文章编号:1671-6833(2022)06-0090-07

# 优化随机森林算法的城市湖泊 DOC 质量浓度遥感反演

李爱民[1], 王海隆[2], 许有成[2]

(1.郑州大学 地球科学与技术学院,河南 郑州 450001;2.郑州大学 水利科学与工程学院,河南 郑州 450001)

**摘 要**: 在城市湖泊的可溶性有机碳(DOC)含量的遥感监测问题中,传统回归模型难以描述非线性关系而不能满足精度的要求,因此,将贝叶斯优化算法引入到随机森林模型的参数优化中,提出一种贝叶斯优化随机森林模型(BO-RF)的城市湖泊 DOC 质量浓度反演方法。以郑州市天德湖水域为例,基于高时空分辨率的 Planet 卫星影像数据和实测的 DOC 水质数据,开展城市湖泊 DOC 质量浓度的遥感反演方法研究。PEARSON 相关性分析结果表明:反演 DOC 质量浓度的 Planet 卫星影像波段最佳波段组合为 B2/B4。利用传统回归方法得到的波段比值模型决定系数 $R^2=0.466$,均方根误差 $RMSE=0.515$ mg/L,无法满足精度要求。利用支持向量机和 BP 神经网络建模精度有所提升,拟合度 $R^2$ 分别为 0.772 和 0.806,均方根误差 $RMSE$ 分别为 0.328 mg/L 和 0.302 mg/L。引入贝叶斯优化算法对随机森林模型进行优化得到 BO-RF 模型,其拟合度 $R^2=0.865$,均方根误差 $RMSE=0.253$ mg/L。优化后的模型 BO-RF 拟合度较好,模型精度显著提高。贝叶斯优化后的随机森林 BO-RF 算法更适合反演天德湖水体 DOC 质量浓度,为城市湖泊水质的遥感监测提供参考。

**关键词**: 可溶性有机碳; 遥感反演; Planet; 随机森林; 贝叶斯优化

**中图分类号**: P237 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2022.06.007

## 0 引言

可溶性有机碳(dissolved organic carbon, DOC)是水质评价的重要指标,是指可以通过 0.45 μm 滤膜的所有有机碳[1]。DOC 质量浓度的增加,会导致水体有机酸含量提高、水体透明度降低、水下光场发生变化,直接或间接地影响水体中各种生物的生存和发展,进而影响到整个水体的生态系统[2]。获取城市湖泊的有机碳含量,对城市水质监测具有重要意义。

目前,城市湖泊的水质监测主要通过实地采样结合实验室分析来获取水质信息[3]。这种方法虽然准确度较高,但需耗费大量的人力和成本,且只能获得各采样点数据。遥感反演技术具备区域化监测能力,可以高效地获取水质情况,而且成本低,在水质监测领域表现突出[4]。卫星遥感用于 DOC 质量浓度的反演已取得一定成果,主要方法有两类。一类方法是先反演有色可溶性有机物(colored dissolved organic matter, CDOM)浓度,再根据 DOC 与 CDOM 的关系计算 DOC 质量浓度[5]。第二类方法是直接利用遥感波段反射率与水体 DOC 质量浓度之间的关系进行反演[6]。DOC 质量浓度的遥感反演研究虽然取得一定进展,但大多采用统计回归的方法构建反演模型。实验发现,统计回归模型难以很好地描述水质参数与遥感数据之间复杂的非线性关系[7],稳定准确的反演模型仍是研究的重点。随着信息时代的到来,机器学习开始应用于各种类型的计算[8]。凭借其自身优势,机器学习在水质遥感反演领域发展迅速,许多研究使用神经网络[9]和支持向量机[10]等方法构建反演模型。神经网络虽然具有较好的鲁棒性和非线性逼近能力,但存在参数较多、学习时间长等缺点[11]。支持向量机同样具备非线性拟合能力,但面临参数选取困难和易陷入局部极值的问题[12]。作为机器学习的主流算法之一,随机森林(random forest, RF)算法近年来逐渐被学者发掘并应用于遥感反演。随机森林是决策树的集合,依靠多个决策树预测组合成一个模型,不易过拟合,抗干扰性强。贝叶斯优化(Bayesian optimization, BO)算法是一种全局优化

**收稿日期**:2022-01-12;**修订日期**:2022-04-01

**基金项目**:国家自然科学基金联合项目(U1704125)

**作者简介**:李爱民(1972—),男,山东菏泽人,郑州大学副教授,博士,主要从事遥感与地理信息技术研究,E-mail:aiminli@zzu.edu.cn。

算法,基于贝叶斯优化框架只需经过少数次的目标函数评估即可获得理想解,对于求解目标函数表达式未知、非凸的复杂优化问题,贝叶斯优化是一种有效方法[13]。

由于传统回归模型不能很好地描述水质参数与遥感数据之间的非线性关系,难以获得满足精度要求的模型,限制了水质反演结果的准确性。常用的神经网络和支持向量机等模型具有较好的非线性逼近能力,但是存在参数选取困难、模型稳健性不足等问题。为了构建精度高、稳健性好的反演模型,本研究利用贝叶斯优化算法对随机森林模型进行优化,提出了一种贝叶斯优化随机森林模型(BO-RF)的城市湖泊 DOC 质量浓度反演方法。应用于 Planet 卫星影像反演天德湖的 DOC 质量浓度并分析 DOC 空间分布情况,探讨适用于城市湖泊的 DOC 遥感反演模型,为城市水体的 DOC 质量浓度遥感监测提供技术支持。

## 1　研究区域与数据

### 1.1　研究区域

天德湖(34°48′~34°49′N,113°29′~113°30′E)为须水河干流上的一个湖泊,水域面积约为 0.306 $km^2$,如图 1 所示。随着城市发展,工业集聚,人类活动加剧,城市湖泊的水质状况备受关注。

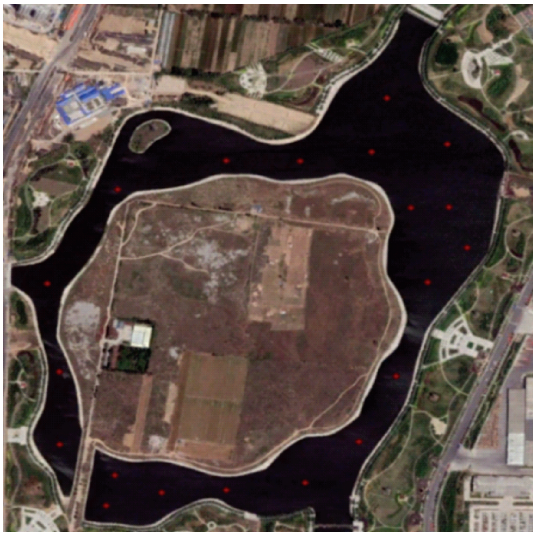

**图 1　研究区采样点位置**

**Figure 1　Location of sampling points in the study area**

### 1.2　实验数据

1.2.1　水质数据

2019 年 4 月 16 日和 2019 年 5 月 22 日两次进入湖区采样得到水质数据。采样点按照均匀分散、特征区域增设的原则布置,采样点位置如图 1 所示。利用专用采水器采集水面下 30~50 cm 深处的水样,共 40 个样本,对水样编号并记录采样点的 GPS 位置坐标。样品采集后马上送至实验室测定 DOC 质量浓度,得到采样点水质参数 DOC 实测数据。

1.2.2　Planet 卫星影像数据

对于城市湖泊 DOC 的监测常常需要多时段,所以具有高时间分辨率、空间分辨率为 3 m 的 Planet 卫星影像是个很好的选择。Planet 拥有 170 余颗 Dove 小卫星,是世界上唯一全球高分辨率、高频次的遥感卫星,影像信噪比高[14]。本文选用 Planet 卫星 2019 年 4 月 16 日和 2019 年 5 月 22 日两期影像数据作为遥感数据源,传感器为 Bayer 滤镜 CCD 相机,成像范围覆盖研究水域,提取出两期影像对应采样点的反射率数据。Planet 卫星基本参数如表 1 所示。

**表 1　Planet 卫星基本参数**

**Table 1　Basic parameters of Planet**

| 参数 | 说明 |
|---|---|
| 传感器高度 | 太阳同步轨道 475 km;国际空间站轨道 400 km |
| 像元大小/m | 3 |
| 区域成像大小 | 24 km×7 km |
| 产品模式 | L1B |
| 波段范围 | B1 蓝波 455~515 nm;B2 绿波 500~590 nm;B3 红波 590~670 nm;B4 近红外 780~860 nm |

L1B 级别产品数据已经过几何校正和辐射校正,本研究主要利用 ENVI 软件对 Planet 影像进行大气校正、图像镶嵌和剪裁等预处理。为提高水体识别的准确性,使用基于绿波段与近红外波段的归一化比值指数 *NDWI* 对水体进行提取,计算式为

$$NDWI = (Green - NIR)/(Green + NIR)。\quad (1)$$

式中:*Green* 为绿波段反射率;*NIR* 为近红外波段反射率。

## 2　研究方法

### 2.1　波段选取

在建模前先对实测 DOC 数据与影像提取的反射率进行 PEARSON 相关性分析,选择敏感波段,相关系数 $R$ 为

$$R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}。\quad (2)$$

式中:$x$ 为各波段及组合反射率;$y$ 为实测 DOC 质

量浓度值;$x_i$、$y_i$ 为变量组内第 $i$ 个数值;$\bar{x}$、$\bar{y}$ 为两组变量的平均值。

通过计算发现单波段与 DOC 质量浓度值的相关性较低,不适合直接建模。对各种波段组合进行比较,结果表明部分波段进行组合可以得到高于单波段的相关系数,统计各组合相关系数如图 2 所示。

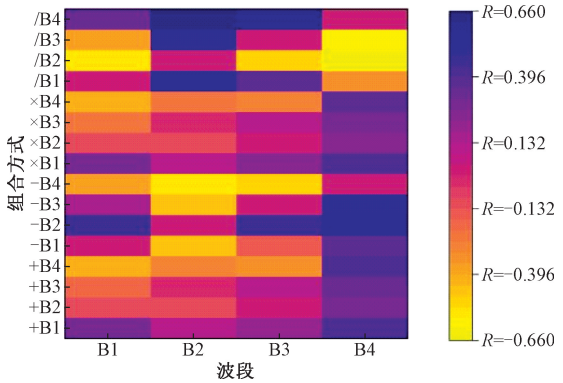


**图 2 波段组合与 DOC 质量浓度相关系数**

**Figure 2 Correlation coefficient between band combination and DOC concentration**

根据图 2 中 PEARSON 相关性分析,选取相关系数较高的波段组合作为输入因子。本文选取波段组合 B2/B4 作为 DOC 质量浓度反演的遥感数据。

## 2.2 常用模型构建

### 2.2.1 传统回归模型

目前,有不少利用传统回归方法进行 DOC 等水质参数遥感反演的研究,其中波段比值模型较为常见[15]。根据 PEARSON 相关性分析结果,本研究选用波段组合 B2/B4 构建回归模型,通过计算发现利用波段组合 B2/B4 构建的三次方程回归模型:$Y = 22.884X^3 - 47.444X^2 + 29.043X + 1.419\ 5$ 效果最好,但拟合度不高。

### 2.2.2 BP 神经网络模型

本文以选定的波段组合作为输入数据,DOC 实测浓度作为输出数据,使用 python 构建 BP 神经网络模型。其中,输入数据的 80%用于训练网络,20%用来测试。利用损失函数二阶导数矩阵即海森矩阵来迭代优化损失函数。对包含不同隐藏层节点数的神经网络模型的测试结果进行比较,从而选出最佳节点数。通过多次实验,发现隐藏层节点数为 8 时效果最好,定为最终模型。BP 神经网络中激活函数为 Relu 函数,学习函数为梯度下降权重函数,其他参数设置如表 2 所示。

### 2.2.3 支持向量机回归模型

以选定的遥感波段组合作为输入数据,实测 DOC 质量浓度为输出变量,在 python 的 scikit learn 开源机器学习库中构建支持向量机。以 80% 的数据为训练集,20% 的数据作为测试集,采用径向基函数(RBF)作为核函数构建支持向量机模型。调用 GridSearchCV 寻找 SVR 的最佳参数 $C$(惩罚系数)和 $gamma$。全局搜索计算得出 $C = 15$、$gamma = 1$ 时拟合效果最佳。

**表 2 BP 神经网络参数设置**

**Table 2 BP neural network parameter setting**

| 参数 | 取值 |
|---|---|
| 隐藏层节点数 | 8 |
| 惩罚系数 | 0.000 1 |
| 最大迭代次数 | 80 000 |

## 2.3 随机森林及优化算法模型构建

### 2.3.1 随机森林 RF 模型

随机森林算法通过随机取样生成多个决策树,综合各决策树得出最终结果,能很好地解决单一决策树过拟合的问题[16]。与人工神经网络相比,随机森林简单高效,在参数优化和变量分析上优势突出[17]。

基于 bagging 框架建立随机森林的步骤如下:

步骤 1 随机有放回地抽样,选出 $N$ 个训练集作为每棵回归树的根节点样本;

步骤 2 用子训练集训练一个 CART 回归树(决策树);

步骤 3 $N$ 棵决策树得到 $N$ 个结果;

步骤 4 对 $N$ 个结果取平均值作为随机森林最终结果。

模型测试结果为

$$E\left(\frac{\sum X_i}{N}\right) = E(X_i)。 \tag{3}$$

式中:$X_i$ 为随机可放回抽样的子数据集的变量,$i = 1, 2, \cdots, N$。

使用选定的反射率波段组合作为输入数据,实测的 DOC 质量浓度作为输出数据,其中,数据的 80%作为训练数据,20%作为验证数据。在 python 的 scikit learn 开源机器学习库利用随机森林算法构建模型。在随机森林回归器中,RF 框架特征有 $n\_estimators$、$oob\_score$、$criterion$。RF 决策树参数有 $max\_features$、$max\_depth$、$min\_samples\_split$、$min\_samples\_leaf$、$random\_state$。各参数设置如表 3 所示。

### 2.3.2 贝叶斯优化的随机森林 BO-RF 模型

在随机森林模型中,$n\_estimators$、$min\_samples\_split$、$max\_features$、$max\_depth$ 等超参数全部使用

**表 3　随机森林算法参数设置**

**Table 3　Random forest algorithm parameters setting**

| 参数 | 取值 | 说明 |
|---|---|---|
| *n_estimators* | 100 | 弱学习器的最大迭代次数(树的数量),默认值 100 |
| *oob_score* | False | 是否采用袋外样本来评估模型的好坏,默认为 False |
| *criterion* | mse | 决策树分裂的标准,默认为 mse |
| *max_ features* | auto | 节点分裂时参与判断的最大特征数 |
| *max_depth* | None | 设置树的最大深度,默认为 None |
| *min_samples_split* | 2 | 划分节点时划分最少的样本数,样本数小于这个值则不会继续进行划分,默认为 2 |
| *min_samples_leaf* | 1 | 叶子节点最少的样本数,默认为 1 |
| *random_state* | 2 | 随机数种子,本实验固定为 2 |

默认值。为了提高模型精度,引入贝叶斯优化算法对随机森林进行优化,优化过程采用贝叶斯定理:

$$P(f \mid D_t) = \frac{P(D_t \mid f)P(f)}{P(D_t)}。\tag{4}$$

式中:$f$ 表示参数模型中的参数;$D_t = \{(\boldsymbol{a}_1, \boldsymbol{b}_1), (\boldsymbol{a}_2, \boldsymbol{b}_2), \cdots, (\boldsymbol{a}_t, \boldsymbol{b}_t)\}$ 表示已观测集合,$\boldsymbol{a}_t$ 表示决策向量,$\boldsymbol{b}_t = f(\boldsymbol{a}_t) + t$ 表示观测值,$t$ 表示观测误差;$P(D_t \mid f)$ 表示 $y$ 的似然分布;$P(f)$ 表示 $f$ 的先验概率分布;$P(D_t)$ 表示 $f$ 的边际似然分布;$P(f \mid D_t)$ 表示 $f$ 的后验概率分布,后验概率分布描述通过已观测数据集对先验进行修正后未知目标函数的置信度。

贝叶斯优化的两个核心过程是先验函数(prior function,PF)和采集函数(acquisition function,AC)。本文基于高斯过程,初始化替代函数的先验分布,根据替代函数的先验分布,采样若干个数据点,再使用采样的值得到目标函数的新值。然后根据新的数据,更新替代函数的先验分布,并开始重复迭代。迭代之后,根据当前的高斯过程找到全局最优解。

贝叶斯优化的主要步骤如图 3 所示。在 python 中导入贝叶斯优化算法,利用贝叶斯优化调节 *n_estimators*、*min_samples_split*、*max_features*、*max_depth* 等对随机森林模型性能和速度影响较大的超参数。具体过程如下:定义目标函数,函数输入为调优的几个参数,输出为模型交叉验证 5 次的 $R^2$ 均值;设置超参数搜索空间 pbounds 如表 4 所示;构建贝叶斯优化器,设置 *n_iter* = 25,*init_points* = 5。通过实验得出最优参数:*max_features* = 0.817、*min_samples_split* = 2、*max_depth* = 8、*n_estimators* = 669,并使用最优参数构建模型。

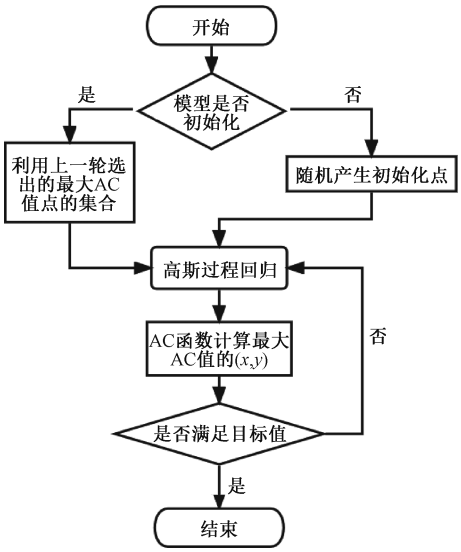


**图 3　贝叶斯优化流程**

**Figure 3　Bayesian optimization process**

**表 4　超参数搜索空间**

**Table 4　Hyperparameter search space**

| 超参数 | 搜索空间取值 |
|---|---|
| *min_samples_split* | 2~20 |
| *n_estimators* | 10~1 000 |
| *max_features* | 0.100~0.999 |
| *max_depth* | 2~15 |

## 3　分析与讨论

### 3.1　模型精度分析

本文构建的模型均使用回归模型常用的评估指标决定系数 $R^2$ 和均方根误差 *RMSE* 来评价模型精度。其中决定系数 $R^2$ 越大表示模型拟合效果越好;*RMSE* 是预测值与真实值的误差平方根的均值,值越小模型精度越高。计算式如下:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2};\tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}。\tag{6}$$

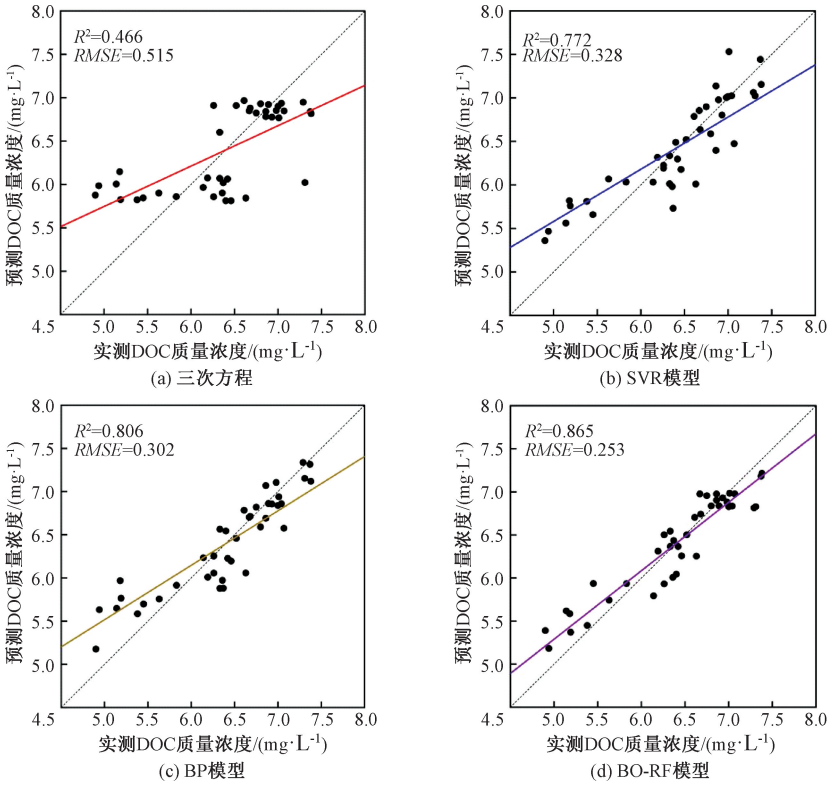


**图 4 模型精度对比**

**Figure 4 Model accuracy comparison**

式中:$n$ 表示样本数量;$Y_i$ 表示第 $i$ 个样本的真实值;$\hat{Y}_i$ 表示第 $i$ 个样本的预测值。

对各模型结果进行对比分析,结果如图 4 所示。由图 4 可以看出,三次方程回归模型的点相对分散,拟合线偏移角度较大,预测效果较差。支持向量机模型(SVR)和 BP 神经网络模型有所提高,但在低浓度区域均有部分点偏离 1:1线较多,说明 SVR 和 BP 虽然能处理复杂非线性问题,但是存在过拟合的情况,模型稳健性不足。BO-RF 模型样点基本在 1:1线附近,虽略有偏差,但从 $R^2$ 和 *RMSE* 来看,BO-RF 模型优势还是比较明显的。由此也验证了贝叶斯优化算法的优越性,说明该方法可以用于 DOC 质量浓度反演。

### 3.2 DOC 质量浓度空间分布分析

将 BO-RF 模型应用于 Planet 遥感影像反演 DOC 质量浓度,得到天德湖 DOC 质量浓度的空间分布图,总体反演结果与实测情况对比分析,匹配度良好,结果如图 5 所示。

由图 5 可以看出,天德湖水域的 DOC 质量浓度集中在 4.0~8.0 mg/L 之间,总体分布大致呈现西高东低,湖泊中部低于沿岸,且随离岸距离增加而降低的空间特征。其中东北方向入水口 DOC 质量浓度较低,这一区域水体流动性大,DOC 质量浓度受来水的影响较大,且水面没有漂浮物和浮萍等水生植物,产生的内源有机碳较少。北部束窄口 DOC 质量浓度处于高位,结合调查发现,该处水面有较多浮游植物,且湖心岛此处地势低,排水相对较多,受陆源输入影响。水体生物产生的内源有机碳、土壤侵蚀等陆源碳输入是造成湖泊有机碳质量浓度差异的关键因素,对碳来源进行分析有利于进一步发掘 DOC 的分布特征。

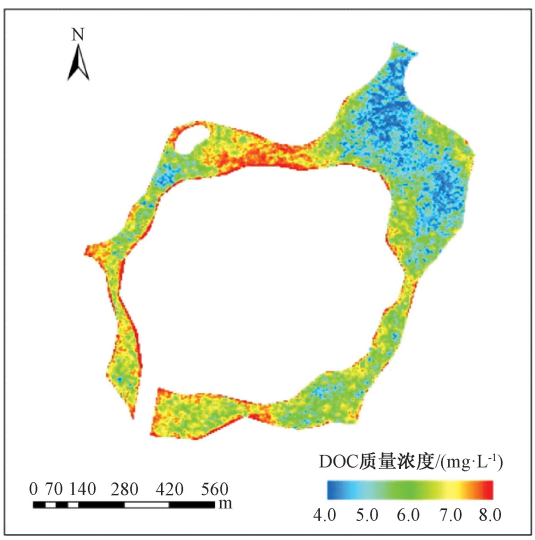


**图 5 天德湖水域 DOC 质量浓度空间分布**

**Figure 5 Spatial distribution of DOC concentration in Tiande lake waters**

## 4　结论

为解决统计回归反演模型拟合度低的问题,本研究基于遥感反射率和实测水质数据,构建了 BP 神经网络、支持向量机和贝叶斯优化随机森林 BO-RF 等多个机器学习模型,通过实验得出 BO-RF 模型效果最好,将实验结果应用于 Planet 卫星影像反演郑州天德湖 DOC 质量浓度。主要结论如下:

(1)通过对 Planet 影像各波段及波段组合进行相关性分析,得出绿波段 B2 与近红外波段 B4 的波段组合 B2/B4 的遥感反射率与天德湖水域的 DOC 水质数据相关性最高。

(2)引入贝叶斯优化对随机森林 RF 模型进行优化,解决了局部最优的问题,模型的反演精度高于其他模型。将 BO-RF 模型应用于 Planet 卫星数据反演 DOC 质量浓度,效果良好,表明利用贝叶斯优化随机森林的优越性,同时也说明了基于实测数据和卫星影像数据的水质参数反演方法具有良好的应用前景和推广价值。

(3)从反演的 DOC 质量浓度分布情况来看,整体效果良好,其中东北部入水口浓度较低,湖中心向周围浓度逐渐升高,符合湖泊水质的分布特点,也体现了反演结果与实际情况的符合程度。

此外,虽然 BO-RF 模型在天德湖区域性能良好,但影响城市水体 DOC 质量浓度的因素复杂多样,今后将结合水体固有光学量、表观光学量和其他水质组分的影响,进一步探讨水体 DOC 的光学遥感机理,提高反演精度。

## 参考文献:


[1] EVANS C D, MONTEITH D T, COOPER D M. Long-term increases in surface water dissolved organic carbon: observations, possible causes and environmental impacts[J]. Environmental pollution, 2005, 137(1): 55-71.

[2] PILLA R M, COUTURE R M. Attenuation of photosynthetically active radiation and ultraviolet radiation in response to changing dissolved organic carbon in browning lakes: modeling and parametrization[J]. Limnology and oceanography, 2021, 66(6): 2278-2289.

[3] 侯迪波, 张坚, 陈泠, 等. 基于紫外-可见光光谱的水质分析方法研究进展与应用[J]. 光谱学与光谱分析, 2013, 33(7): 1839-1844.
HOU D B, ZHANG J, CHEN L, et al. Water quality analysis by UV-vis spectroscopy: a review of methodology and application[J]. Spectroscopy and spectral analysis, 2013, 33(7): 1839-1844.

[4] 彭保发, 陈哲夫, 李建辉, 等. 基于 GF-1 影像的洞庭湖区水体水质遥感监测[J]. 地理研究, 2018, 37(9): 1683-1691.
PENG B F, CHEN Z F, LI J H, et al. Monitoring water quality of Dongting Lake region based on GF-1 image[J]. Geographical research, 2018, 37(9): 1683-1691.

[5] CHERUKURU N, FORD P W, MATEAR R J, et al. Estimating dissolved organic carbon concentration in turbid coastal waters using optical remote sensing observations[J]. International journal of applied earth observation and geoinformation, 2016, 52: 149-154.

[6] KUTSER T, VERPOORTER C, PAAVEL B, et al. Estimating Lake carbon fractions from remote sensing data[J]. Remote sensing of environment, 2015, 157: 138-146.

[7] 徐良将, 黄昌春, 李云梅, 等. 基于高光谱遥感反射率的总氮总磷的反演[J]. 遥感技术与应用, 2013, 28(4): 681-688.
XU L J, HUANG C C, LI Y M, et al. Deriving concentration of TN, TP based on hyper spectral reflectivity[J]. Remote sensing technology and application, 2013, 28(4): 681-688.

[8] 蔡婉贞, 黄翰. 基于 BP-RBF 神经网络的组合模型预测港口物流需求研究[J]. 郑州大学学报(工学版), 2019, 40(5): 85-91.
CAI W Z, HUANG H. A model based on the combination of BP and RBF neural network for port logistic demand forecasting[J]. Journal of Zhengzhou university (engineering science), 2019, 40(5): 85-91.

[9] 马丰魁, 姜群鸥, 徐藜丹, 等. 基于 BP 神经网络算法的密云水库水质参数反演研究[J]. 生态环境学报, 2020, 29(3): 569-579.
MA F K, JIANG Q O, XU L D, et al. Retrieval of water quality parameters based on BP neural network algorithm in Miyun reservoir[J]. Ecology and environmental sciences, 2020, 29(3): 569-579.

[10] 夏晓芸, 解启蒙, 杨国范, 等. 大伙房水库叶绿素 a 浓度反演模型研究[J]. 节水灌溉, 2018(8): 39-42, 46.
XIA X Y, XIE Q M, YANG G F, et al. A study on remote sensing inversion model of chlorophyll-A in Dahuofang reservoir based on HJ-1A/1B data[J]. Water saving irrigation, 2018(8): 39-42, 46.

[11] 张明慧, 苏华, 季博文. MODIS 时序影像的福建近岸叶绿素 a 浓度反演[J]. 环境科学学报, 2018, 38(12): 4831-4839.
ZHANG M H, SU H, JI B W. Retrieving nearshore

chlorophyll-a concentration using MODIS time-series images in the Fujian Province (China)[J]. Acta scientiae circumstantiae, 2018, 38(12): 4831-4839.

[12] 盛辉, 池海旭, 许明明, 等. 改进SVR的内陆水体COD高光谱遥感反演[J]. 光谱学与光谱分析, 2021, 41(11): 3565-3571.
SHENG H, CHI H X, XU M M, et al. Inland water chemical oxygen demand estimation based on improved SVR for hyperspectral data[J]. Spectroscopy and spectral analysis, 2021, 41(11): 3565-3571.

[13] SHAHRIARI B, SWERSKY K, WANG Z Y, et al. Taking the human out of the loop: a review of Bayesian optimization[J]. Proceedings of the IEEE, 2016, 104(1): 148-175.

[14] 韩中含, 徐白山, 杨成林, 等. 基于Planet多光谱影像的南海岛礁水深反演研究[J]. 测绘与空间地理信息, 2020, 43(12): 139-142, 146.
HAN Z H, XU B S, YANG C L, et al. Research on reef depth retrieval of South China Sea Island based on planet multispectral image[J]. Geomatics & spatial information technology, 2020, 43(12): 139-142, 146.

[15] 吴志明, 李建超, 王睿, 等. 基于随机森林的内陆湖泊水体有色可溶性有机物(CDOM)浓度遥感估算[J]. 湖泊科学, 2018, 30(4): 979-991.
WU Z M, LI J C, WANG R, et al. Estimation of CDOM concentration in inland lake based on random forest using Sentinel-3A OLCI[J]. Journal of lake sciences, 2018, 30(4): 979-991.

[16] 王奕森, 夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术, 2018, 12(1): 49-55.
WANG Y S, XIA S T. A survey of random forests algorithms[J]. Information and communications technologies, 2018, 12(1): 49-55.

[17] 方馨蕊, 温兆飞, 陈吉龙, 等. 随机森林回归模型的悬浮泥沙浓度遥感估算[J]. 遥感学报, 2019, 23(4): 756-772.
FANG X R, WEN Z F, CHEN J L, et al. Remote sensing estimation of suspended sediment concentration based on random forest regression model[J]. Journal of remote sensing, 2019, 23(4): 756-772.

## Remote Sensing Retrieval of Urban Lake DOC Concentration Based on Optimized Random Forest Algorithm

LI Aimin [1], WANG Hailong[2], XU Youcheng[2]

(1. School of Geo-science and Technology, Zhengzhou University, Zhengzhou 450001, China; 2. School of Water Conservancy Engineering, Zhengzhou University, Zhengzhou 450001, China)

**Abstract**: In remote sensing monitoring of dissolved organic carbon (DOC) content in urban lakes, the traditional regression model is difficult to describe the nonlinear relationship and can not meet the accuracy requirements. In this study, Bayesian optimization algorithm was introduced into the parameter optimization of random forest model, and a DOC concentration inversion method of urban lakes based on Bayesian optimization random forest model (BO-RF) was proposed. Taking the water area of Tiande Lake in Zhengzhou city as an example, the remote sensing inversion method of DOC concentration in urban lakes was studied based on high spatial-temporal resolution Planet satellite image data and measured DOC water quality data. Through PEARSON correlation analysis, the results showed that the best band combination of Planet satellite image band for retrieving DOC concentration was B2/B4. The determination coefficient $R^2$ of the band ratio model obtained by the traditional regression method was 0.466, and the root mean square error *RMSE* was 0.515 mg/L, which could not meet the accuracy requirements. The modeling accuracy was improved by using support vector machine and BP neural network, the fitting $R^2$ was 0.772 and 0.806, respectively, and the root mean square error *RMSE* was 0.328 mg/L and 0.302 mg/L, respectively. Bayesian optimization algorithm was introduced to optimize the random forest model to obtain the BO-RF model, and its fitting degree $R^2$ was 0.865 and root mean square error *RMSE* was 0.253 mg/L. The BO-RF fit of the optimized model was good, and the accuracy of the model was significantly improved. The Bayesian optimized random forest BO-RF algorithm was more suitable for retrieving DOC concentration in Tiande lake, which could provide a reference for remote sensing monitoring of urban lake water quality.

**Keywords**: dissolved organic carbon; remote sensing inversion; Planet; random forest; Bayesian optimization