第42卷，第7期
2022年7月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 7, pp2181-2186
July, 2022

透射光谱的水体亚硝酸盐含量模拟估算

王彩玲[1]，王 波[2]，纪 童[3]，徐 君[4]，剧 锋[5]，王洪伟[6*]

1. 西安石油大学计算机学院，陕西 西安 710065
2. 盐池县草原实验站，宁夏 盐池 751506
3. 甘肃农业大学草业学院，甘肃 兰州 730070
4. 西安航空学院，陕西 西安 710077
5. 中华人民共和国银川海关，宁夏 银川 750000
6. 西北工业大学光电与智能研究院，陕西 西安 710072

摘 要 亚硝酸盐是水体中重要的必测指标之一，对于水体质量的评估有着重要意义。但传统的检测方法操作复杂、受干扰因素多、测定时间长、不能及时反映水质变化、无法及时有效地预警突发水污染事件。鉴于此，探索准确、实时、环保的环境水体和饮用水中的亚硝酸盐含量检测办法具有重要意义。采用优级纯试剂配制10种浓度的亚硝酸盐氮标准溶液(0.02，0.04，0.06，0.08，0.10，0.12，0.14，0.16，0.18和0.20 mg·L^{-1})，采用OCEAN-HDX-XR微型光纤光谱仪扫描10次各浓度亚硝酸盐溶液在181.1～1 023.1 nm范围内的透射光谱，取平均值作为各浓度亚硝酸盐溶液原始透射光谱，之后以亚硝酸盐含量作为因变量，全波段原始透射光谱作为自变量，采用随机森林回归中特征变量重要性方法，筛选特征变量，再此基础上利用交叉验证法，挑选最为稳定的模型变量个数，建立亚硝酸盐优化随机森林反演模型。结果如下：(1)利用全波段建立的随机森林模变量解释率(var explained)＝76.49%，均方残差(mean of squared residuals)＝0.000 688；(2)随机森林变量重要性方法筛选对亚硝酸盐反演的敏感波段，其中195.1 nm重要性值最高，并利用留一交叉法发现，当利用19个光谱特征变量时随机森林模型的均方根误差最低，以筛选光谱特征变量建立的优化随机森林变量解释率(var explained)＝83.45%，均方残差(mean of squared residuals)＝0.000 552。变量筛选有效减少了光谱数据量，对优化模型的建立提供了基础；(3)对建立模型进行模型检验，其中全波段随机森林模型测试集R^2＝0.820 3，RMSE＝0.03，检验集R^2＝0.979 3，RMSE＝0.01，优化随机森林模型测试集R^2＝0.873 4，RMSE＝0.022，检验集R^2＝0.979 8，RMSE＝0.008，对比全波段随机森林模型与优化后随机森林模型后发现，优化随机森林模型测试集与检验集模型解释度、模型精度均要高于全波段随机森林模型，说明优化方法不仅可有效降低光谱维度，对于寻找亚硝酸盐光谱敏感波段，建立精度较高的亚硝酸盐反演模型有着积极意义。基于以上试验结果，提出了一种优化随机森林模型高光谱水质亚硝酸盐参数的反演方法，为水质亚硝酸盐参数动态检测提供了新方法。

关键词 高光谱；亚硝酸盐；模型；随机森林

中图分类号：P237 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)07-2181-06

引 言

随着人类物质生活水平的提高和工业化的发展，水污染已经成为当今社会普遍存在的问题，其监测与治理也备受关注。在pH＜6.5时亚硝酸盐会与仲氨反映生成具有强致癌性的亚硝胺基，是水质监测的必测指标之一[1]。“分光光度计法”、紫外-分光光度法为现下普遍接受的测定亚硝酸盐指标的方法，但测定时间长、不能及时反映水质变化，不适合现场监测[2]。

收稿日期：2021-08-19，**修订日期**：2022-02-07

基金项目：国家自然科学基金项目(31160475，61401439)，陕西省重点研发计划项目(2019GY-112)，西安航空学院高教研究项目(2018GJI005)，陕西西安教育科学“十三五”规划2018年度课题(SGHI8H435)资助

作者简介：王彩玲，女，1984年生，西安石油大学计算机学院副教授 e-mail：azering@163.com

*通讯作者 e-mail：whwdyx@163.com

原始光谱反射数据有着数据量大，指标彼此高度相关的特性，原始指标高度相关的特性经常会导致多重共线性问题的产生，从而导致模型失真[3]；因此如何对大量光谱数据进行处理和挑选一直是光谱反演的重点。随机森林(random forest，RF)作为常用机器学习算法在分类、指标反演、筛选指标上应用广泛[4]，国内许多学者将随机森林等机器学习新方法作为典型计量模型的代表广泛应用到水质预测领域，促进水质分析向多参数测试趋势发展。张颖等[5]利用随机森林分类算法对巢湖区域水质进行类别判定，监测断面水质分类准确率可达 96.15%；吴志明等[6]基于随机森林对太湖湖泊水体有色可溶性有机物(CDOM)浓度进行遥感估算，根据随机森林算法的特征重要性参数提供的各自变量影响力结果，发现 709 和 560 nm 波段贡献率最大，是反演 CDOM 的敏感波段，并建立了精度较高的随机森林反演模型；

现有文献报道中，利用透射光谱估测水质参数亚硝酸盐指标的报道较少；基于此，试验利用光谱数据进行水体指标亚硝酸盐的反演，测定水体样本的光谱数据，将采集到的光谱数据与标液亚硝酸盐含量建立亚硝酸盐随机森林反演模型，由于光谱指标之间的高度相关，为避免模型失真，在建立反演模型之前，利用随机森林变量重要性法挑选敏感光谱指标，并将筛选指标利用留一交叉法进一步筛选，最终利用筛选的变量组合建立亚硝酸盐随机森林反演模型，比较全波段(未筛选)与优化(筛选变量)随机森林模型精度，选出更加适合反演亚硝酸盐指标的建模方法。探索利用高光谱估测水体亚硝酸盐含量的可行性与最优方法，为实时诊断水体状况提供关键技术与可行的途径。

1 实验部分

1.1 供试亚硝酸盐标液

称取在 105～110 ℃下烘干约 4 h 的亚硝酸钠($NaNO_2$) 0.492 8 g 溶于水，准确定容至 1 000 mL，此溶液含 NO_2-N 100 mg·L^{-1}。实验前，用移液管吸取此溶液 20.00 mL 用水稀释至 1 000 mL，此溶液含 NO_2-N 0.2 mg·L^{-1}。用此方法配制 0.02，0.04，0.06，0.08，0.10，0.12，0.14，0.16，0.18 和 0.20 mg·L^{-1}的亚硝酸盐标液[7]。

1.2 光谱仪参数

试验用仪器为 Ocean Optics 公司出品的 OCEAN-HDX-XR 微型光纤光谱仪，该光谱仪采用高清晰度光学系统，具有高通量、低杂散光和高热稳定性的特点，适用于精确测量溶液中的分析物，具有体积小，容易集成到许多工业应用的生产过程环境的优势。仪器参数见表 1。

表 1 光谱仪参数

Table 1 Spectrometer parameters

探测器	光谱范围/nm	光学分辨率/nm	信噪比	尺寸/mm	杂散光/AU	操作温度/℃	狭缝接口
Back-thinned CCD	181.1～1 030.1	0.61～0.9	400∶1	88.9×63.5×52.4	＞3	0～40	SMA 905

1.3 光谱数据获取

样品为 0.02，0.04，0.06，0.08，0.10，0.12，0.14，0.16，0.18 和 0.20 mg·L^{-1}的亚硝酸盐标液，光谱仪狭缝为 10 μm，相同时间间隔重复采集十次上述标液 181.1～1 030.1 nm 范围内的高光谱透射率数据，共计得到 100 条光谱数据。

采用白板校正分别得到所采集的高光谱数据的光谱透射率值[8]，如式(1)所示

$$TC = TO/TW \quad (1)$$

式(1)中：TC 为光谱透射率，TO 为原始光谱数据，TW 为白板数据。

1.4 数据处理

随机森林(RF)算法[9]结构清晰、易于解释、运行效率高，对于数据要求低，且具有很好的抗噪声能力，能够处理高维度数据，训练速度快，泛化能力强，比较容易实现并行计算，不易出现过拟合问题。随机森林模型的建立通过调用 R 语言中“randomForest”程序包[10]来实现。该方法首先完成两个随机采样过程，即通过自助法重采样技术有放回的在 100 组训练数据中重复随机抽取 67 个训练样本(总样本容量的三分之二)，未被抽取到的数据被称为“袋外”(outofbag)数据。

随机森林模型建立时有两个重要参量[11]，分别为随机森林决策树数目(mtry)与指定节点中用于二叉树的变量个数(ntree)，其中 mtry 一般取值为变量的二次方根，ntree 的取值需要逐一尝试，当模型内误差稳定时，即为 ntree 数值。

模型评价方面，通过计算解释方差百分比(% Var explained)与模型拟合精度(R^2)来评定模型稳定能力与预测能力。

2 结果与讨论

2.1 亚硝酸盐原始透射光谱

图 1 为 10 种浓度亚硝酸盐原始透射光谱，从图中可以看出不同浓度溶液的亚硝酸盐光谱曲线的趋势类似，在紫外波段 180.1～400 nm 亚硝酸盐光谱曲线呈先下降后上升的趋

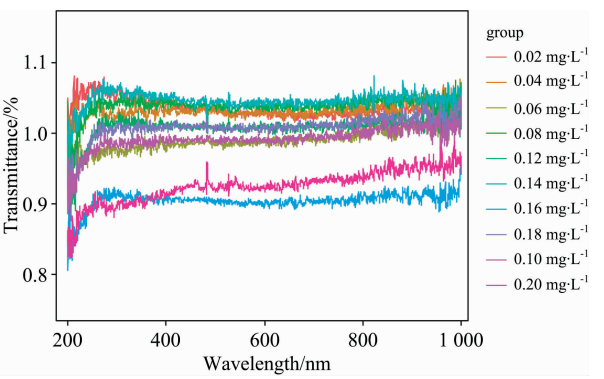

图 1 原始透射光谱图

Fig. 1 Original transmission spectra

势，光谱曲线波谷分布于 185～197 nm 范围内，且谱线均在在紫外短波段有强吸收，图中在 210 nm 波长周围处有极大的吸收峰，浓度不同峰的高度也有所不同，主要表现为随着亚硝酸盐含量的增加，亚硝酸盐在各波段的光谱透射率逐渐降低。

2.2　随机森林反演模型

原始光谱共有 2 049 个变量，对所有光谱变量进行随机森林建模，其中参数 ntree 设定为 500，mytry 设定为 40，随机森林反演模型参数见表 2，其中残差平方均值为 0.000 69，变量解释率为 76.49%。拟合结果见图 2 训练集(train)，其中拟合精度(R^2)为 0.820 3，均方根误差为 0.03，说明随机森林模型对于水体亚硝酸盐含量能够做出很好的预测。

利用测试集 test，对建立的随机森林模型进行模型检验，检验结果见图 2，通过对预测值与真实值进行线性拟合，进行模型检验，R^2＝0.979 3，RMSE＝0.01，说明建立的随机森林模型有着很强的预测能力。

表 2　随机森林模型参数

Table 2　Spectrometer parameters

Tree	Mytry	Mean of squared residuals	%Var explained	R^2	RMSE
500	40	0.000 69	76.49	0.820 3	0.03

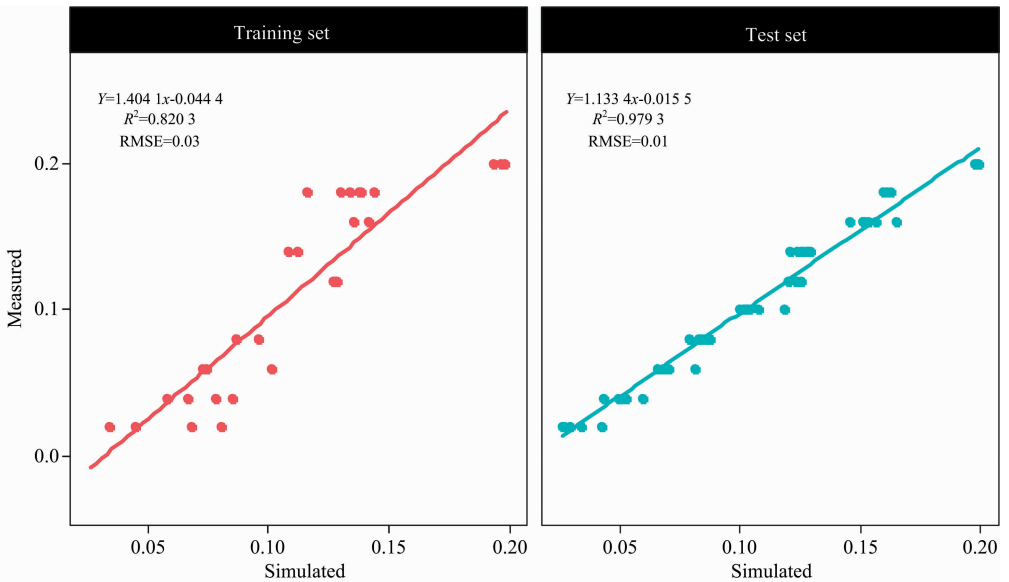

图 2　全波段随机森林模型在测试集与训练集的预测结果

Fig. 2　The prediction results of the test set and training set using the full-band random forest model

2.3　随机森林变量重要性

原始光谱数据量繁杂，变量间存在多重共线性问题，研究亚硝酸盐光谱敏感波段，对于分析水体亚硝酸盐光谱特征，降低光谱冗余，以及提升模型精度有着重要意义。随机森林算法中变量重要性算法，可以分析各个自变量对因变量的影响程度，以方差增量(IncMSE)指标来定性表征[12]。方差增量指将某一变量替换成随机变量后对预测结果造成的影响，若用于替换的随机变量显著改变了方差，则认为原变量重要性很高。在建立全波段随机森林模型过程中得出的随机森林变量重要性结果如图 3 所示；25 个光谱变量(IncMSE≥3)中 195.1 nm 变量重要性最高，IncMSE 值为 4.6，说明 195.1 nm 波段对反演水体亚硝酸盐含量有着重要作用。

2.4　优化随机森林模型

按照变量重要性大小，将指标由大到小依次输入随机森林模型，并采用交叉验证方法比较输入不同变量时模型均方误差的大小，结果如图 4 所示，发现模型输入变量为 19 个时，模型均方误差值最低(RMSE＝0.02)，且随变量数增多，模型均方误差趋于稳定，故选用筛选出的 19 个光谱变量作为优化随机森林模型的初始变量。

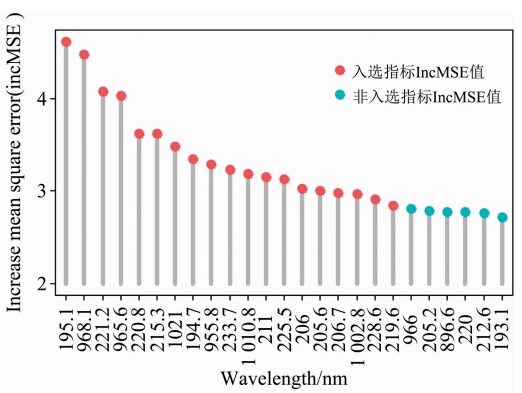

图 3　随机森林变量重要性(IncMSE)图

Fig. 3　Random forest variable importance (IncMSE) graph

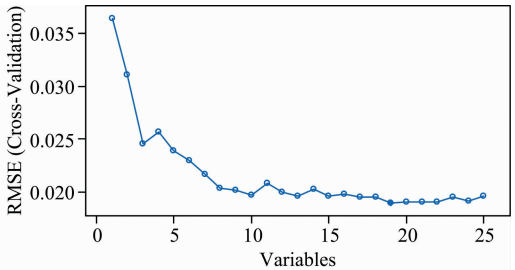

图 4 交叉验证

Fig. 4 Cross-validation

利用筛选出的 19 个光谱变量进行随机森林建模，其中参数 ntree 设定为 500，因参与建模的光谱变量仅有 19 个，因此 mytry 设定为 4，随机森林反演模型参数见表 3，其中残差平方均值为 0.000 55，变量解释率为 83.45%，拟合结果见图 5 训练集(training set)，其中拟合精度(R^2)为 0.873 4，均方根误差(RMSE)为 0.022，说明优化随机森林模型对于水体亚硝酸盐含量能够做出很好的预测。

表 3 优化随机森林模型参数

Table 3 Optimize random forest model parameters

Tree	Mytry	Mean of squared residuals	%Var explained	R^2	RMSE
500	4	0.000 55	83.45	0.873 4	0.022

利用袋测试集 test，对建立的随机森林模型进行模型检验，检验结果见图 5，通过对预测值与真实值进行线性拟合，进行模型检验，R^2=0.9798，RMSE=0.008，说明建立的随机森林模型有着很强的预测能力。

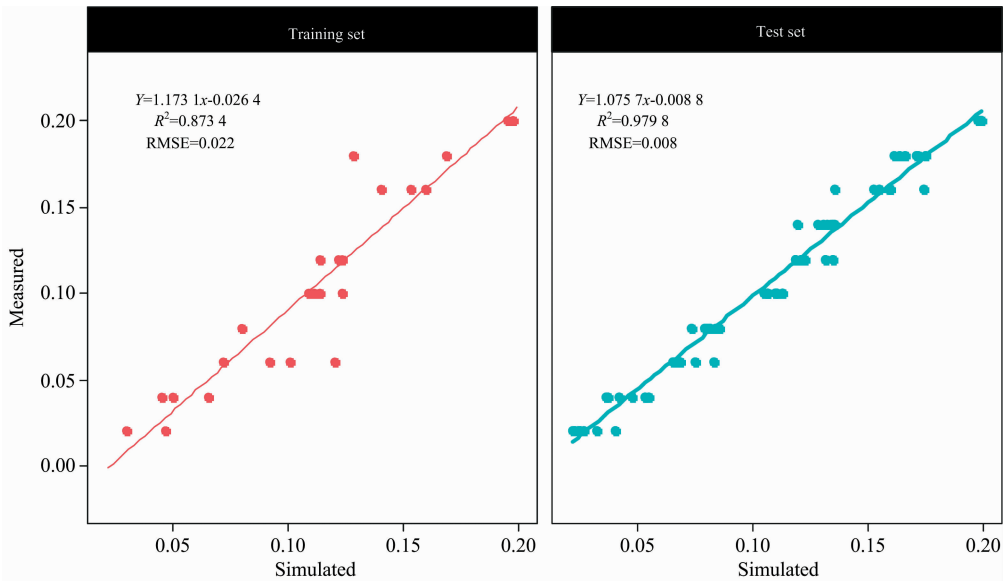

图 5 优化随机森林模型在测试集与训练集的预测结果

Fig. 5 The prediction results of the test set and training set of the random forest model

2.5 模型精度对比

通过对比全波段随机森林模型与优化随机森林模型参数，挑选最为适合监测水体亚硝酸盐的光谱反演方法，模型参数结果见表 4。

表 4 模型参数对比

Table 5 Model accuracy test

模型	Mean of squared residuals	%Var explained	R^2	RMSE
全波段随机森林模型	0.000 69	76.49	0.820 3	0.03
优化随机森林模型	0.000 55	83.45	0.873 4	0.022

从表 4 可以看出，优化随机森林模型在各项指标上均优于全波段随机森林模型，方差解释率增加了 7 个百分点，且优化随机森林模型建模变量要远低于全波段建模变量，大大提高了机器学习的运算速率，降低了数据的冗余度，说明提取特征波段对水体中亚硝酸盐含量进行预测可以大大减少干扰信息的影响，提高预测模型的性能，可适用于水体亚硝酸盐含量的反演。

3 结 论

物质的光谱强度与物质的组成成分和性质之间存在一定的联系，从而可以建立光谱强度与样品含量之间的关系模型。基于透射光谱研究水体亚硝酸盐含量的研究较少，多在紫外吸收光谱中研究，其中硝酸盐氮(NO_3-N)的紫外吸收峰在 202.0 nm 左右，而亚硝酸盐氮(NO_2-N)的紫外吸收峰在 210 nm 左右[7]。在建立全波段随机森林模型时，利用随机森林变量重要性得出 191.5，968.1 和 221.2 nm 等 19 个重要性较高变量，得出的波段与亚硝酸盐氮(NO_2-N)的紫外吸收峰 210nm 结果相近。

利用一种优化后的随机森林模型方法进行水体亚硝酸盐

指标的反演，通过随机森林变量重要性法筛选的光谱指标，并利用交叉验证法进一步缩小了变量个数，建立了优化随机森林模型，优化后随机森林模型具有以下优点：(1)通过波长或波长区间选择，可以有效减少参与建模的自变量数量，从而简化模型，降低建模预测时的计算量；(2)对待测组分具有光谱特征的波段处的信息进行提取强化，同时弱化待测组分吸收不明显或干扰物质影响显著的波段，以此提升模型的预测精度；(3)消除或减弱由于仪器和环境带来的噪声以及谱线中存在的冗余信息对回归建模的影响。

优化随机森林模型不仅模型精度，稳定性、预测能力显著高于全波段随机森林模型，而且有效降低了光谱数据维度，综合了有效波段的光谱特性。结果表明本优化方法，模型精度较高，可适用于反演水体亚硝酸盐含量反演。

以上试验结果为水质亚硝酸盐指标的快速估算提供了理论基础，为水体质量评估提供更便利的方案。

References

[1] SUN Dai-hua, GOU Xiao-dong(孙代华, 勾效东). Shandong Environment(山东环境), 2000, (S1): 71.

[2] LUO Ji-yang, WEI Biao, TANG Bin, et al(罗继阳, 魏　彪, 汤　斌, 等). Environmental Science & Technology(环境科学与技术), 2015, 38(S2): 246.

[3] GONG Cai-lan, YIN Qiu, KUANG Ding-bo, et al(巩彩兰, 尹　球, 匡定波, 等). National Remote Sensing Bulletin(遥感学报), 2006, (6): 910.

[4] XING Xiao-yu, YANG Xiu-chun, XU Bin, et al(邢晓语, 杨秀春, 徐　斌, 等). Geo-Information Science(地球信息科学学报), 2021, 23(7): 1312.

[5] ZHANG Ying, GAO Qian-qian(张　颖, 高倩倩). Chinese Journal of Environmental Engineering(环境工程学报), 2016, 10(2): 992.

[6] WU Zhi-ming, LI Jian-chao, WANG Rui, et al(吴志明, 李建超, 王　睿, 等). Journal of Lake Sciences(湖泊科学), 2018, 30(4): 979.

[7] WANG Jing-min, ZHANG Jing-chao, ZHANG Zun-ju(王静敏, 张景超, 张尊举). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(1): 161.

[8] ZHANG Xue, ZHENG Xiao-shen(张　雪, 郑小慎). Ocean Technology(海洋技术学报), 2018, 37(6): 79.

[9] XING Xiao-yu, YANG Xiu-chun, XU Bin, et al(邢晓语, 杨秀春, 徐　斌, 等). Geo-Information Science(地球信息科学学报), 2021, 23(7): 1312.

[10] ZHAO Bei-geng(赵北庚). Computer CD Software and Applications(计算机光盘软件与应用), 2015, 18(2): 152.

[11] WANG Teng-jun, FANG Ke, YANG Yun, et al(王腾军, 方　珂, 杨　耘, 等). Bulletin of Surveying and Mapping(测绘通报), 2021, (11): 92.

[12] YOU Jie-wen, ZOU Bin, ZHAO Xiu-ge, et al(游介文, 邹　滨, 赵秀阁). China Environmental Science(中国环境科学), 2019, 39(3): 969.

Simulated Estimation of Nitrite Content in Water Based on Transmission Spectrum

WANG Cai-ling[1], WANG Bo[2], JI Tong[3], XU Jun[4], JU Feng[5], WANG Hong-wei[6]*

1. College of Computer Science, Xi'an Shiyou University, Xi'an 710065, China
2. Grassland Experiment Station of Yanchi, Yanchi 751506, China
3. College of Grass Industry, Gansu Agricultural University, Lanzhou 730070, China
4. Xi'an Aeronautical University, Xi'an 710077, China
5. Yinchuan Customs District P. R. China, Yinchuan 750000, China
6. School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an 710072, China

Abstract NO_2-N is an important parameter in water bodies and can quickly detect organic pollution parameters. It is of great significance to the assessment of water quality. However, traditional methods are complicated in operation, subject to many interference factors, long measurement time, cannot reflect water quality changes in time, and cannot provide timely and effective early warning. For sudden water pollution incidents, because of the shortcomings of traditional methods, it is of great significance to explore accurate, real-time, and environmentally friendly detection methods for the NO_2-N content in environmental water bodies and drinking water. This experiment is to study the use of superior grade pure reagents to prepare 10 concentrations of NO_2-N nitrogen standard solutions (0.02, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18 and 0.2 mg·L^{-1}), using the OCEAN-HDX-XR micro-fiber spectrometer to scan 10 times the transmission spectrum of the NO_2-N solution of each concentration in the range of 181.1～1 023.1 nm. Take the average value as the original transmission spectrum of the NO_2-N solution of each concentration, and then take the NO_2-N content of the solution as the dependent variable and the original transmission spectrum as the independent variable. Use the method of variable feature importance in random forest regression to screen the feature variables. Based on the cross-validation method, the number of the most stable model variables is selected, and the NO_2-N optimization random forest inversion model is established. The results of the study are as follows: (1) The variable explained rate (Var Explained) of the random forest model established by the whole band (Var Explained)=76.49%, and the mean squared residuals (Mean of squared residuals)=0.000 688; In the sensitive band of salt inversion, 195.1 nm has the highest importance value, and the leave-one-out crossover method is used to find that the random forest model has the lowest root mean square error when 19 spectral characteristic variables are used to screen the optimized random forest established by spectral characteristic variables Variable Explained rate (Var Explained)=83.45%, Mean of squared residuals (Mean of squared residuals)= 0.000 552. Variable screening effectively reduces the amount of spectral data and provides a basis for the establishment of the optimization model; (3) Model verification of the established model, including the full-band random forest model test set R^2=0.820 3, RMSE=0.03, test set R^2=0.979 3, RMSE=0.01, optimized random forest model test set R^2=0.873 4, RMSE=0.022, test set R^2=0.979 8, RMSE=0.008, after comparing the full-band random forest model with the optimized random forest model, it is found that the optimized random forest model test set and test The interpretation and accuracy of the set model are higher than the full-band random forest model, indicating that the optimization method can not only effectively reduce the spectral dimension, but also has positive significance for finding the sensitive band of NO_2-N spectrum and establishing a high-precision NO_2-N inversion model. . Based on the above test results, an inversion method for optimizing the hyperspectral water quality NO_2-N parameters of the random forest model is proposed, which provides a new method for the dynamic detection of water quality NO_2-N parameters.

Keywords Hyperspectral; Nitrite; Model; Random forest

(Received Aug. 19, 2021; accepted Feb. 7, 2022)

* Corresponding author