

白洋淀水体总磷总氮高光谱估算模型比较

陈洁^{1,2,3}, 张立福¹, 张红明¹, 张琳珊^{1,2}, 岑奕¹, 童庆禧¹

1. 中国科学院空天信息创新研究院 遥感科学国家重点实验室, 北京 100101;

2. 中国科学院大学, 北京 100049;

3. 中国自然资源航空物探遥感中心, 北京 100083

摘要: 总磷(TP)、总氮(TN)是水质富营养化的重要指标,亦是水质监测的主要参数。具有快速高效、无二次污染等特点光谱法水质监测是当今水环境遥感分析研究的热点。针对水体TP、TN反演模型采用实验室标准液或野外全样本进行建模时,各种水质参数的相互影响及预测值超出建模样本值域的可能,使得实际的预测效果并不理想的情况。本文以白洋淀实验区的实际水体样本为反演模型的输入值,在确定最优相关波段和最佳反演模型的基础上,讨论了5种不同浓度范围场景下的样本建模对反演模型的影响,同时剖析了模型对超出建模浓度值样本的预测能力。结果表明:建模样本浓度覆盖预测样本时,反演模型决定系数 $R^2 > 0.6$, TP、TN浓度预测值的平均偏离度 $ARE < 20\%$;建模浓度高于预测样本时, R^2 在0.6左右,对超过建模浓度范围12%以内的预测值,其 $ARE < 25\%$;建模浓度低于预测值时, R^2 介于0.4—0.5,预测值超过建模样本浓度一倍时, $ARE \leq 30\%$;建模样本浓度位于预测值两侧时, R^2 可达0.8, $ARE < 25\%$;建模样本浓度值介于预测值之间时, $0.45 < R^2 < 0.55$, $ARE > 35\%$ 。通过本文的研究与讨论,可为水质参数监测的实际工程应用提供科学依据和参考。

关键词: 水质监测, 总磷, 总氮, 浓度反演, 偏最小二乘法

中图分类号: P2

引用格式: 陈洁,张立福,张红明,张琳珊,岑奕,童庆禧.2023.白洋淀水体总磷总氮高光谱估算模型比较.遥感学报,27(7):1642–1652

Chen J, Zhang L F, Zhang H M, Zhang L S, Cen Y and Tong Q X. 2023. Comparative study on the hyperspectral estimation models of TP and TN in Baiyangdian water body. National Remote Sensing Bulletin, 27(7): 1642–1652 [DOI:10.11834/jrs.20210575]

1 引言

水体富营养化是当今水环境恶化的主要表现之一,严重危害水生生物生命,对人类的水资源造成严重影响(Xia和Chen, 2015; Zhai等, 2017; Zhang等, 2018)。水体中总磷TP(Total Phosphorus)与总氮TN(Total Nitrogen)是衡量水质污染程度和水体富营养化的重要指标,其中TP是指以元素磷、磷酸盐和有机团结合的磷酸盐等形式存在于水中的磷总和,TN是指水体中各种形态无机和有机氮的总量,它们的主要来源为生活污水、化肥、有机磷农药及近代洗涤剂(Rixon等, 2020)。TP、TN是地表水中藻类生长的关键元素,当其超标时

会出现水体的富营养化,即微生物大量繁殖,浮游生物生长旺盛,内陆水体富营养化、海洋出现赤潮。常规的测定方法是利用能与水中磷、氮元素反应的化学试剂与其发生氧化,再使用分光光度法的单色光测定TP、TN含量。该方式对实验条件、操作人员水平等有较高要求,精度高、周期长、水体样本经测定后无法反复使用,其测定的水质参数值通常作为后续反演建模的“真值”。

随着高光谱技术的发展,利用光谱法进行水质参数的反演逐步从实验研究走向实用化,具有操作简单、无污染、实时在线、多参数同步监测的特点,代表了未来水环境监测的发展方向(童庆禧等, 2016)。在基于光谱技术的水质TP、TN

收稿日期: 2021-01-06; 预印本: 2021-02-25

基金项目: 国家自然科学基金(编号:41830108,41977154);兵团重大科技项目(编号:2018AA004);兵团重点领域创新团队(编号:2018CB004);国家重点研发计划(编号:2017YFC1500901)

第一作者简介: 陈洁,研究方向为高光谱遥感水质监测。E-mail:chenjie@aircas.ac.cn

通信作者简介: 张立福,研究方向为高光谱遥感。E-mail:zhanglf@radi.ac.cn

的反演研究上, 龚绍琦等 (2008) 在实验室环境下利用配置的 TP、TN 标准液, 建立了单波段反射率与 TP、TN 浓度的线性回归方程, 虽然其模型 R^2 优于 0.9, 但预测精度较低, 其研究还停留在以清水中进行氮磷的实验阶段, 忽略了实际水体中藻类、悬浮物的影响; 吕航等 (2013) 利用高光谱反射率的双波段减法构建了 TP、TN 的浓度反演模型, 其模型的建立是基于预测浓度与建模样本浓度为同一范围的假设之上的; 王云霞 (2017) 利用最小二乘支持向量机模型 (LS-SVM) 估算了沈阳清水水库的 TP、TN 含量并取得了较好的结果, 但其精度验证选用的是留一法, 对于超出建模样本浓度范围的 TP、TN 预测能力则缺少评价; 林剑远和张长兴 (2019) 同样利用全样本建模的方式, 分别采用指数和二次多项式的非线性模型对嘉兴市河网的 TP、TN 等 4 项水质参数进行了定量反演, 但其仅使用了一组相关性最强的波长比值, 忽略了其他特征波长的影响; 邹兴红等 (2020) 仅利用与 TP、TN 具有最大相关性的单波段敏感波长的归一化和一阶微分光谱反射率为自变量, 建立了 TP、TN 估算方程, 输入样本的选取采用了随机法, 没有顾及不同浓度范围样本对其反演能力的影响; 余祝媛 (2019) 在对红枫湖的 TP 含量进行预测时, 仅采用了 823 nm 处的一阶微分光谱值。上述研究均表明, 反射率光谱与水中 TP、TN 含量存在线性或非线性的回归关系, 利用与水质参数敏感的特征波长可以较好地预测目标区的 TP、TN 含量。但它们普遍存在以下两方面的问题:

(1) 反演模型的建立往往只使用单一的特征波段, 忽略了其他敏感波长对反演模型的影响。对于 TP、TN 来说, 与反射率具有高相关性的波长和组合很多, 前人的研究多取具有最大相关性的波长, 但该波长往往又与其他水质参数含量具有高相关性, 如 TP 与溶解氧、TN 与 TP 的相关系数分布图具有相似的特点, 从而在一定程度上降低了反演模型的精度。所以需要尝试利用更多的相关波长作为输入值构建反演模型, 并与常规的单波段法进行比较, 评价其预测精度。

(2) 建模样本的选取时, 其水质参数含量的变化或呈均匀梯度变化, 或涵盖了预测样本的含量值。而在实际的同一地区的某个水质参数浓度的变化量往往较小, 对于 TP、TN 来说, 内陆水体的含量值本身就很很小 (II 类水体的 $TP \leq 0.1 \text{ mg/L}$,

$TN \leq 0.5 \text{ mg/L}$), 因此利用实际水样反射率与水质参数浓度建立的反演模型, 往往面临预测非建模样本浓度范围内的问题。如何评价反演模型对非建模样本浓度范围内的水体参数能力, 亦是本文研究讨论的问题。

基于如上考虑, 本文以河北白洋淀实验区的 62 个观测点数据为研究对象, 在比较单波段建模法和基于多波段输入的偏最小二乘法建模效果的基础上, 讨论了不同样本浓度建模的反演效果, 确定了各种情况下对超出建模浓度范围的预测样本估算值的偏离度, 为水质参数发生异常变化时光谱法反演 TP、TN 含量的置信度提供了参考依据。

2 研究区与数据预处理

2.1 研究区概况

实验区位于河北白洋淀, 使用水质光谱在线监测系统的光谱探测器采集了 62 个点的水体光谱数据, 并取水样进行化学分析以得到浊度、总磷、总氮等水质参数。实验区地理位置和观测采样点分布如图 1 所示。

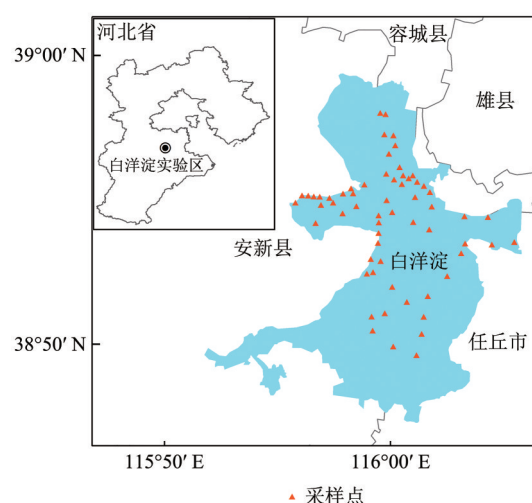


图1 白洋淀实验区水体光谱采集点分布示意图

Fig. 1 Schematic diagram of the distribution of water spectrum collection points in Baiyangdian

2.2 数据预处理

光谱探测器采集的水体原始光谱中存在许多随机噪声干扰信息, 这些干扰信息会降低所建立模型的预测准确性, 因此通常采用平滑算法对光谱图进行预处理, 提升输出光谱的真实性和可靠性。有关平滑处理的数学方法很多, 如傅里叶变换

(Fuller和Ogilvie, 2015)、奇异值分解 (Golyandina, 2010) 等, 但最常用的方法是Savitsky与Golay提出的多项式平滑方法 (也称卷积平滑), SG滤波法 (Savitzky Golay Filter) 的核心思想也是对窗口内的数据进行加权滤波, 但是它的加权重是对给定的高阶多项式进行最小二乘拟合得到 (宁鸿章等, 2020)。它的优点在于, 在滤波平滑的同时, 能够更有效地保留信号的变化信息。本研究采用此方法对光谱进行平滑处理。该方法基于最小二乘法原理, 能够保留分析信号中的有用信息, 消除随机噪音。SG滤波法对于数据的观测信息保持的更好, 在一些注重数据变化的场合会比较适用, 经SG处理后的实验区水体光谱曲线如下图2所示。

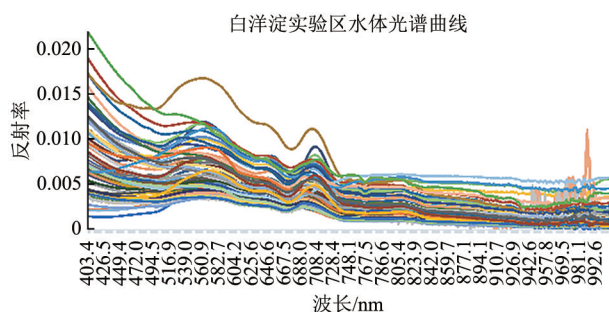


图2 经SG处理后的62个观测点反射率曲线
Fig. 2 Reflectance curve of 62 observation points after SG treatment

3 实验结果与讨论

3.1 特征提取与模型构建

3.1.1 特征波段提取

按照实验室对采集的水体样本进行的化学法测量结果, 逐波长地计算反射率值、反射率一阶微分值和反射率比值的相关系数 R 。 R 的计算公式如下:

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

式中, x_i 和 y_i 分别为光谱观测值和水质参数测量值, $\bar{x}$ 和 $\bar{y}$ 是光谱值与水质参数值的平均值, n 为样本数,

相关系数 R 的绝对值在0.8以上时认为两变量间有强的相关性; R 在0.5至0.8之间时认为两变量中度相关; R 在0.3至0.5时认为两变量间相关性较

弱; R 在0.3以下时可认为两变量间没有相关性。对采集的62组光谱数据进行总磷、总氮浓度和单波长反射率、单波长反射率一阶、反射率比值进行相关性分析, 得到的 R 随波长变化的结果如图3。

由图3(a)可知, 单波长反射率与TP含量的相关系数 R 值不大, 正、负 R 的绝对值最大不足0.3, 说明TP的含量与400—1000 nm的各个波长反射率基本不存在相关性。反射率一阶与TP浓度的最大正相关 R 为0.7625, 最大负相关 R 为-0.5968, 表现为中度相关性。由图3(b)可知, 比值法相关系数 R 明显大于上述单波长法和一阶微分法, 其中最大正相关系数 R 超过0.8, 说明该比值与TP浓度具有高度的相关性。因此在建模时可以考虑不用考虑单波长的建模, 只选用一阶微分法和比值法进行。由图3(a)可以看出, TN浓度与单波长反射率、反射率一阶的相关系数均较小, 仅有一阶微分的负相关 R 的绝对值超过了0.5。而图3(c)则说明TN的含量与波长反射率的比值运算值具有很好的相关性。具体来看, TN浓度与单波段反射率的最大正相关 R 为0.3516, 最大负相关 R 为-0.0619, 因此在后续TN反演建模中, 不考虑单波段的情况; 反射率一阶和比值法与TN浓度的相关系数均在0.5左右, 一阶微分的负相关的 R 最大, 为-0.5217, 因此在后续TN浓度反演建模只考虑一阶微分和比值法进行。表1统计了不同反射率模型下的最大相关系数值及其对应的特征波长。

3.1.2 线性模型构建与比较

按照上述最优相关性对应的波长模型, 利用不同的方式建立TP、TN反演模型, 其中TP、TN的含量为纵坐标 y 、反射率运算值为横坐标 x 。回归模型及其决定系数 R^2 的统计如表2所示。

决定系数 R^2 可以评价回归模型的拟合优度, 其最大值为1, 定义 R^2 愈接近1则表示拟合性愈佳。对于TP反演模型, 高次线性回归的效果要优于低次、指数及对数函数的结果, 总磷浓度与反射率一阶微分正相关波长建立的反演模型要达到六次幂的线性函数其 R^2 才高于0.6; TP浓度与反射率一阶微分负相关波长建立的反演模型其 R^2 最大值仅为0.2639, 因此就不在上表中进行罗列; 比值法中正相关波段所有回归模型的 R^2 均大于0.6, 说明这些反演模型都具有较好的预测效果, 随着线性回归阶数的增加 R^2 也增大, 但增大的程度较

缓慢; 比值法负相关波长的建模效果总体上与正相关波长类似, 除指数回归外其余反演模型 R^2 均大于0.6。对于TN反演模型, 利用反射率一阶进行TN浓度的反演建模时, 不论是正相关波长还是负相关波长, 模型的 R^2 最大不超过0.3, 回归效果不

佳; 比值法建模的高次线性回归的效果要优于低次、指数及对数函数的结果, 正相关的回归模型的 R^2 高于正相关的模型。但总体上建模效果不理想, 高阶线性回归的 R^2 最大值不到0.6。

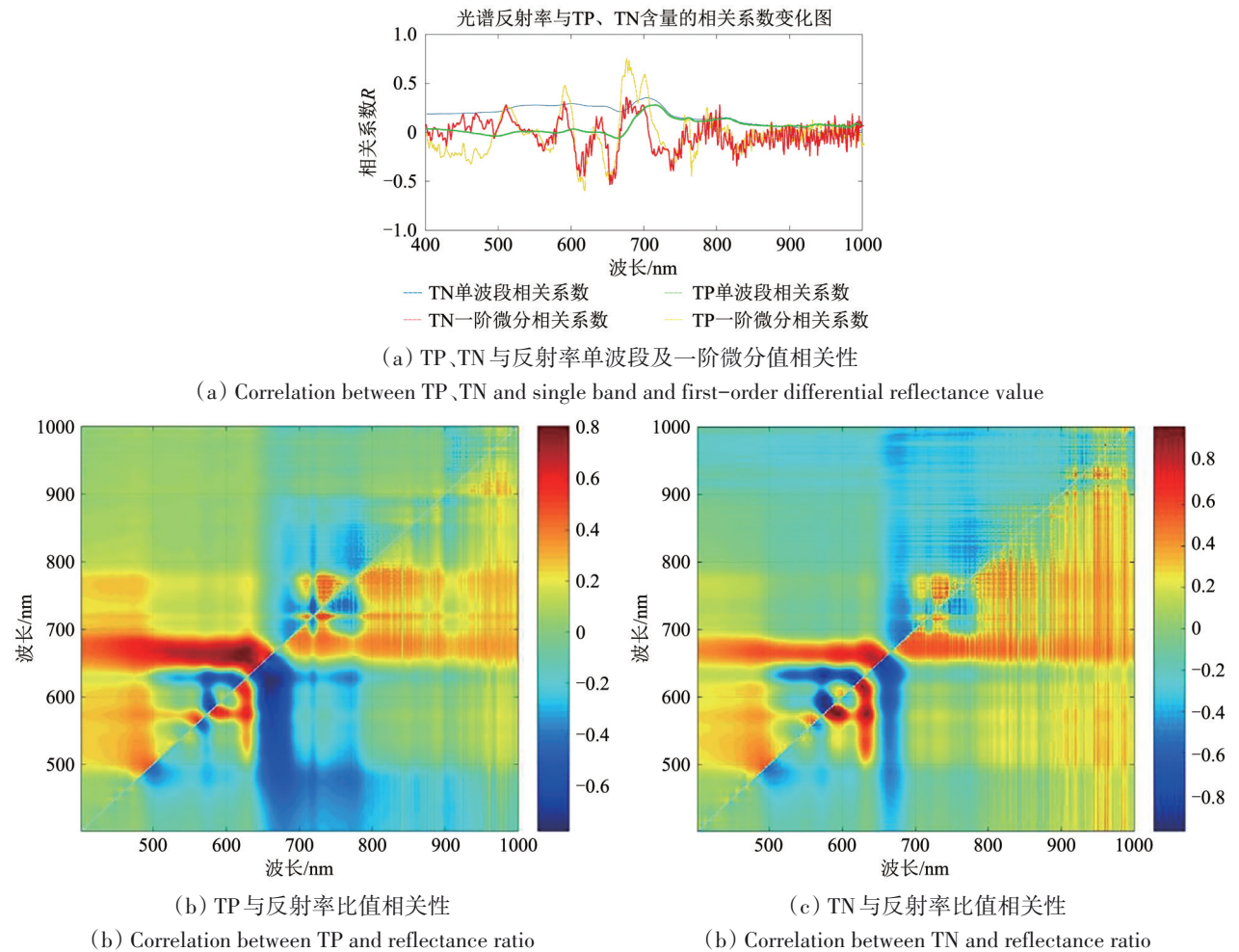


图3 反射率与TP、TN含量的相关系数变化图

Fig. 3 Change graph of correlation coefficient between reflectance and TP and TN content

表1 不同反射率运算值与TP、TN含量的最大相关系数(R)统计

Table 1 Statistics of the maximum correlation coefficient (R) between different reflectance calculation values and TP and TN content

反射率模型	TP敏感 波长/nm	R_{TP}	TN敏感 波长/nm	R_{TN}
单波段	716.40	0.2587	704.3	0.3516
	666.12	-0.0796	976.2	-0.0619
一阶微分	677.13	0.7625	675.7	0.3832
	619.92	-0.5968	653.6	-0.5217
比值法	703/657	0.8077	701/679	0.7550
	657/700	-0.7800	696/697	-0.7269

通过对河北白洋淀实验区TP、TN水质参数的反演模型建立和回归效果评价, 可以得出如下结论:

(1) 单波长反射率与TP、TN浓度的相关性不高。这主要是由于实际水体中所含的各种水质参数可达十几种, 其相互间的相互影响可能会淹没单一水质参数本身的特征波长, 所以仅依赖单波长反射率进行回归分析, 其效果不佳。

(2) 反射率一阶微分与TP、TN水质参数的相关性不强。这可能是由于水体反射率本身就不高, 而且最大与最小反射率相差较小。由经SG处理后的反射率图可以看出, 反射率分布在3%的范围

内，一阶微分反映的是数据的变化情况，而反射率如此平缓的变化使用微分方式进行就显现不出其特点和优势。

(3) 高阶的线性回归效果更优。TP、TN的反演模型中，线性函数要优于指数和对数，高阶的优于低阶的。但随着阶数的增加， R^2 增加速度放缓。

表 2 基于最优相关系数的特征波长建立的 TP 线性反演模型比较

Table 2 Comparison of TP linear inversion models established based on the characteristic wavelength of the optimal correlation coefficient

反射率模型	回归方式	TP 反演模型	R^2 (TP)	TN 反演模型	R^2 (TN)
一阶微分(正相关)	指数	$y=0.0628e^{17245x}$	0.5778	/	/
	对数	$y=0.0652\ln(x)+0.8107$	0.4394	/	/
	一次幂	$y=2400.6x+0.0434$	0.5644	/	/
	二次幂	$y=5\times10^6x^2+1985.1x+0.0501$	0.5654	/	/
	三次幂	$y=1\times10^{12}x^3+2\times10^8x^2-3393.7x+0.1$	0.5977	/	/
比值法(正相关)	指数	$y=0.0163e^{1.8286x}$	0.6047	$y=9\times10^{-32}e^{70.562x}$	0.4369
	对数	$y=0.3047\ln x+0.1116$	0.6343	$y=73.528\ln x+0.0885$	0.4071
	一次幂	$y=0.267x-0.1576$	0.6498	$y=72.564x-72.469$	0.4070
	二次幂	$y=0.0227x^2+0.2127x-0.1262$	0.6501	$y=1576.7x^2+3267.4x-1690.9$	0.4102
	三次幂	$y=-0.6531x^3+2.4462x^2-2.7015x+1.0108$	0.6729	$y=-2\times10^6x^3+6\times10^6x^2-6\times10^6x+2\times10^6$	0.4406
比值法(负相关)	指数	$y=1.3279e^{-2.578x}$	0.5975	$y=8\times10^6e^{-16.74x}$	0.5235
	对数	$y=-0.334\ln x+0.1064$	0.6363	$y=-16.53\ln x+0.1363$	0.5049
	一次幂	$y=-0.3662x+0.475$	0.6078	$y=-17.542x+17.651$	0.5062
	二次幂	$y=0.6017x^2-1.4674x+0.9672$	0.6618	$y=139.22x^2+245.26x-106.34$	0.5157
	三次幂	$y=1.218x^3-2.7387x^2+1.5227x-0.0952$	0.6697	$y=15430x^3-43788x^2+41394x-13034$	0.5527

3.1.3 偏最小二乘法反演模型构建

1983 年，阿巴诺和伍德首次提出了一种新型的多元统计回归模型，它同时汲取了主成分分析方法、典型相关分析方法、多元线性回归分析方法的优点，建立了一种全新的回归分析模型，也就是偏最小二乘 PLS（Partial Least Squares）回归模型（杨燕明 等，2005）。PLS 不仅能够消除模型中变量的多重共线性，还能辨识模型中变量的系统信息与噪声。在水质参数监测模型的建立中，很好地消除了吸光度数据的连续相关性，提供了一种更为稳健和高效的建模方法。PLS 在原有的研究基础上加入了成分提取的思想，结合多元线性回归分析、典型相关分析和主成分分析，构建了一种既能够简化数据维度和数据结构，又能够解决多变量对多变量问题的回归模型，不仅降低了回归模型建立过程中对变量筛选和数据质量的要求，还解决了多因变量的回归建模难题。PLS 通过对模型变量的综合筛选，提取一些对模型中变量具有代表性的综合变量，再对提取出来的综合变量进行回归模型建立。

对白洋淀实验区的 62 组数据进行 TP、TN 浓度的特征波长提取。根据 3.1.2 章节的研究结果，反

射率比值与 TP、TN 含量的相关性好且建模效果优，因此提取比值法中 R 最大前 10 个波长参与后续的 PLS 法建模。而 R 有正负之分，因此提取的与 TP 浓度正相关对应的波长用 TP⁺代表，相关系数用 R^+ ；与 TP 浓度负相关对应的波长用 TP⁻代表，相关系数用 R^- 。TN 浓度相关的特征波长提取方法与 TP 一致。统计结果如下表 3 所示。

完成特征波长提取后，利用相关系数最大的波长完成后续的 PLS 反演模型建立。分别使用前 10 个最大正相关和负相关系数对应的波长比值进行 PLS 建模。反演模型的预测性能可用预测均方根误差和预测偏离度来进行评价。其中预测值与实测值的预测均方根误差 RMSEP（Root Mean Square Error of Prediction）表示两者的离散程度，可用来表征所建立的回归模型的预测能力，反映预测结果的整体精度。其值越小表示该反演模型的预测精度越高，RMSEP 可按式（2）进行计算：

$$RMSEP = \sqrt{\frac{\sum_{i=1}^n (y_{iP} - y_{iM})^2}{n}} \tag{2}$$

式中， y 的下标 P（prediction）和 M（measurement）分别代表实测值和预测值， n 是参与预测的样本数目。

表3 白洋淀TP、TN最优特征波长统计表
Table 3 Statistics table of TP, TN optimal characteristic wavelength in Baiyangdian

相关系数	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
TP ⁺	703/653	700/655	701/655	703/655	700/656	703/656	700/657	701/657	703/657	703/659
R^+	0.8058	0.8070	0.8063	0.8076	0.8062	0.8073	0.8064	0.8065	0.8077	0.8065
TP ⁻	657/697	656/698	657/698	659/698	660/698	656/700	657/700	659/700	660/700	657/701
R^-	-0.7789	-0.7778	-0.7794	-0.7787	-0.7779	-0.7782	-0.7800	-0.7792	-0.7788	-0.7775
TN ⁺	704/678	700/679	701/679	703/679	704/679	705/679	700/681	701/681	704/681	705/681
R^+	0.7478	0.7525	0.7550	0.7522	0.7549	0.7527	0.7483	0.7504	0.7507	0.7482
TN ⁻	618/608	621/608	615/610	696/697	679/700	694/700	696/700	679/701	686/701	679/704
R^-	-0.7202	-0.7203	-0.7204	-0.7269	-0.7206	-0.7196	-0.7221	-0.7217	-0.7192	-0.7194

预测偏离度，也称之为平均相对误差 ARE (Absolute Relative Error)，表示所有预测样本的预测值 y_{ip} 与实测值 y_{im} 之差的绝对值之和的均值。预测偏离度越大，说明模型预测能力越弱，反之则越强。ARE 的计算式 (3) 如下：

$$ARE = \frac{1}{n} \sum_{i=1}^n \frac{|y_{ip} - y_{im}|}{y_{im}} \times 100\% \quad (3)$$

(1) 设正相关系数对应的波长为 x_1 — x_{10} ，TP 或 TN 浓度变量为 y 。利用 PLS 法进行建立 TP 和 TN 的反演模型。得到的回归方程分别为

$$y_{TP} = -0.1624 - 13.8984 x_1 + 14.6031 x_2 + 4.7714 x_3 + 11.3264 x_4 - 8.6886 x_5 - 8.6886 x_6 + 4.0671 x_7 - 6.3795 x_8 - 0.3856 x_9 + 3.5356 x_{10}$$

$$y_{TN} = 0.9667 + 0.9445 x_1 - 10.7344 x_2 + 5.6788 x_3 - 10.7344 x_4 + 34.3033 x_5 - 19.6436 x_6 - 5.1196 x_7 + 6.2572 x_8 + 26.987 x_9 - 28.0811 x_{10}$$

由相关数据计算，利用 62 个样本的正相关系数进行 PLS 法 TP 反演建模，模型 R^2 为 0.7070，RMSEP 为 0.03515，总体的 ARE 为 23.79%。TN 反演建模，模型 R^2 为 0.6128，RMSEP 为 0.2267，总体的 ARE 为 18.38%。

(2) 设负相关系数对应的波长为 x_1 — x_{10} ，TP 或 TN 浓度变量为 y 。利用 PLS 法进行建立 TP 和 TN 的反演模型。得到的回归方程分别为

$$y_{TP} = 0.4926 + 0.9077 x_1 + 3.9384 x_2 - 2.5987 x_3 - 3.0923 x_4 + 5.0926 x_5 + 0.0787 x_6 - 5.8956 x_7 - 5.898 x_8 + 1.6347 x_9 + 5.4466 x_{10}$$

$$y_{TN} = 28.5781 - 0.6070 x_1 - 7.566 x_2 - 7.392 x_3 - 6.9867 x_4 - 17.6599 x_5 - 19.4375 x_6 - 7.1903 x_7 - 1.728 x_8 + 7.6424 x_9 + 16.7842 x_{10}$$

由相关数据计算可知，PLS 法 TP 反演建模，模型 R^2 为 0.6354，RMSEP 为 0.03921，总体的 ARE 为 27.21%。TN 反演建模，模型 R^2 为 0.6492，RMSEP 为 0.2158，总体的 ARE 为 17.03%。

(3) 选取 TP、TN 相关系数最大的正、负各 5 个对应的波长比值，进行 PLS 法反演模型建立。得到：

$$y_{TP} = 0.1195 + 0.5139 x_1 + 0.5904 x_2 - 2.3045 x_3 + 0.4377 x_4 + 0.9114 x_5 + 0.6337 x_6 + 1.8488 x_7 - 0.6015 x_8 - 2.512 x_9 - 0.4747 x_{10}$$

$$y_{TN} = 10.7848 + 16.8082 x_1 + 28.4899 x_2 - 23.3021 x_3 - 10.5821 x_4 - 10.5821 x_5 - 6.146 x_6 - 0.1199 x_7 - 9.8228 x_8 - 8.8769 x_9 - 5.594 x_{10}$$

PLS 法 TP 反演建模，模型 R^2 为 0.6688，RMSEP 为 0.03737 mg/L，总体的 ARE 为 23.83%。TN 反演建模，模型 R^2 为 0.6296，RMSEP 为 0.2218 mg/L，总体的 ARE 为 17.83%。

通过上述分析，分别使用单波段、一阶微分、比值、PLS 法进行了 TP、TN 反演建模。从预测值与实测值的数据对比结果来看，基于波长比值的 PLS 法的反演效果有显著提高，不论是正、负相关系数建立的反演模型，其决定系数 R^2 均大于 0.6 (图 4)，优于对应的单波长、一阶和比值法的线性回归模型，只有当线性回归模型达到 6 次幂时，其回归决定系数 R^2 才与 PLS 法的接近。TP、TN 反演模型的预测偏离度 ARE 控制在 20% 左右，具有良好的 TP、TN 浓度预测能力。

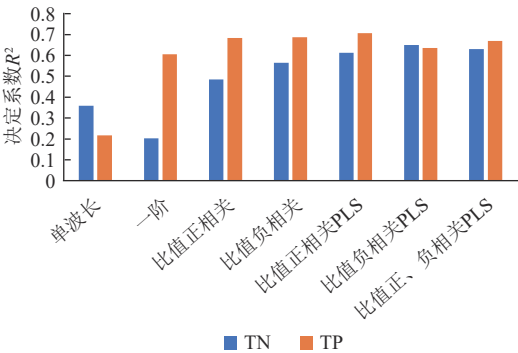


图4 不同反演模型决定系数 R^2 的对比示意图
Fig. 4 Comparison diagram of R^2 of different inversion models

因此,在后续的TP、TN反演模型的建立时,选用正相关系数的PLS法进行TP含量预测,选用负相关系数的PLS法进行TN含量预测。

3.2 样本浓度对建模的影响分析

实验室理想环境下的建模往往都先针对某个水质参数进行等间隔浓度的标准液配置,然后进行光谱测量,以此数据为样本进行建模。这样的方式虽有较好的反演效果,但在实际应用中存在两点不足:一是实际水体样本中所含的水质参数众多,其相互间的影响机理还有待进一步研究与论证,因而使用单参数标准液进行反演建模仅能讨论其方法的科学性和技术流程的可行性;二是

对于某个固定的研究区域,其水质参数的浓度往往变化不大,建模时仅能利用已有浓度范围内的数据进行。因此能否反演非样本浓度范围内的水质参数浓度,以便在水质参数发生突变时可进行在线监测并提供预警信息。所以本小节主要讨论面向含有多种水质参数的水体样本,在进行不同浓度范围的样本选取时,建立的反演模型是否会对非样本浓度范围内的水质参数进行预测。本次研究使用水质光谱在线检测系统在河北白洋淀地区实地采集62个点的光谱数据,对观测的62个点的水体样本进行了化学分析,得到其总磷、总氮的浓度值如下图5所示。

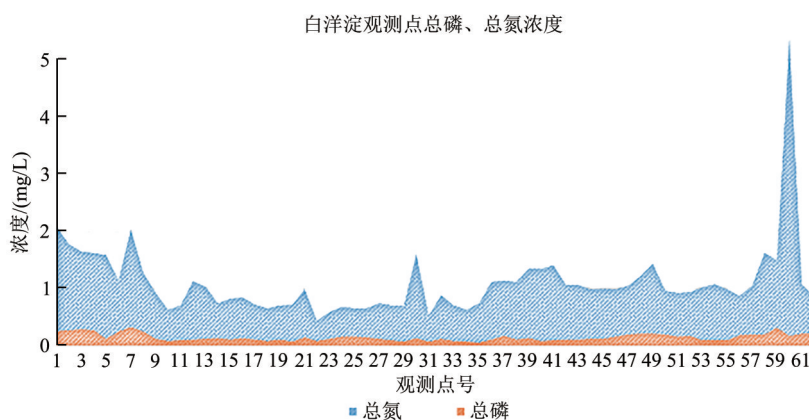


图5 白洋淀62个观测点TP、TN含量变化图

Fig. 5 Changes of TP and TN content at 62 observation points in Baiyangdian

TP的浓度范围为0.05—0.31 mg/L,均值0.156 mg/L,总体变化幅度平缓;TN浓度范围0.43—5.35 mg/L,但第60观测点的浓度远高于其余点,因此后续TN建模时不采用该异常点,其他观测点的浓度变化平缓,均值为1.03 mg/L。建模样本选取时,先将样本按照总磷或总氮的浓度大小进行排序,然后均匀分组,选取2/3的样本进行建模,参与建模的样本浓度涵盖了实验区总磷和总氮浓度范围。然后根据此模型对剩余的1/3样本进行总磷和总氮的浓度预测。

3.2.1 均匀浓度样本建模结果

对获取的62组数据,按照TP、TN的浓度分布,均匀选取41组数据进行比值法的最大正相关系数的PLS法建模,可称之为均值法。提取出的前10个最大正相关比值波段记为 x_1-x_{10} ,建立的TP回归模型 R^2 为0.7249。使用该模型对剩余的21组数

据进行TP含量估算,实测值与预测值的RMSEP为0.03815 mg/L,说明预测值随着浓度的增大,偏离度随之减小,总体的ARE为27.66%。21个预测点中,最大偏离发生在实测TP含量最小的0.05 mg/L的观测点,其原因可能由于TP本身含量较低,而水体中存在的浊度、叶绿素等其他水质参数对其反射率造成了影响,从而导致预测结果偏差较大。但随着TP浓度的增加至0.11 mg/L时,后续预测的ARE降至12.84%,表现出较好的预测能力。TN的模型 R^2 为0.6234,21组实测值与预测值的RMSEP为0.2457 mg/L,说明预测值随着浓度的增大,偏离度随之减小,总体ARE为23.01%。21个预测点中,最大偏离同样发生在TN含量最小的的观测点,其原因可能由于TN本身含量较低,而水体中存在的浊度、叶绿素等其他水质参数对其反射率造成了影响,从而导致预测结果偏差较大。若将该点去除,剩余20个观测点总氮浓度的ARE为

20.01%, 预测能力较强。对于TN异常点的估算值为1.34 mg/L, 误差较大。

3.2.2 高值浓度样本建模结果

对获取观测点数据按照TP、TN含量进行排序, 使用含量最大的前41组数据作为输入样本, 可称之为高值法。浓度范围为0.1—0.31 mg/L, 采用PLS法进行反演模型的建立, TP模型的 R^2 为0.6781, 使用该模型对剩余的21组数据的TP进行反演, 浓度范围为0.05—0.09 mg/L。实测值与预测值的RMSEP为0.04763, 总体ARE为60.59%。当预测浓度超过样本浓度值20%, 即0.08—0.09 mg/L时, ARE为45.07%。虽然预测偏差较大, 但预测样本的TP浓度为0.05—0.09 mg/L, 浓度本身和变化均非常小, 《GB 11893-89 水质总磷的测定 钼酸铵分光光度法》中的有关TP测定下限值为0.1 mg/L, 所以获取的光谱数据可能本身就不具有定量分辨如此小浓度的TP的能力。

对TN浓度进行排序发现有1组突变值(5.35 mg/L), 因此建模时不使用该组数据。使用TN浓度最大的前40组数据作为输入样本, 浓度范围为0.83—2.07 mg/L, PLS法反演模型的 R^2 为0.5805。使用该模型对剩余的22组数据的总磷进行反演, 浓度范围为0.43—0.83 mg/L和5.35 mg/L。21个样本点(除突变点外)的预测RMSEP为0.2822, 总体ARE为39.48%, 此时TN预测浓度超出建模浓度范围的50%; 当预测浓度超出建模浓度25%, 即0.64—0.83 mg/L时, ARE降至31.61%; 当预测浓度超出12%时, 即0.73—0.83 mg/L时, ARE为23.76%。对于异常突变点的预测值为1.19 mg/L, 偏差达78%, 说明对于超过一定建模样本浓度范围内的TN参数, 不具有反演可行性。

同时发现, TP和TN的预测值除个别点外均不同程度的大于实测值, 这说明当TP、TN浓度较低时, 水体中的浊度、叶绿素会对它们的敏感特征波长对应的反射率造成提升的影响, 导致预测结果偏差较大, 所以对于低浓度的水质参数进行反演前, 有必要开展相应的浊度等水质参数的补偿处理。

3.2.3 低值浓度样本建模反演

使用TP、TN浓度较小的前41组数据作为输入样本, 可称之为低值法。其中TP的浓度范围为0.05—

0.15 mg/L, 采用PLS建立的反演模型 R^2 为0.4105, 21组预测样本的浓度范围为0.15—0.31 mg/L, 实测值与预测值的RMSEP为0.09028, 总体ARE为36.96%, 此时预测的TP浓度超出了建模浓度的一倍; 当预测浓度超过建模浓度25%, 即0.15—0.18 mg/L时, ARE为35.44%; 超过50%时ARE为36.52%; 超过75%时ARE为34.8%。由此可见, 由TP低浓度样本建模预测高浓度时, 预测偏在35%左右, 并且所有的21个预测点的预测值均不同程度的小于实测值, 呈现出浓度越大偏离越大的趋势。其原因可能是由于水体样本种其他水质参数的存在, 增大了TP特征波长处的吸收度, 使其反射率降低, 最终导致预测值低于实测值。

TN建模输入样本的浓度范围为0.43—1.06 mg/L, 反演模型 R^2 为0.4632。使用该模型对剩余的浓度范围为1.06—2.07 mg/L和5.35 mg/L的22组数据的总磷进行反演, 21个样本点的RMSEP为0.5040, 总体ARE为28.31%。21个预测点中的最大浓度值已经超过建模样本TN浓度的一倍, 但仍具有30%以内的总体平均偏离度; 在超过建模样本浓度50%, 即1.06—2.07 mg/L范围时, ARE为22.84%, 说明在此范围内TN模型的反演能力更佳。对于异常突变点的预测值为0.98 mg/L, 这是由于该样本的TN浓度超过了建模样本浓度范围的5倍以上, 致使其反演效果不具有可行性。

3.2.4 最值浓度样本建模结果

使用TP浓度最大和最小的41组数据作为输入样本, 可称之为最值法。浓度范围为0.05—0.09 mg/L与0.16—0.31 mg/L, 建立的反演模型 R^2 为0.8146。对0.1—0.15 mg/L的21组数据进行反演可得RMSEP为0.03228 mg/L, 总体ARE为19.75%, 反演模型预测效果表现优良。并且21个观测点的预测值位于实测值的两侧, 呈现出随机误差的特性, 偏离值均值为仅为0.0043 mg/L。同时, 预测样本的TP浓度变化仅为0.05 mg/L, 说明基于最大、最小浓度样本建立的反演模型具有优于0.05 mg/L的分辨力。

使用TN浓度范围为0.43—0.81 mg/L和1.10—2.07 mg/L的前40组数据作为输入样本, PLS法建立的反演模型 R^2 为0.7902。对剩余的浓度范围为0.83—1.06 mg/L和5.35 mg/L的22组数据的进行TN反演, 除异常点外的21个样本的预测RMSEP为

0.3409 mg/L, 总体ARE为25.34%, 预测值位于实测值的两侧, 具有随机误差的特征, 21个点的偏离度均值为0.07368 mg/L, 整体预测效果较优。对21个预测值的统计发现, TN浓度变化仅为0.2 mg/L, 在此范围内建立的模型仍具有较好的反演效果, 说明水质光谱在线监测系统具有优于0.2 mg/L的TN浓度变化监测能力。对于异常突变点的预测值为1.2479 mg/L, 反演效果不理想。

3.2.5 中间浓度样本建模结果

使用TP浓度介于最大、最小之间的41组数据采用PLS法进行反演模型的建立, 可称之为中值法。该模型 R^2 为0.4573。对21组浓度范围为0.05—0.08 mg/L和0.2—0.31 mg/L的TP数据进行反演, RMSEP为0.06368, 总体ARE为45.80%; 而仅就高浓度范围0.2—0.31 mg/L进行TP反演, 其ARE可

迅速下降至27.46%, 说明影响预测准确度的样本主要集中在低浓度范围内。使用介于0.70—1.40 mg/L的40组数据建立的反演模型 R^2 为0.5589。对0.43—0.69 mg/L、1.42—2.07 mg/L和5.35 mg/L浓度范围的TN预测的结果显示, RMSEP等于0.4250, 总体ARE为35.52%, 21个点的偏离度均值为-0.1489 mg/L, 整体预测效果良好; 当预测浓度值控制在建模浓度值的20%范围内, 即0.56—1.68 mg/L时, ARE可降至28.93%; 对于异常突变点的预测值为1.0074 mg/L, 反演效果不理想。对TP和TN预测值的统计发现, 低浓度的均高于实测值, 而高浓度的均低于实测值, 且低浓度的预测偏离度要大于高浓度的偏离度, 再次验证了水体中其他水质参数对低浓度水质参数反演的影响。

基于上述5种不同样本建立模型的反演效果如下图6所示。

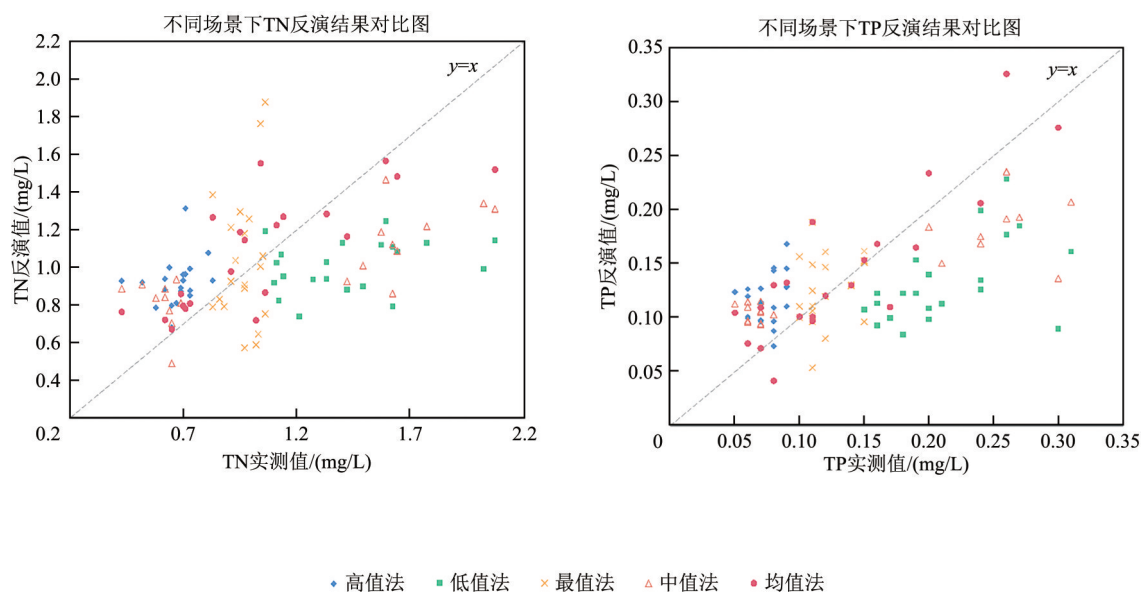


图6 不同浓度样本建模反演效果示意图

Fig. 6 Schematic diagram of the inversion model effect based on different concentrations

4 结论

本文基于实测的水样反射率光谱与水质参数含量数据, 在对比不同反演模型决定系数的基础上, 确定了使用PLS法可获得较好的TP和TN含量回归模型。针对实测建模样本浓度受地域影响变化幅度较小的问题, 系统深入地探讨了5种不同浓度范围样本构建反演模型的稳定性和预测准确性, 同时对超出建模样本浓度的预测值的置信度进行

了分析评价, 并给出了光谱法可能具有的最大TP、TN含量反演分辨率。具体结论如下:

(1) 水中TP、TN含量与波长反射率的比值表现出良好的相关性。具体来说, TP含量与波长反射率的相关性较弱, $R < 0.3$, 而与波长反射率比值的 R 可达0.8以上, 对应的特征波长位于657 nm和703 nm左右。TN含量与波长反射率的 R 最大仅为0.35; 与一阶微分的负相关性最好, 但后期建模效果不佳; 与反射率比值的正负相关性相当, R 值均为

0.49, 对应的敏感波长处于 611 nm 和 620 nm 左右。

(2) TP、TN 的建模效果均呈现 PLS 法优于反射率比值法和一阶微分法, 线性反演模型优于指数和对数模型, 随着线性自变量幂次的增加, 模型 R^2 随之增大, 一般 3 次幂后 R^2 趋于稳定。基于 PLS 法的 TP、TN 反演模型的 R^2 分别为 0.7070 和 0.6492, 远高于应用反射率其他运算值建立的反演模型的决定系数, 只有当 TP 的线性回归模型达到 6 次幂时, 才与之相当。

(3) 建模样本的水质参数含量不同时, 模型的 R^2 有最值法>均值法>高值法>中值法>低值法。均值法模型稳定性好, $R^2>0.6$, 预测的 ARE<20%, 预测浓度普遍大于实测浓度, 当 TP、TN 浓度值过小 (小于 0.1 mg/L) 时, 受浊度等其他水质参数及光谱探测器本身分辨力影响, 模型对低浓度的预测能力减弱。高值法模型 $R^2\approx 0.6$, 由于 TP 的预测浓度值和其范围本身较小, 光谱反射率变化量已不能反映微小浓度的变化, TP 反演的 ARE 基本为 50%; 当 TN 预测浓度的范围超过建模浓度范围的 12% 时, 其 ARE 可控制在 25% 以内。低值法的模型 R^2 介于 0.4—0.5, 预测样本浓度在超过建模样本浓度一倍时, 仍能保持 30% 左右的总体平均偏离度。预测浓度一般均小于实测浓度。最值法具有最佳的反演效果, R^2 可达 0.8, 预测值与实测值呈现出随机误差的特, ARE<25%; 对于总磷 0.05 mg/L 和总氮 0.2 mg/L 的微弱变化, 反演模型仍具有良好的预测效果, 说明基于此样本建立的模型不仅精度较高, 分辨力也较强。中值法模型的 R^2 大于 0.45 而小于 0.55, 低浓度的预测结果要大于实测值, 高浓度的预测结果要小于实测值, ARE 一般大于 35%。但当仅对高浓度样本进行反演时, ARE 可控制在 30% 以内。

在下一阶段的研究中, 可以尝试将本文提出的反演模型应用于机载、星载平台, 以实现水质监测“由点及面”的技术跨越。但在实际应用时, 不同平台数据的匹配融合、辐射和几何校正也是需要解决的首要问题, 只有当各种平台的数据都归一化至统一的框架下, 才能更好地发挥地面及实验室研究所建立的模型价值。

参考文献 (References)

Fuller F D and Ogilvie J P. 2015. Experimental implementations of two-dimensional fourier transform electronic spectroscopy. *Annual*

Review of Physical Chemistry, 66: 667-690 [DOI: 10.1146/annurev-physchem-040513-103623]

Golyandina N. 2010. On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Statistics and Its Interface*, 3(3): 259-279 [DOI: 10.4310/SII.2010.v3.n3.a2]

Gong S Q, Huang J Z, Li Y M, Lu W N, Wang H J and Wang G X. 2008. Preliminary exploring of hyperspectral remote sensing experiment for nitrogen and phosphorus in water. *Spectroscopy and Spectral Analysis*, 28(4): 839-842 (龚绍琦, 黄家柱, 李云梅, 陆皖宁, 王海君, 王国祥. 2008. 水体氮磷高光谱遥感实验研究初探. *光谱学与光谱分析*, 28(4): 839-842) [DOI: 10.3964/j.issn.1000-0593.2008.04.029]

Lin J Y and Zhang C X. 2019. Inversion of water quality parameters of urban river network using airborne hyperspectral remote sensing. *Remote Sensing Information*, 34(2): 23-29 (林剑远, 张长兴. 2019. 航空高光谱遥感反演城市河网水质参数. *遥感信息*, 34(2): 23-29) [DOI: 10.3969/j.issn.1000-3177.2019.02.004]

Lü H, Ma W C, Zhou L G, Tang L and Huai H Y. 2013. Estimation for water quality parameters in dianshan lake based on HJ-1A HSI image. *Journal of Fudan University (Natural Science)*, 52(2): 238-246 (吕航, 马蔚纯, 周立国, 汤琳, 怀红燕. 2013. 淀山湖微量水质参数卫星高光谱遥感估算. *复旦学报 (自然科学版)*, 52(2): 238-246) [DOI: 10.15943/j.cnki.fdxbs-jns.2013.02.015]

Ning H Z, Tan X, Li Y H, Jiao Q B and Li W H. 2020. Joint space-spectrum SG filtering algorithms for hyperspectral images and its application. *Spectroscopy and Spectral Analysis*, 40(12): 3699-3704 (宁鸿章, 谭鑫, 李宇航, 焦庆斌, 李文昊. 2020. 空一谱维联合 Savitzky-Golay 高光谱滤波算法及其应用. *光谱学与光谱分析*, 40(12): 3699-3704) [DOI: 10.3964/j.issn.1000-0593(2020)12-3699-06]

Rixon S, Levison J, Binns A and Persaud E. 2020. Spatiotemporal variations of nitrogen and phosphorus in a clay plain hydrological system in the Great Lakes Basin. *Science of the Total Environment*, 714: 136328 [DOI: 10.1016/j.scitotenv.2019.136328]

Tong Q X, Zhang B and Zhang L F. 2016. Current progress of hyperspectral remote sensing in China. *Journal of Remote Sensing*, 20(5): 689-707 (童庆禧, 张兵, 张立福. 2016. 中国高光谱遥感的前沿进展. *遥感学报*, 20(5): 689-707) [DOI: 10.11834/jrs.20166264]

Wang Y X. 2017. Study on the Water Quality of Qinghe Reservoir Inversion Based on Landsat8. Shenyang: Shenyang Agricultural University (王云霞. 2017. 基于 Landsat8 的清水水库总氮、总磷反演研究. 沈阳: 沈阳农业大学)

Xia R and Chen Z. 2015. Integrated water-quality assessment of the huai river basin in China. *Journal of Hydrologic Engineering*, 20(2): 05014018 [DOI: 10.1061/(ASCE)HE.1943-5584.0001030]

Yang Y M, Liu Z W, Chen B Q and Tang J W. 2005. Retrieval of oceanic color constituents from case II water reflectance by partial least squares regression. *Journal of Remote Sensing*, 9(2): 123-130 (杨燕明, 刘贞文, 陈本清, 唐军武. 2005. 用偏最小二乘法反演二类水体的水色要素. *遥感学报*, 9(2): 123-130) [DOI: 10.3321/j.issn.1007-4619.2005.02.003]

Yu Z Y. 2019. A Study on Eutrophication Evaluation of Lake Based on Hyper-Spectral Remote Sensing—A Case on Lake Hongfeng. Guiyang: Guizhou Normal University (余祝媛. 2019. 基于高光谱遥感的湖泊富营养化评价研究——以贵阳市红枫湖为例. 贵阳: 贵州师范大学)

Zhai X Y, Xia J and Zhang Y Y. 2017. Integrated approach of hydrological and water quality dynamic simulation for anthropogenic dis-

- turbance assessment in the Huai River Basin, China. *Science of the Total Environment*, 598: 749-764 [DOI: 10.1016/j.scitotenv.2017.04.092]
- Zhang C, Shan B Q, Zhao Y, Song Z X and Tang W Z. 2018. Spatial distribution, fractionation, toxicity and risk assessment of surface sediments from the Baiyangdian Lake in northern China. *Ecological Indicators*, 90: 633-642 [DOI: 10.1016/j.ecolind.2018.03.078]
- Zou X H, Wu Y B, Zhu X C, An Y T, Zhou Z Y and Zhu J X. 2020. Assessment of nutritional status of water bodies in Jiangsu Dafeng milu national nature reserve based on hyperspectral reflectance. *Wetland Science*, 18(2): 219-227 (邹兴红, 吴永波, 朱晓成, 安玉亭, 周子尧, 朱嘉馨. 2020. 基于高光谱反射率的江苏大丰麋鹿国家级自然保护区水体营养状态评价. *湿地科学*, 18(2): 219-227) [DOI: 10.13248/j.cnki.wetlandsci.2020.02.011]

Comparative study on the hyperspectral estimation models of TP and TN in Baiyangdian water body

CHEN Jie^{1,2,3}, ZHANG Lifu¹, ZHANG Hongming¹, ZHANG Linshan^{1,2}, CEN Yi¹, TONG Qingxi¹

1. State Key Laboratory of Remote Sensing Science, Aero Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing 100083, China

Abstract: Total Phosphorus (TP) and Total Nitrogen (TN) are important indicators of water quality eutrophication and the main parameters for water quality monitoring. Water quality monitoring by spectroscopy has become a hot spot in the current remote sensing water environment research because it is rapid and efficient and has no secondary pollution. The usual TP and TN inversion models are established based on the laboratory configuration standard solution for spectral measurement or the modeling based on the full samples. The model constructed in this way has a good regression effect. However, the actual water body causes the mutual influence of various water quality parameters, the concentration distributions of TP and TN are not uniform, and the predicted value may exceed the training sample range, making the actual prediction effect always unsatisfactory.

In this study, the actual water samples in the Baiyangdian area are used as the input values of the inversion model. First, the measured spectral data and the chemical analysis values of TP and TN are used to compare the relationship between the correlation values of different reflectances and water quality parameters. Various inversion models have been constructed for the best relevant bands. The most stable and accurate modeling method has been determined through comparison. Therefore, the modeling samples are divided into uniform, high-value, low-value, median-value, and max-min-value samples according to concentration. Then, the influence of the sample modeling with different concentration ranges on the inversion model is discussed. The model's predictive ability for samples with concentration values beyond modeling is determined.

The extraction results of the characteristic wavebands in the range of 400—1000 nm indicate that the reflectance correlation coefficient of TP and TN corresponding to a single wavelength is less than 0.3, which is not high; the maximum correlation coefficient with the first-order value of reflectance is 0.76, which is a moderate correlation; the correlation coefficients with the reflectance ratio are all over 0.8, which is highly correlated. In the inversion effect of the linear regression model, the exponential method and the logarithmic method are inferior to the multiple power method. Moreover, the effect of high power is better than low power. However, the overall effect is not ideal. The model's R^2 is less than 0.6. When the range of the modeling sample concentration is different, the R^2 of the model is also different. The result is as follows: max-min method > uniform method > high-value method > middle-value method > low-value method. When the modeling sample concentration covers the predicted sample, the inversion model determination coefficient is $R^2 > 0.6$. The average deviation of the predicted value (ARE) of TP and TN concentrations is less than 20%. When the modeled concentration is higher than the predicted sample, the R^2 is approximately 0.6. The predicted value within 12% of the overconcentration range has an ARE of <25%. When the modeling concentration is lower than the predicted value, the R^2 is between 0.4 and 0.5. The ARE is $\leq 30\%$ when the predicted value exceeds the modeling sample concentration. When the sample concentration is on both sides of the predicted value, R^2 can reach 0.8, and ARE is <25%. The following is obtained when the modeled sample concentration is between the predicted values: $0.45 < R^2 < 0.55$; ARE > 35%.

When the reflectance method for TP and TN inversion was used, the ratio method can be given priority to the characteristic band when modeling. The regression effect of the partial least square method is significantly better than that of multiple power and exponential models. The model also has a clear physical meaning. Thus, it can be used in the regression study of TP and TN based on reflectance. For the predicted value that is not within the range of the modeled sample concentration, the credibility of the inversion results can be judged based on the relative relationship between its concentration value and the modeled sample concentration value.

Key words: water quality monitoring, total phosphorus, total nitrogen, concentration inversion, partial least square method

Supported by National Natural Science Foundation of China (No. 41830108, 41977154); XPCC Major Science and Technology Projects (No. 2018AA004); XPCC Innovation Team in Key Areas (No. 2018CB004); The National Key Research and Development Program of China (No. 2017YFC1500901)