目最少,其波段数目最多为 4 个,模型的效果也最差,表明经由 LR 处理后的光谱数据与水质综合评价因子间的相关性较差。

2.5 ELM 建模及预测

极限学习机(ELM)是由 HUANG 等[32]提出来的求解单隐层神经网络的算法。ELM 的网络训练模型由输入层、隐含层和输出层组成。其中,模型的训练效果受隐含层的神经元数量影响较大,且隐含层的神经元数量需人为确定。输入层和输出层的神经元数量取决于所分析问题的自变量和因变量数量。具体推导过程及训练步骤详见文献[33]。ELM 最大的特点是可以在保证学习精度的前提下相对于传统的神经网络的学习算法速度更快。

以经过逐步回归 SR 降维后的 4 种光谱指标(SG、LR、SNV、CR)的光谱数据作为自变量,水质综合评价因子为因变量建立 SR - ELM 模型。建模以及验证结果见表 6。

表 6 水质指标的极限学习机模型

Tab. 6 ELM model of water quality indicators

光谱	最佳隐含	建模集		验证集		
指标	层单元数	R_c^2	RMSEC	R_p^2	RMSEP	RPD
SG	484	0.827	1.016	0.836	0.833	2.507
SNV	309	0.956	0.500	0.954	0.500	4.651
CR	168	0.808	1.089	0.733	1.157	1.943
LR	148	0.670	1.374	0.651	1.294	1.523

由表 6 可知,基于 SR 降维后的 4 种光谱指标(SG、LR、SNV、CR)所建立的极限学习机模型中,SG - SR - ELM 和 SNV - SR - ELM 具有很好的效果,验证集的决定系数 R_p^2 均在 0.82 以上,相对分析误差均在 2.5 以上,具有很好的定量预测能力。LR - SR - ELM 模型和 CR - SR - ELM 模型相对效果较差,验证集的决定系数 R_p^2 和 RPD 分别为 0.651、0.733 和 1.523、1.943,预测能力一般,只能对样本进行粗略的估计。其中 SNV - SR - ELM 模型具有最高的 R_p^2、RPD 值和最小的 RMSE 值,其值分别为 0.954、4.651、0.500。

作为机器学习的一种算法,ELM 在本组数据建模中,隐含层的神经元数量对训练效果的影响较大。从表 6 可以看出,建模预测效果较好的 SG - SR - ELM 模型和 SNV - SR - ELM 模型的相对最佳隐含层单元数高于建模预测效果较差的 LR - SR - ELM 模型和 CR - SR - ELM 模型,反映了 SG 和 SNV 数据和水质综合评价因子间更具有相关性。总体来说,ELM 模型在本组数据中对于光谱的拟合和预测具有很好的效果。

2.6 模型对比

运用 3 种不同的回归方法对 4 种光谱指标进行建模,各个模型对污水水质的预测效果见图 3。

由图 3 可以看出,在 3 种回归模型中,基于 SR 降维后的波段建立的 SNV - SR - ELM 模型反演精度最高。其建模集和验证集的决定系数 R_c^2 和 R_p^2 最高,分别达到 0.956 和 0.954,同时具有最高的相对分析误差 RPD 和最低的均方根误差 RMSEP,分别为 4.651 和 0.500。这进一步验证了极限学习机 ELM -高光谱遥感模型在定量预测污水水质参数方面的可行性,以及降维处理对于 ELM 模型在去除冗余信息、提高预测效率和精度的重要作用,这与张峥等[33]的研究相符合。同时针对本组试验数据,发现并不是每种预处理方法都能够使建模效果得到改善,有些预处理方法(CR、LR)破坏原来较好的数据,使建模效果变差,这与吴元清等[34]的研究结果较为一致。而在本组数据中,经过 SNV 处理后的光谱数据,在 SR 模型、PLSR 模型、ELM 模型中具有很好的效果,与吴元清等[34]的研究结果出现差异,本研究认为这是由于建模时的因变量是由主成分分析降维得到的水质综合评价因子,SNV 处理可以减小光谱的绝对强度,对光谱数据进行均衡化,更有利于处理类似于本组数据这种由多因素综合作用所得到的光谱数据,以分析与不同水质参数综合作用得到的综合评价因子的内在联系和特点,降低因不同的水质参数对光谱数据带来的特异性交互影响。这进一步证明 SNV 预处理可以消除光程变化、散射和颗粒大小之间的干扰,提高模型精度[14]。

2.7 讨论

高光谱遥感在定量反演水质参数时,由于具有光谱分辨率高和波段连续性强等特点,可以获得更为全面广泛的光谱波段数据,而由于光谱测量中的某些人为和自然因素的干扰,光谱数据需要进行不同的数学变换以增强信噪比,从而提高光谱数据与水质参数的相关性,进而提高模型的预测精度[22]。

虽然高光谱遥感在实际应用中,可以获得更为精细的光谱信息,但因此也造成了数据和计算量的

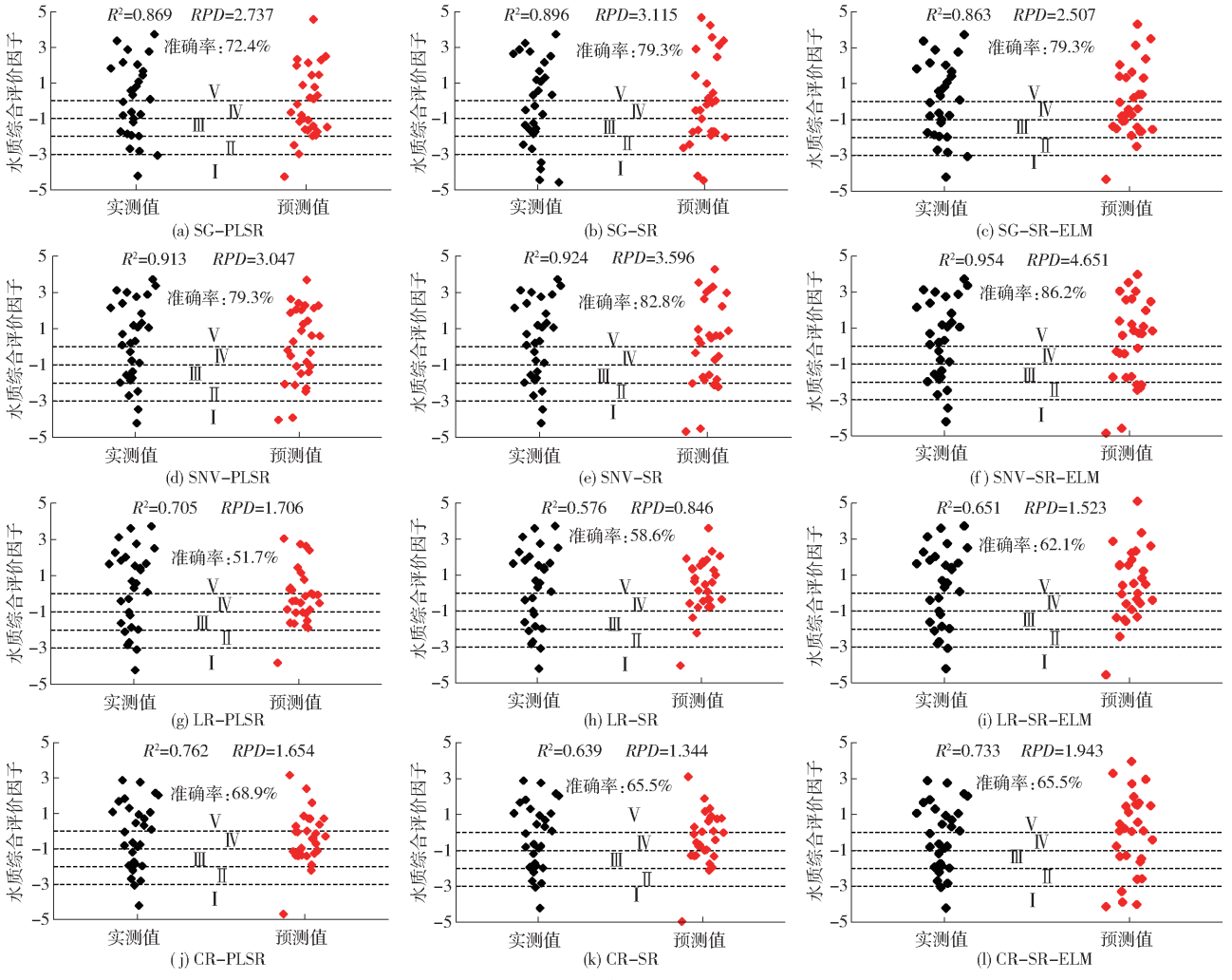

图 3　水质综合评价反演模型

Fig. 3　Inversion model for comprehensive evaluation of water quality

冗杂，为实现数据的筛选和模型的简化，本文通过逐步回归方法对光谱数据进行分析筛选建模。结果表明，逐步回归方法能够有效地对数据进行大幅度的降维（可达到 99%），使得筛选后留下的数据既是重要的，又没有严重的多重共线性。同时 SR 模型也有着很好的精度和预测效果，而以经过逐步回归筛选后的波段所建立的 SR－ELM 模型的精度和预测效果明显比 PLSR 模型和 SR 模型更优越，这为下一步的对高光谱数据通过波段筛选和数据降维以实现模型的简化提供了可行性。由水质综合评价因子和光谱数据建立的水质综合反演模型较由单项水质参数建立的反演模型精度有所下降，没有能够很好地表征各不同的单项水质参数对于光谱数据各波段的影响和作用，有待进一步探索各水质参数对于光谱不同波段数据的单独影响和作用以及综合的交互影响和作用。另外，由于内陆水体光学特征的复杂性、水质影响因子的多样性，如何更好地将水质信息状况的综合评价与高光谱技术相结合，以达到通过高光谱遥感技术实时全面地反映水体的污染程度，主要污染物的类别、来源、成因、时空分布规律以及变化趋势，值得去进一步深入地研究和探索，是下一步研究的方向和目标。

3　结论

（1）PLSR 模型、SR 模型、SR－ELM 模型均能得到具有很好精度和预测效果的模型，其中 SR－ELM 模型的精度高于其他两个模型，更适用于处理本组的水质综合评价因子。

（2）SR 模型较其他两个模型具有明显的简便性和快速性，在数据降维筛选中，筛选了与水质综合评价因子相关的显著性波段。但其模型受不同的预处理和数学变换的影响较大，SNV－SR 模型具有最好的预测效果（$RPD = 3.596$），而 LR－SR 模型和 CR－SR 模型决定系数 R_p^2 以及相对分析误差（RPD）大部分都小于 0.65 和 1.5，验证效果较差。

（3）ELM 模型预测水质综合评价因子时具有很好的效果，为机器学习在水质参数反演预测方面

的应用验证了可行性,其中 SNV - SR - ELM 模型为水质的综合评价方法和高光谱反演模型的结合提供了参考。

(4)基于标准正态化变换指标建立的偏最小二乘模型、逐步回归模型与极限学习机模型,其决定系数和 RPD 均最高,反演精度最优,SNV 为本组光谱数据的最佳预处理方法。其中 SNV - SR - ELM 模型决定系数为 0.954,RPD 为 4.651,为本组数据的最佳模型。

(5)水质综合评价方法和高光谱技术反演水质参数的结合具有可行性,其模型的建立可以为水体的快速监测和综合评价提供参考。

参 考 文 献

[1] 张渊智,聂跃平,蔺启忠,等. 表面水质遥感监测研究[J]. 遥感技术与应用,2000(4):214 - 219.
ZHANG Yuanzhi, NIE Yueping, LIN Qizhong, et al. Surface water quality monitoring using remote sensing[J]. Remote Sensing Technology and Application, 2000(4): 214 - 219. (in Chinese)

[2] YE S, WANG D, MIN S. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection[J]. Chemo Metrics and Intelligent Laboratory Systems, 2008, 91(2): 194 - 199.

[3] 吕航, 马蔚纯, 周立国,等. 淀山湖微量水质参数卫星高光谱遥感估算[J]. 复旦大学学报(自然科学版), 2013, 52(2):238 - 246.
LÜ Hang, MA Weichun, ZHOU Liguo, et al. Estimation for water quality parameters in Dianshan lake based on HJ - 1A HSI image[J]. Journal of Fudan University(Natural Science), 2013, 52(2): 238 - 246. (in Chinese)

[4] 曹引,冶运涛,张小娟,等. 南四湖水体浊度高光谱定量反演模型[J]. 南水北调与水利科技,2015,13(5):883 - 887.
CAO Yin, YE Yuntao, ZHANG Xiaojuan, et al. Quantitative inversion model of hyperspectral for turbidity in the Nansi Lake [J]. South-to-North Water Transfers and Water Science & Technology, 2015, 13(5): 883 - 887. (in Chinese)

[5] BANSOD B, SINGH R, THAKUR R. Analysis of water quality parameters by hyperspectral imaging in Ganges River[J]. Spatial Information Research, 2018, 26(2): 203 - 211.

[6] BARUCH A. Water quality measurements from hyperspectral remote sensing: the case of the river ganga[C]// American Geophysical Union Fall Conference, 2014.

[7] BAGHERI S. Hyperspectral remote sensing of nearshore water quality[M]. Springer International Publishing, 2017.

[8] WIANGWANG N, QI J. Water quality assessment using hyperspectral remote sensing[C]// IEEE International Geoscience & Remote Sensing Symposium. IEEE, 2005.

[9] GIARDINO C, BRANDO V E, GEGE P, et al. Imaging spectrometry of inland and coastal waters: state of the art, achievements and perspectives[J]. Surveys in Geophysics, 2019, 40(3): 401 - 429.

[10] 马小雪, 王腊春, 廖玲玲. 温瑞塘河流域水体污染时空分异特征及污染源识别[J]. 环境科学, 2015(1):64 - 71.
MA Xiaoxue, WANG Lachun, LIAO Lingling. Spatio-temporal characteristics and source identification of water pollutants in Wenruitang River watershed[J]. Environmental Science, 2015(1): 64 - 71. (in Chinese)

[11] 徐国宾,翟晶. 基于模糊标识指数的水功能区水质评价方法[J]. 天津大学学报(自然科学与工程技术版),2017, 50(7):710 - 716.
XU Guobin, ZHAI Jing. Water quality evaluation method based on fuzzy identification index in water function zone[J]. Journal of Tianjin University(Science and Technology), 2017, 50(7): 710 - 716. (in Chinese)

[12] 王云鹏,招赞铭. 两种测定水的室内可见-近红外反射光谱的方法及应用比较[J]. 遥感信息,2000(1):2 - 5.
WANG Yunpeng, ZHAO Zanming. Comparison of two methods and applications for determination of indoor visible-near infrared reflectance spectra of water[J]. Remote Sensing Information, 2000(1): 2 - 5. (in Chinese)

[13] 褚小立,袁洪福,陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展,2004(4):528 - 542.
CHU Xiaoli, YUAN Hongfu, LU Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. Progress in Chemistry, 2004(4): 528 - 542. (in Chinese)

[14] 彭小婷,高文秀,王俊杰. 基于包络线去除和偏最小二乘的土壤参数光谱反演[J]. 武汉大学学报(信息科学版),2014, 39(7):862 - 866.
PENG Xiaoting, GAO Wenxiu, WANG Junjie. Inversion of soil parameters from hyperspectral based on continuum removal and partial least squares regression[J]. Geomatics and Information Science of Wuhan University, 2014, 39(7): 862 - 866. (in Chinese)

[15] 于雷,洪永胜,耿雷,等. 基于偏最小二乘回归的土壤有机质含量高光谱估算[J]. 农业工程学报,2015,31(14):103 - 109.

[16] KENNARD R W, STONE L A. Computer aided design of experiments[J]. Technometrics, 1969, 11(1): 137 - 148.

[17] 刘伟,赵众,袁洪福,等. 光谱多元分析校正集和验证集样本分布优选方法研究[J]. 光谱学与光谱分析,2014,34(4): 947 - 951.
LIU Wei, ZHAO Zhong, YUAN Hongfu, et al. An optimal selection method of samples of calibration set and validation set for spectral multivariate analysis [J]. Spectroscopy and Spectral Analysis, 2014, 34(4): 947 - 951. (in Chinese)

[18] RYAN K, ALI K. Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data[J]. Ocean Science Journal, 2016, 51(2): 209 - 221.

[19] 朱亚星,周桢津,洪永胜,等. 耦合高光谱数据估算土壤含水率的方法[J]. 华中师范大学学报(自然科学版),2017, 51(1):123 - 129.
ZHU Yaxing, ZHOU Zhenjin, HONG Yongsheng, et al. Prediction of soil moisture content based on coupled hyperspectral data

[J]. Journal of Central China Normal University(Natural Sciences),2017,51(1):123 - 129. (in Chinese)
[20] 于雷,朱亚星,洪永胜,等. 高光谱技术结合 CARS 算法预测土壤水分含量[J]. 农业工程学报,2016,32(22):138 - 145.
YU Lei, ZHU Yaxing, HONG Yongsheng, et al. Determination of soil moisture content by hyperspectral technology with CARS algorithm[J]. Transactions of the CSAE,2016,32(22):138 - 145. (in Chinese)
[21] 叶勤,姜雪芹,李西灿,等. 基于高光谱数据的土壤有机质含量反演模型比较[J/OL]. 农业机械学报,2017,48(3):164 - 172.
YE Qin, JIANG Xueqin, LI Xican, et al. Comparison on inversion model of soil organic matter content based on hyperspectral data[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2017,48(3):164 - 172. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 20170321&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2017.03.021. (in Chinese)
[22] 张智韬,王海峰,KARNIELI Arnon,等. 基于岭回归的土壤含水率高光谱反演研究[J/OL]. 农业机械学报,2018,49(5):240 - 248.
ZHANG Zhitao, WANG Haifeng, KARNIELI Arnon, et al. Inversion of soil moisture content from hyperspectra based on ridge regression[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2018,49(5):240 - 248. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 20180528&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2018.05.028. (in Chinese)
[23] 唐贤伦,周家林,张娜,等. 基于极限学习机的非线性内模控制[J]. 电子科技大学学报,2016,45(1):96 - 101.
TANG Xianlun, ZHOU Jialin, ZHANG Na, et al. Nonlinear internal model control system based on weighted regularized extreme learning machine[J]. Journal of University of Electronic Science and Technology of China,2016,45(1):96 - 101. (in Chinese)
[24] 张颖,李梅. 基于粒子群优化极限学习机的水质评价新模型[J]. 环境科学与技术,2016,39(5):135 - 139.
ZHANG Ying, LI Mei. A novel evaluation model of water quality based on PSO - ELM method[J]. Environmental Science & Technology,2016,39(5):135 - 139. (in Chinese)
[25] 周鹏,杨玮,李民赞,等. 基于灰度关联-极限学习机的土壤全氮预测[J/OL]. 农业机械学报,2017,48(增刊):271 - 276.
ZHOU Peng, YANG Wei, LI Minzan, et al. Soil total nitrogen content prediction based on gray correlation-extreme learning machine[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2017,48(Supp.):271 - 276. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 2017s041&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2017.S0.041. (in Chinese)
[26] 王敬哲,塔西甫拉提·特依拜,丁建丽,等. 基于分数阶微分预处理高光谱数据的荒漠土壤有机碳含量估算[J]. 农业工程学报,2016,32(21):161 - 169.
WANG Jingzhe,TASHPOLAT Tiyip,DING Jianli, et al. Estimation of desert soil organic carbon content based on hyperspectral data preprocessing with fractional differential[J]. Transactions of the CSAE,2016,32(21):161 - 169. (in Chinese)
[27] 姚焕玫, 黄仁涛, 甘复兴,等.用改进的主成分分析法对东湖的水质污染进行评价[J]. 武汉大学学报(信息科学版), 2005, 30(8):732 - 735.
YAO Huanmei, HUANG Rentao, GAN Fuxing, et al. Principal component analysis of the water quality evaluation in east lake [J]. Geomatics and Information Science of Wuhan University,2005, 30(8):732 - 735. (in Chinese)
[28] 王敏. 污水水质毒性评价及排污河道生态修复效果研究[D]. 天津:南开大学,2012.
WANG Min. Study on the toxicity evaluation of sewage quality and ecological restoration effect of sewage channel[D]. Tianjin: Nankai University,2012. (in Chinese)
[29] 林海明,杜子芳. 主成分分析综合评价应该注意的问题[J]. 统计研究,2013,30(8):25 - 31.
LIN Haiming, DU Zifang. Some problems in comprehensive evaluation in the principal component analysis[J]. Statistical Research,2013,30(8):25 - 31. (in Chinese)
[30] 陈俊英,张智韬,LEIONID Gillerman,等. 影响土壤斥水性的污灌水质主成分分析[J]. 排灌机械工程学报,2013,31(5):434 - 439.
CHEN Junying, ZHANG Zhitao,LEIONID Gillerman, et al. Analysis of principal components of wastewater affecting soil water repellency[J]. Journal of Drainage and Irrigation Machinery Engineering,2013,31(5):434 - 439. (in Chinese)
[31] 邹海明,蒋良富,李粉茹. 基于主成分分析的水质评价方法[J]. 数学的实践与认识,2008(8):85 - 90.
ZOU Haiming, JIANG Liangfu, LI Fenru. Water quality evaluation method based on principal component analysis[J]. Mathematics in Practice and Theory,2008(8):85 - 90. (in Chinese)
[32] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: a new learning scheme of feedforward neural networks in neural networks[C]// Proceedings 2004 IEEE International Joint Conference on IEEE,2004:985 - 990.
[33] 张峥,魏彪,汤戈,等. 一种紫外-可见光谱法水质 COD 检测的预测模型研究[J]. 激光杂志,2016,37(4):21 - 24.
ZHANG Zheng, WEI Biao, TANG Ge, et al. A prediction model for the determination of water COD by using UV - visible spectroscopy[J]. Laser Journal,2016,37(4):21 - 24. (in Chinese)
[34] 吴元清,杜树新,严赟. 水体有机污染物浓度检测中的紫外光谱分析方法[J]. 光谱学与光谱分析,2011,31(1):233 - 237.
WU Yuanqing, DU Shuxin, YAN Yun. Ultraviolet spectrum analysis methods for detecting the concentration of organic pollutants in water[J]. Spectroscopy and Spectral Analysis,2011,31(1):233 - 237. (in Chinese)