文章编号:1001-9014(2018)05-0621-10 DOI:10.11972/j.issn.1001-9014.2018.05.016

基于哨兵3A-OLCI影像的内陆湖泊藻蓝蛋白浓度反演算法研究

苗 松[1], 王 睿[1], 李建超[1], 吴志明[1], 时 蕾[1], 吕 恒[1,2,3*], 李云梅[1,2,3], 赵少华[4], 刘思含[4]

(1. 南京师范大学 虚拟地理环境教育部重点实验室,江苏 南京 210023;
2. 江苏省地理环境演化国家重点实验室培育建设点,江苏 南京 210023;
3. 江苏省地理信息资源开发与利用协同创新中心,江苏 南京 210023;
4. 环境保护部卫星环境应用中心,北京 100029)

摘要:蓝藻是内陆富营养水体水华发生的主要优势藻种,而藻蓝蛋白(Phycocyanin,PC)是蓝藻的标志性色素,因此利用遥感估算水体中藻蓝蛋白浓度从而对蓝藻水华预警具有重要意义。利用太湖、滇池、洪泽湖的实测数据,构建藻蓝蛋白随机森林遥感估算模型,并将模型应用到哨兵3A-OLCI影像。通过对随机森林的输入自变量进行重要性分析,发现第7波段(620 nm)、第8波段(665 nm)和第9波段(675 nm)三个波段对藻蓝蛋白反演的影响程度最大。同时,反演结果表明,随机森林反演的藻蓝蛋白浓度平均绝对百分比误差(MAPE)为34.86%,均方根误差(RMSE)为38.67 μg/L,与Simis等半分析算法和齐琳的PCI(Phycocyanin Index)指数模型相比,平均绝对百分比误差(MAPE)分别提高了85.06%和15.65%,均方根误差分别提高了26.08 μg/L和19.86 μg/L。利用地面实测数据对同步卫星影像大气校正进行精度评价,发现MUMM(The Management Uint Mathematical Model)算法可以用于OLCI影像的大气校正,尤其在560~779 nm处共8个波段的MAPE低于30%,光谱曲线与实测光谱曲线形状保持一致。结果表明所构建的基于哨兵3A-OLCI影像的藻蓝蛋白随机森林反演模型,可以成功的应用于我国的内陆富营养化湖泊,为我国内陆湖泊藻蓝蛋白浓度的遥感反演提供一个新的算法。

关 键 词:藻蓝蛋白;OLCI;随机森林;遥感;反演
中图分类号:X524;X87 **文献标识码**:A

Retrieval algorithm of phycocyanin concentration in inland lakes from Sentinel 3A-OLCI images

MIAO Song[1], WANG Rui[1], LI Jian-Chao[1], WU Zhi-Ming[1], SHI Lei[1], LYU Heng[1,2,3*], LI Yun-Mei[1,2,3], ZHAO Shao-Hua[4], LIU Si-Han[4]

(1. Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China;
2. State Key Laboratory Cultivation Base of Geographical Envir-onment Evolution (Jiangsu Province), Nanjing 210023, China;
3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China;
4. Satellite Environment Application Center, Ministry of Environment Protection, Beijing 100029, China)

Abstract: Cyanobacteria is the dominant algae species in inland eutrophic water bodies, and the phycocyanin (PC) is its unique pigment which can be used as an indicator of its presence. Therefore, the retrieval of PC concentration by remote sensing is of great significance to early warning of cyanobacteria bloom. In this paper, the Random Forest retrieval Model for estimating PC concentration based on the sentinel 3A-OLCI bands was developed using in situ data collected from Taihu Lake, Dianchi Lake and

收稿日期:2017-08-04,**修回日期**:2017-11-30 **Received date**: 2017-08-04, **revised date**: 2017-11-30
基金项目:国家重点研发计划项目"城市水体水质高分遥感与地面协同监测关键技术研究"(2017YFB0503902);国家自然科学基金(41471282);水体污染控制与治理科技重大专项,"城市水环境遥感监管及定量评估关键技术研究"(2017ZX07302003);民用航天技术预先研究项目"5m红外多谱段卫星系统环保领域研究"
Foundation items: Supported by Nation Key R&D Program of China (2017YFB0503902); the National Natural Science Foundation of China (41471282); Major Science and Technology Program for Water Pollution Control and Treatment (2017ZX07302003); Civil Aerospace Technology Research Project in Advance

作者简介(**Biography**):苗 松(1992-),男,河南焦作人,博士研究生,主要研究方向为水色遥感. E-mail:njnums1214@163.com
* **通讯作者**(**Corresponding author**): E-mail: Heng.Lyu@njnu.edu.cn

Hongzehu Lake. The results of the importance analysis of input variables in random forest demonstrated that the seventh band(674 nm), the eighth band(665 nm) and the ninth band (620 nm) have significant impact on the PC estimation. The accuracy assessment showed that the Mean Absolute Percentage Error(MAPE) of this PC retrieval model is only 34.86% with the Root Mean Square Error(RMSE) of 38.67 μg/L. The comparison between the mode developed by this paper and other models, i.e., Simis semi-analytic algorithm and PCI exponential model was extensively conducted, and it was found that compared with other two models, the MAPE was improved by 85.65% and 15.65% respectively, and the RMSE was improved by 26.08 μg/L and 19.86 μg/L respectively. The atmospheric correction accuracy was further analyzed using the in situ samples and synchronous satellite image, and the result showed that the Management Uint Mathematical Model (MUMM) method can be successfully used for the OLCI image. The atmospheric corrected spectral curves are consistent with the measured spectral curves, and the MAPEs of 8 bands are all less than 30% at the wavelength range between 560 and 779 nm. The random forest model developed for estimating PC concentration in this paper can be successfully applied to Sentinel 3A-OLCI images, which provides a new algorithm for remote estimation of phycocyanin concentration in inland lake.

Key words: phycocyanin(PC), OLCI, Random Forest(RF), remote sensing, inversion
PACS: 42.68.WT, 92.40.qj

引言

湖泊富营养化和蓝藻暴发已经日益成为全球环境关注的焦点[1-3]. 我国内陆湖泊如太湖、巢湖、滇池等均出现过不同程度上的水华现象. 已有研究表明[4-7]藻蓝蛋白(PC)是蓝藻的特征色素,因此可以将藻蓝蛋白色素的浓度作为表征水体中蓝藻含量的指标,为预警蓝藻水华提供了新的指征. 由于藻蓝蛋白在620 nm附近具有区别其他藻类的吸收峰,因此国内外众多学者利用这一光学特征对蓝藻进行识别、监测和定量估算. Dekker[8]曾提出基线算法,通过624 nm与附近两个波段600 nm与648 nm所连基线的相对高度来计算藻蓝蛋白的浓度. Schalles[9]建立了650 nm与620 nm波段附近的遥感反射率的比值和藻蓝蛋白色素浓度之间关系的单一反射比算法(Single reactance ratio algorithm). Simis[10-11]等提出了嵌套波段比值(709 nm/665 nm,709 nm/620 nm)的方法,首先利用波段比值的方法计算藻蓝蛋白在620 nm的吸收值,然后同通过比吸收系数计算藻蓝蛋白浓度,并证实算法可用于以蓝藻为主的浑浊水体的藻蓝蛋白反演,目前该方法使用最为广泛. 尹斌[12]利用Simis半分析模型对滇池的藻蓝蛋白浓度进行反演,反演精度良好,并指出季节性差异导致的同生长期蓝藻细胞内色素浓度和组分的变化是导致模型误差的主要原因. 齐琳[13]定义藻蓝素指数(Phycocyanin Index,PCI)为560 nm和665 nm之间的基线高度减去620 nm遥感反射率,然后利用PCI指数估算藻蓝蛋白浓度.

随机森林(RF)算法[14]是美国科学家Leo breiman将Bagging集成学习理论[15]和随机子空间方法[16]相结合于2001年发表的一种机器学习算法,其中随机森林回归算法(Random Forests for regression,RFR)是重要应用类型. 大量的理论以及实测数据验证均表明随机森林方法能够克服决策树过拟合现象,对噪声和异常值有较好的容忍性,并处理大数据集(如高光谱数据)时非常高效,对于结果和解决反演问题具有可解释性和优势性[17]. 白琳[18]利用随机森林模型构建反演近地表气温模型,并对随机森林模型的输入参数进行重要性分析,认为地表温度是气温反演模型的最大影响因素;张颖[19]基于监测数据及随机森林分类算法对巢湖水质进行评价,结果表明该算法具有稳健性较高、实用性强、泛化性能好等特点,可以有效进行水质评价;侍昊[20]利用遥感影像光谱指数与图像变换方法构建多个特征变量,并结合随机森林模型,对太湖流域水生植被的空间分布进行研究;江佳乐[21]等基于实测数据提出了基于随机森林反演海盐的算法模型并表明基于多因子参数的随机森林反演海表盐度是可行高效的. 故随机森林是一种预测精度高、泛化能力高和对数据集中的噪声有较强鲁棒性的模型.

哨兵3A卫星是欧洲航空局于2016年2月发射的多光谱中分辨率卫星,该卫星载有4个传感器:海洋与陆地彩色成像光谱仪(OLCI)、海洋和陆地表面温度辐射计(SLSTR)、合成孔径雷达高度计(SRAL)和微波辐射计(MWR),可满足环境监测、海况评估、火灾探测、植被监测、海洋污染、湖泊河流高度变化

以及与云、气溶胶和污染相关的大气研究等一系列用途需求. 所搭载的海洋与陆地彩色成像光谱仪(OLCI)传感器是一种中分辨率线阵推扫成像光谱仪,幅宽为1300 km,视场68.5°,海洋上空分辨率为1.2 km,沿海区和陆地上空的分辨率为0.3 km,其21个光谱波段的设置和300 m分辨率为内陆湖泊水质参数遥感监测提供了一个新的数据源.

本文以太湖、滇池、洪泽湖藻蓝蛋白为研究对象,利用哨兵3A-OLCI光谱响应函数模拟后的光谱数据,构建随机森林藻蓝蛋白浓度估算模型,并与同期数据建立的simis半分析模型和PCI指数模型进行对比与分析,旨在为提升我国内陆湖泊藻蓝蛋白浓度遥感反演精度提供新的方法和思路.

1　研究区与数据源

1.1　研究区

以滇池、太湖、洪泽湖为研究区,这三个湖泊均曾有不同程度的富营养化现象发生,其中太湖和滇池是我国重点治理监测和治理的富营养化湖泊. 由于其各自地理位置、形成原因、气候条件、经济发展因素、入湖河流等因素的影响,水体的光学特性也呈现出较大的差异性[22-24].

1.2　实验数据

2016年7月22~23日和2016年12月6~9日和2017年4月13~14日分别对太湖、洪泽湖、滇池按照点位图(如图1所示)进行水面光谱测量和水样的采集,共采集109个实测数据. 采集的样品放入存储箱内保存,当天带回实验室进行过滤处理,测定各种水质参数,包括藻蓝蛋白浓度、叶绿素a浓度、总悬浮物浓度等.

1.2.1　遥感反射率测量

水面光谱数据的测量利用ASD公司生产的ASD Field Spec Pro便携式光谱仪进行采集,采集方法是采用水面以上测量法,同时记录了GPS点位、风速、气压及天气状况. 水面以上测量法是按照唐军武[25]等介绍的方法换算成水面遥感反射率. 具体计算遥感反射率过程如下:

(1)离水辐亮度 L_w 的计算:

$$L_{sw} = L_w + rL_{sky} \quad , \quad (1)$$

式中,L_w 为离水辐亮度,L_{sw} 为水体总辐射亮度,L_{sky} 为天空漫散射光,不携带任何水体信息,需去除;r 为气—水界面对天空光的反射率,取决于太阳位置、观测几何、风速风向或海面粗糙度等因素. 在上述几何观测条件下,平静水面可取 $r=0.022$,在5 m/s左右风速的情况下,r 可取0.025,10 m/s左右风速的情况下,r 为0.026~0.028.

(2)水表面入射总辐照度 $E_d(0^+)$ 可由测量标准板(plaque)的反射计算得到:

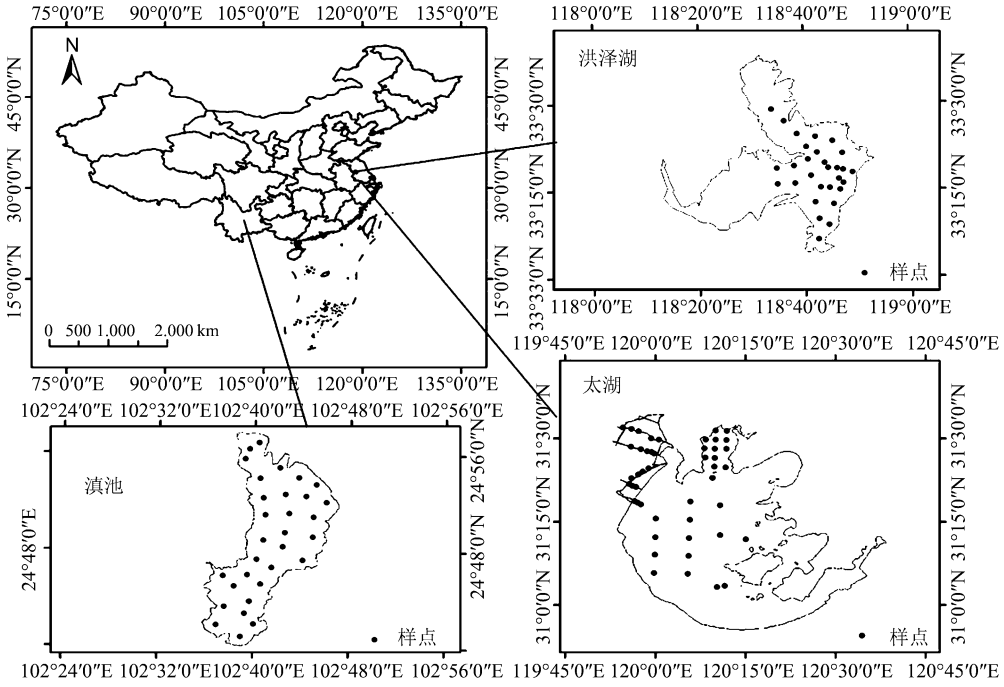

图1　太湖、洪泽湖、滇池采样分布示意图
Fig. 1　The sampling distribution of Taihu Lake, Hongzehu Lake and Dianchi Lake

$$E_d(0^+) = L_p\pi/\rho_P \quad , \quad (2)$$

式中,L_P 为测量标准版的反射率,ρ_P 为标准版的反射率.

(3)遥感反射率的导出:

$$R_{rs} = L_w/E_d(0^+) \quad . \quad (3)$$

1.2.2 室内水质参数的测量

(1)叶绿素 a 浓度(Chla)的测量

叶绿素 a 浓度的测量采用热乙醇—分光光度计法. 用直径 47 nm 的 GF/F 滤膜过滤一定的水样,将滤膜置于置于冰箱中冷冻 48 个小时,温度设为 -20℃. 冷冻结束后,用 90% 的热乙醇萃取,避光保存 4 ~ 6 小时,再利用 25 nm 的 GF/F 膜过滤萃取的溶液,在岛津 UV2450 紫外-分光光度计上测量 665 nm 和 750 nm 处的吸光度(E_{665} 和 E_{750})后,加入 1 滴 1% 稀盐酸化 1 min,再次测量 665 nm 和 750 nm 处的吸光度(A_{665} 和 A_{750}),进去利用公式(4)计算得到叶绿素 a 的浓度.

$$\text{Chla} = 27.9 \times [(E_{665} - E_{750}) - (A_{665} - A_{750})] \times V_B/(V \times \delta) \quad , \quad (4)$$

式中,V_B 为乙醇定容的体积,V 为水样体积,δ 为比色皿光程.

(2)藻蓝蛋白浓度(PC)的测量

藻蓝蛋白浓度的测量采用荧光法. 用直径 47 mm 的 GF/F 滤膜过滤一定体积的水样,滤膜放入离心管中并置于冰箱中冷冻 48 小时,温度设置为 -20℃. 冷冻结束后,加入 8 ml 的 PH = 7 的磷酸缓冲液并定容,然后将已加入缓冲液的离心管反复冻融 5 次以上,冻融后对样品进行离心,并取上清液通过 Fluo Imager M53 荧光仪测定样品的荧光强度,并使用相同条件下测量标准溶液得到的标准工作曲线求得样品的藻蓝蛋白浓度.

(3)悬浮物浓度的测量

悬浮物浓度采用国际标准测量法 GB 11901—89 烘干称重法进行测量. 用预先烘烤(550℃)并冷却后称重的直径 47 mmGF/F 滤膜过滤一定体积的水样,然后将滤膜进行二次烘干(110℃)并称重,将烘干后的滤膜的重量减去过滤前滤膜的重量即为总悬浮物重量,除以过滤水样的体积即可得到总悬浮物浓度.

1.3 遥感数据与预处理

OLCI 传感器是在 MERIS 传感器基础上改进和创新,增加的 6 个波段信息(400、673、764、767、910、1 020 nm)将更有利于影像的大气校正,共有 21 个波段,光谱范围从可见光到近红外(400 ~ 1 020 nm),OLCI 传感器携带有覆盖藻蓝蛋白吸收特征峰的 620 nm 波段和叶绿素对红光强吸收导致反射特征峰的 673 nm 波段,将更好地适用于藻蓝蛋白浓度和叶绿素 a 浓度的反演[26](https://sentinel.esa.int). 选取 4 幅影像,成像时间为洪泽湖 2016 年 12 月 7 日和 2016 年 12 月 8 日两景、滇池 2017 年 4 月 12 日一景和太湖 2017 年 5 月 18 日一景,数据下载来自欧洲太空局(https://scihub.copernicus.eu). 利用 SEADAS 7.4 对影像进行预处理,主要包括几何校正、大气校正,其中大气校正采用的算法为 MUMM 模型[27]. MUMM 算法是一种二类水体大气校正方法,是 Gordon 标准大气校正算法的扩展,将 Gordon 标准大气校正算法中假设红光以及近红外波段的离水辐射率为零代之以假设研究区域内 765 和 865 nm 的气溶胶散 = 射比和离水辐射率的比率为确定值;同时假设经过大气漫射透过率校正的 765 和 865 nm 处的水体反射率的比值为定值,在此基础上求解辐射传输方程,再根据标准大气校正算法进行校正.

2 研究方法

2.1 随机森林回归原理

随机森林回归算法是利用 Bagging 重抽样方法从原始样本中抽取多个样本,对每个新的样本进行决策树建模,然后组合多棵决策树的预测,通过投票得出最后的预测值. 同时在建立回归树时会有一部分样本数据不被选中,其作为检验样本的出现,起到了样本内部交叉验证的作用,可减少过度拟合情况的出现.

随机森林回归模型的输出结果是数值,假设从独立且随机分布的变量(X,Y)中抽取,输入量是 X,输出量是 Y,则预测值 $h(X)$ 的均方泛化误差为 $E_{X,Y}\left[\begin{matrix}Y-h\\(X)\end{matrix}\right]^2$. 其最终预测输出的结果是通过对 k 棵决策树$\{h(X,\theta_k)\}$取平均值得到的,其具有如下定理:

定理 1 当 $K\to\infty$ 时,

$$E_{X,Y}[Y - \alpha\nu k_h(X,\theta_k)]^2 \to E_{X,Y}[Y - E_\theta(X,\theta_k)]^2 \quad , \quad (5)$$

其中,式(5)中右边记为 $PE^*(forest)$,即随机森林的泛化误差. 每颗决策树的平均泛化误差定义为:

$$PE^*(tree) = E_\theta E_{X,Y}[Y - h(X,\theta)]^2 \quad . \quad (6)$$

定理 2 假设对所有的 θ,$Y = E_x h(X,\theta)$. 则:

$$PE^*(forest) \leqslant \bar{\rho} PE^*(tree) \quad , \quad (7)$$

式(7)$\bar{\rho}$ 是残差 $Y-h(X,\theta)$ 和 $Y-h(X,\theta')$ 的加权相关系数,且 θ 和 θ' 相互独立.

其中,定理 2 给出了随机森林精确回归的要求,即残差之间的低相关性和低的错误决策树. 随机森林通过因素$\bar{\rho}$ 来降低决策树的平均误差,同时其随机性需要关注低相关性[28].

2.2　随机森林反演模型构建

通过 Python 语言中的 random forest 数据包来实现随机森林反演模型的构建. 首先是确定自变量的选取. 结合哨兵 3A-OLCI 波段设置以及光谱响应函数,选取部分波段的遥感反射率作为输入自变量(具体波段的选取见后文). 其次需要确定的是反演过程中涉及到的两个参数:ntree 和 mtry. 其中 ntree 为决策树的数量,即使用 bootstrap 重采样的次数;mtry 为随机特征的数量,即每个树节点随机采样的数目,其大小通常为输入自变量的 1/3[19,29]. 因变量为藻蓝蛋白(PC)浓度. 随机森林反演模型具体步骤如下:

(1)利用 bootstrap 方法重采样,随机产生 K 个训练集;并利用每个训练集生成对应的决策树.

(2)假设特征有 M 维,从 M 维特征中随机抽取 m 个特征作为当前节点的分裂特征集,并以这 m 个特征中最好的分裂方式对该节点进行分裂. 同时每个决策树得到最大限度的增长且不进行剪枝.

(3)对于新的数据,单棵决策树的预测则通过节点的观测值取其平均值获得. 最后便可获得随机森林回归的预测值.

2.3　模型验证方法

随机森林在样本选取上具有较强的随机性,因此其本身具有着交叉验证的优点. 当决策树即 ntree 的数量足够多,可以保证每个样本分别作为训练样本和测试样本,有效地避免过度拟合的效果. 同时,为了进一步验证随机森林反演模型,文中随机抽取 76 个样本作为训练数据集,余下的 33 个样点作为随机特征的数量即验证数据. 首先利用训练样本数据建立随机森林模型,然后利用验证样本对随机森林模型进行评价,根据均方根误差(RMSE)、平均绝对百分比误差(MAPE)、相对均方根误差(rRMSE)和决定系数(R^2)评价模型优劣.

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i}-X_{modei,i})^2}{n}},\quad(8)$$

$$\mathrm{MAPE}=\sum_{i=1}^{n}\left|\frac{X_{modei,i}-X_{obs,i}}{X_{obs,i}}\right|\times\frac{100\%}{n},\quad(9)$$

$$\mathrm{rRMSE}=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{obs,i}-X_{modei,i})^2}}{\frac{1}{n}\sum_{i=1}^{n}X_{obs,i}}.\quad(10)$$

3　结果与讨论

3.1　随机森林变量重要性分析

为了分析自变量中不同波段遥感反射率对藻蓝蛋白反演的重要性,利用 Python 语言中的 feature_important 函数分析了各个波段遥感反射率,其 feature_imoportant 函数值越大,则表明该波段对藻蓝蛋白反演模型影响程度越大. 由图 2 可得知各遥感波段在随机森林回归模型中的重要性. 在输入的各遥感反射率波段中 620 nm、665 nm 、674 nm 三个波段的影响程度最大相比于其他波段,这可能是跟藻蓝蛋白和叶绿素 a 的光学特征有关,从图 3 实测样点的遥感反射率光谱曲线图可知(红色竖线分别代表 620 nm、665 nm 、674 nm 波段),叶绿素 a 在 674 nm 波长处对红光的强吸收作用会形成一个反射谷,620 nm 波长处的反射谷则是由于水样中蓝藻的藻蓝蛋白在 620 nm 波长附近的吸收引起的. 由于蓝藻的存在,藻蓝蛋白和叶绿素 a 在 620 nm 和 674 nm 波长处对光的较强吸收导致 650 nm 波长附近有反射峰的存在. 在内陆湖泊中,620 nm 波长处的反射谷和 650 nm 波长处的反射峰通常作为判断水体中是否含有蓝藻的主要光学特征. 而其他遥感反射率波段的重要性相对于 620 nm、665 nm 、674 nm 三个波段明显较低.

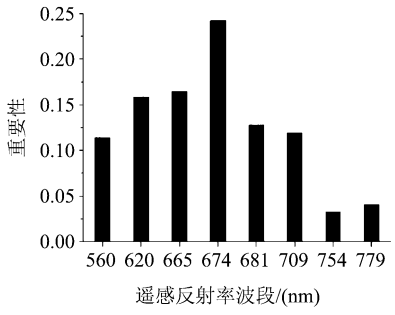

图 2　随机森林变量重要性
Fig. 2　Importance of random forest variables

3.2　随机森林反演模型验证与分析

前文中表明建立随机森林反演模型重要的两个参数是决策树和随机特征数量即 ntree 和 mtry,通过确定这两个参数即可构建最优随机森林反演模型. 在模型中,将 mtry 设定为样本量的 1/3(即 33 个样

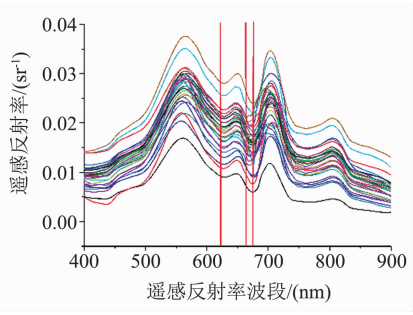

图 3 实测样点遥感反射率光谱
Fig. 3 Remote sensing reflectance spectrum of sample points

点). 通过统计学评判指标 MAPE 确定决策树 ntree 的值. 由图 4 可以看出,当 ntree 小于 100 时模型误差出现较大的波动. 当 ntree 大于 100 后模型误差趋于平稳. 因此将 ntree 的数目设置为 100.

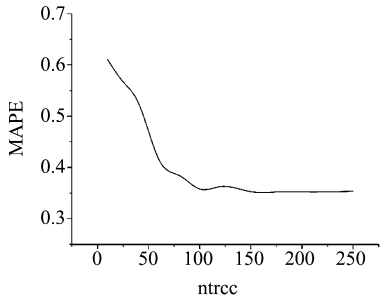

图 4 平均绝对百分比误差(MAPE)随决策树数目的变化
Fig. 4 The average absolute percentage error (MAPE) varies with the number of decision trees

因此可利用 76 个实测的多波段遥感反射率与藻蓝蛋白浓度构建随机森林反演模型、Simis 半分析模型、PCI 指数模型,同时通过 33 个实测样点对三种模型进行验证与分析. 其中,Simis 半分析模型主要采用 709 nm/665 nm、709 nm/620 nm 波段比值分别求出叶绿素 a 在 650 nm 的吸收系数和 620 nm 处藻蓝蛋白与叶绿素 a 的吸收系数和,进而求出藻蓝蛋白在 620 nm 处的吸收系数,最后通过藻蓝蛋白比吸收系数值求出藻蓝蛋白的浓度,本文在应用 Simis 算法过程中重新率定相关参数;齐琳 PCI 指数方法则是基于藻蓝蛋白在 620 nm 附近的反射谷特征,定义藻蓝素指数(PCI)为 560 nm 和 665 nm 之间的基线高度减去 620 nm 的波段反射率,随后使用非线性最小二乘拟合建立的 PCI 与藻蓝蛋白之间的关系. Simis 半分析模型、PCI 指数模型均使用相同的 76 个数据建模,33 个数据进行验证分析.

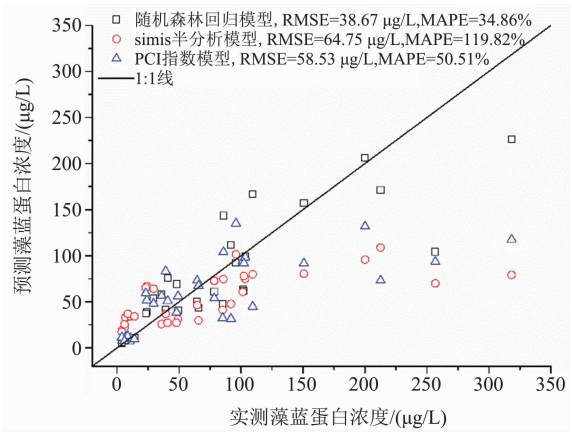

图 5 三种算法反演藻蓝蛋白浓度散点图
Fig. 5 Scatter diagram of phycocyanin concentration retrieved by three algorithms

如图 5 所示,随机森林反演模型拟合线大多数样点聚集在 1∶1 线周围,有较高的拟合度,同时 MAPE 为 34.86%,RMSE 为 38.67 μg/L,反演精度较好,以 100 μg/L 浓度为界限,浓度低于 100 μg/L 的样本分布比高于 100 μg/L 浓度的样本更贴近 1∶1线,表明随机森林在低中浓度时反演精度更好;但是 Simis 半分析模型中样点在高浓度时分布散乱,其 R^2 为 0.49,RMSE 为 64.75 μg/L,MAPE 为 119.92 %,当藻蓝蛋白(PC)浓度较低时(<25 μg/L),出现明显高估现象,当藻蓝蛋白浓度较高(>150 μg/L)时,其 Simis 半分析模型低估现象严重;PCI 指数模型样点大多分布在 1∶1线附近,与 Simis 算法一样,PCI 指数模型在高浓度(>150 μg/L)时,误差明显增大,但其在低浓度时精度明显优于 Simis 半分析模型,R^2、RMSE 和 MAPE 都有不同程度的提高,但是与随机森林反演模型相比,其 R^2 降低了 0.23 左右,RMSE 降低了 20%,尤其 MAPE 变化显著,总体来看随机森林反演模型的反演精度都要优于 simis 半分析模型和 PCI 指数模型. 这是由于随机森林并不是简单的数值拟合,在分析多个波段遥感反射率时,具有更好的灵活性和预测性,无论藻蓝蛋白浓度低或高,其反演精度都较好.

3.3 随机森林反演模型的普适性验证

为了评价随机森林反演模型,利用 2017 年 08 月巢湖采集的 19 个样点的遥感反射率和藻蓝蛋白浓度数据进行模型的验证和精度评价.

首先利用随机森林反演模型,将太湖、洪泽湖和滇池的 109 个实测数据的遥感反射率和藻蓝蛋白浓

度作为训练集得到回归模型,输入 2017 年 08 月巢湖的 19 个样点的遥感反射率,估算得到各样点相应的藻蓝蛋白浓度,然后利用 RMSE 和 MAPE 对估算得到的藻蓝蛋白浓度和实测藻蓝蛋白浓度进行对比分析. 如图 6 所示,藻蓝蛋白的实测值和估算值的散点图较好的分布在 1∶1线附近,RMSE = 10.72 μg/L,MAPE =22%,这表明利用随机森林反演模型反演藻蓝蛋白浓度在巢湖是可行的.

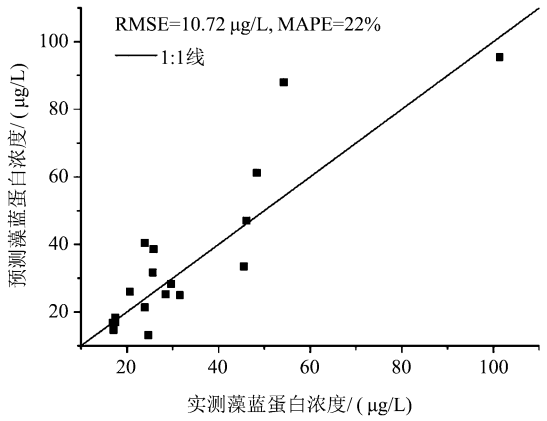

图 6　随机森林反演藻蓝蛋白浓度散点图
Fig. 6　Random forest inversion of Phycocyanin concentration scatter plot

3.4　哨兵 3A-OLCI 大气校正评价

为验证基于光谱响应函数的波段模拟反射率值和卫星数据真实遥感反射率值的一致性,利用实测同步点数据对大气校正效果进行定量评估,从而确定构建的随机森林反演模型能够直接用于影像上,以期说明模型在遥感影像上的通用性.

利用 SEADAS 7.4 对 OLCI 影像进行大气校正,并通过与卫星获取时间准同步的 10 个实测样点对大气校正后的影像进行精度验证与评价,大气校正算法主采用的是 MUMM 算法. 结果如下表所示.

如表 1 所示,影像各波段遥感反射率 MAPE 值和 rRMSE 变化较大,在可见光 400 ~ 510 nm 即蓝绿波段处其 rRMSE 与 MAPE 值较大,校正效果较差,同时近红外波段 865 ~ 885 nm 处 MAPE 值高达 64.53%,误差较大,但在可见光 560 ~ 779 nm 处 MAPE 均值为 16.70%,rRMSE 均值为 0.1903,表明经实测遥感反射率(Rrs)模拟得到 560 ~ 779 nm 波段处的 Rrs 与大气校正后的 OLCI 影像各波段 Rrs 拟合度较好. 同时利用光谱角度匹配(SAM)的方法对大气校正效果进行评价,即利用成像光谱数据提取的地物光谱曲线和光谱库(实验室或者野外测量的标准曲线)的光谱曲线的进行匹配分析. 该方法充分利用了光谱维的信息,强调了光谱的相似特征. 该方法表示为如下:

$$\theta = \arccos \frac{\sum_{i=1}^{n} X_i \gamma_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \sqrt{\sum_{i=1}^{n} \gamma_i^2}} \quad \theta \in [0, \frac{\pi}{2}] \tag{11}$$

式(11)中,χ_i 为第 i 个波段值,γ_i 为第 i 波段遥感影像光谱值,θ 为光谱角度.

计算遥感影像光谱曲线与实测样点光谱曲线的角度,当光谱夹角即 θ 值越小,相似度越大,则实测样点光谱曲线的形状与遥感影像光谱曲线保持一致. 对 560 ~ 779 nm 的 8 个波段的实测光谱曲线与 OLCI 影像光谱曲线分析,10 个同步验证点的光谱角度 θ 值分别为 1.86°、3.24°、2.97°、2.62°、3.47°、4.54°、3.24°、4.11°、3.83°、3.90°,刘天乐[30]等人认为当 $\theta = 0°$ 时,表示两个光谱完全相似;当 $\theta = \pi/2$ 时,则光谱完全不同. 因此光谱角结果显示大气校正后的的光谱曲线与测光谱曲线形状大致保持一致(见图 7).

因此根据哨兵 3A-OLCI 波段设置及大气校正结果,发现 560 nm、620 nm、665 nm、674 nm、681 nm、709 nm、754 nm、779 nm 八个波段的实测遥感反射率与大气校正后的 OLCI 影像 Rrs 拟合度较好,同时光谱曲线形状保持一致,证明经过 MUMM 大气校正算法处理后 OLCI 遥感影像能够利用实测数据所构建的随机森林反演模型对藻蓝蛋白浓度进行估算,并同时将这 8 个波段的遥感反射率作为随机森林反演模型的输入自变量.

3.5　藻蓝蛋白空间分布格局分析

利用随机森林反演藻蓝蛋白浓度模型对已经过大气校正的 2016 年 12 月 8 日、2017 年 4 月 12 日和 2017 年 5 月 18 日的三景哨兵 3A-OLCI 影像进行分

表 1　大气校正后影像各波段遥感反射率精度评价
Table 1　Estimation of reflectance of remote sensing images after atmospheric correction

波段/nm	400	412	442	490	510	560	620	665	674	681	709	754	779	865	885
MAPE/%	46.49	73.78	48.14	32.01	27.06	14.43	12.92	11.87	11.76	11.68	16.50	27.46	26.94	37.87	64.53
rRMSE	0.4718	0.7598	0.5016	0.3334	0.3130	0.1693	0.1514	0.1366	0.1375	0.1357	0.1884	0.3040	0.2994	0.4121	0.631

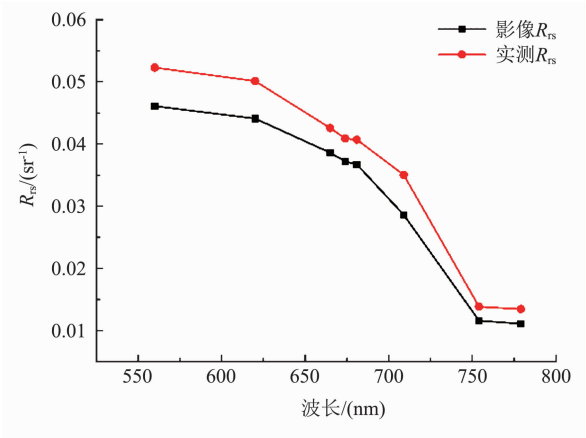

图 7 同步样点实测光谱曲线与影像光谱曲线

Fig. 7 Measured spectral curve and image spectral curve of synchronous sampling points

析. 从图 8 中可以看出,洪泽湖、滇池和太湖藻蓝蛋白分布明显. 洪泽湖藻蓝蛋白浓度整体偏低,平均值为 20.71 μg/L,其中成子湖区域(西北区域)藻蓝蛋白浓度最高,成子湖相对封闭,是主要的围网养殖区域,换水较慢,富营养化程度较高[31-32]. 近岸区域和洪泽湖河口区域藻蓝蛋白浓度值也较高,这是因为河口区水域换水周期频繁,水草较多,有利于藻类的生长,近岸区域水深较浅,较低的风速会减少水体的扰动,导致更多的光进入水体,藻类的初级生产力较高,导致了藻蓝蛋白浓度增加湖心湾区域即洪泽湖主要的过水区域水体浑浊且湖流扰动剧烈,其藻蓝蛋白浓度最低[31-32];太湖藻蓝蛋白浓度一般在 10 ~ 20 μg/L 之间,太湖贡湖湾藻蓝蛋白浓度普遍较高,最高值达到 80 μg/L. 西部藻蓝蛋白浓度偏低,由于镇湖湾和光福湾附近水深较浅,属于水生植被区,这些区域的沉水植被、浮叶植被交替生长,对遥感反射率比造成影响,因而这些区域的色素浓度可能与实际情况有些偏差[4,33-34];竺山湾和梅梁湾区域是太湖污染或水华常发地带,从图 7 可知,该区域藻蓝蛋白浓度平均值为 40 μg/L,在藻蓝蛋白浓度较高区域,通过影像真彩色显示,可以发现有水华现象发生,同时在 2017 年 5 月份中旬太湖盛行西北风,导致竺山湾区域表层藻蓝蛋白被西北风输送到湾内;4 月份滇池蓝藻水华面积开始逐渐升高[35],其中藻蓝蛋白浓度平均值在 60 μg/L,滇池北部的龙门村、海埂、盘龙江入湖河口一带和南部晋宁县附近的藻蓝蛋白浓度高度 80 μg/L,同时滇池流域常年盛行西南风,从而导致滇池东部呈贡区藻蓝蛋白浓度值较高. 张虎才[36]研究结果表明滇池藻蓝蛋白浓度整体北部高于南部,这在哨兵 3A-OLCI 影像中得到了响应.

4 结论

研究基于哨兵 3A-OLCI 影像波段,构建了适合我国富营养化水体的藻蓝蛋白随机森林反演模型,并对随机森林反演模型与常用的 Simis 半分析模型、PCI 指数模型在内陆湖泊水域藻蓝蛋白浓度估算精度进行对比分析,发现随机森林的 MAPE 仅为 34.86%,远低于 Simis 半分析模型和 PCI 指数模型,RMSE 为 38.67ug/L ,明显的优于其他两种模型. 同时通过独立的巢湖数据集验证表明,该模型在巢湖的反演精度令人满意,表明该模型具有较好的通用性. 本研究所构建的随机森林反演模型可以用来反演我国内陆富营养化湖泊的藻蓝蛋白浓度,取得了令人满意的效果.

对遥感影像进行大气校正是水质参数反演的前提. 哨兵 3A-OLCI 影像作为一种全新的水色遥感数据源,目前还缺少比较成熟的大气校正算法,对 OLCI 影像进行有效的大气校正是直接关系到 OLCI 影像光谱匹配的正确性和反演结果的正确性. 本研究虽采用目前常用的大气校正算法 MUMM,但是部分波段校正效果不是十分理想,对反演结果产生一定的影响.

References

[1] Guo L. Doing Battle With the Green Monster of Taihu Lake [J]. *Science*, 2007, **317**(5842):1166.

[2] Stone R. Ecology. China aims to turn tide against toxic lake pollution[J]. *Science*, 2011, **333**(6047):1210.

[3] Richardson L L. Remote Sensing of Algal Bloom Dynamics [J]. *Bioscience*, 1996, **46**(7):492 - 501.

[4] Ma Rong-Hua, KONG Wei-Juan, DUAN Hong-Tao, *et al*. Quantintative estimation of Phycocyanin concentration using MODIS imagery during the period of cyanobacterial blooming in Taihu Lake [J]. *China Environmental Science*, (马荣华, 孔维娟, 段洪涛, 等. 基于 MODIS 影像估测太湖蓝藻暴发期藻蓝素含量. **中国环境科学**), 2009, **29**(3):254 - 260.

[5] Jupp D, Kirk J, Harris G P, *et al*. Detection, identification and mapping of cyanobacteria—Using remote sensing to measure the optical quality of turbid inland waters[J]. *Marine & Freshwater Research*, 1994, **45**(5):135 - 53.

[6] Kutser T, Metsamaa L, Strömbeck N, *et al*. Monitoring cyanobacterial blooms by satellite remote sensing[J]. *Estuarine Coastal & Shelf Science*, 2006, **67**(1-2):303 - 312.

[7] Hunter P D, Tyler A N, Carvalho L, *et al*. Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes. [J]. *Remote Sensing of Environment*, 2010, **114**(11):2705 - 2718.

[8] Dekker A G. Detection of optical water quality parameters

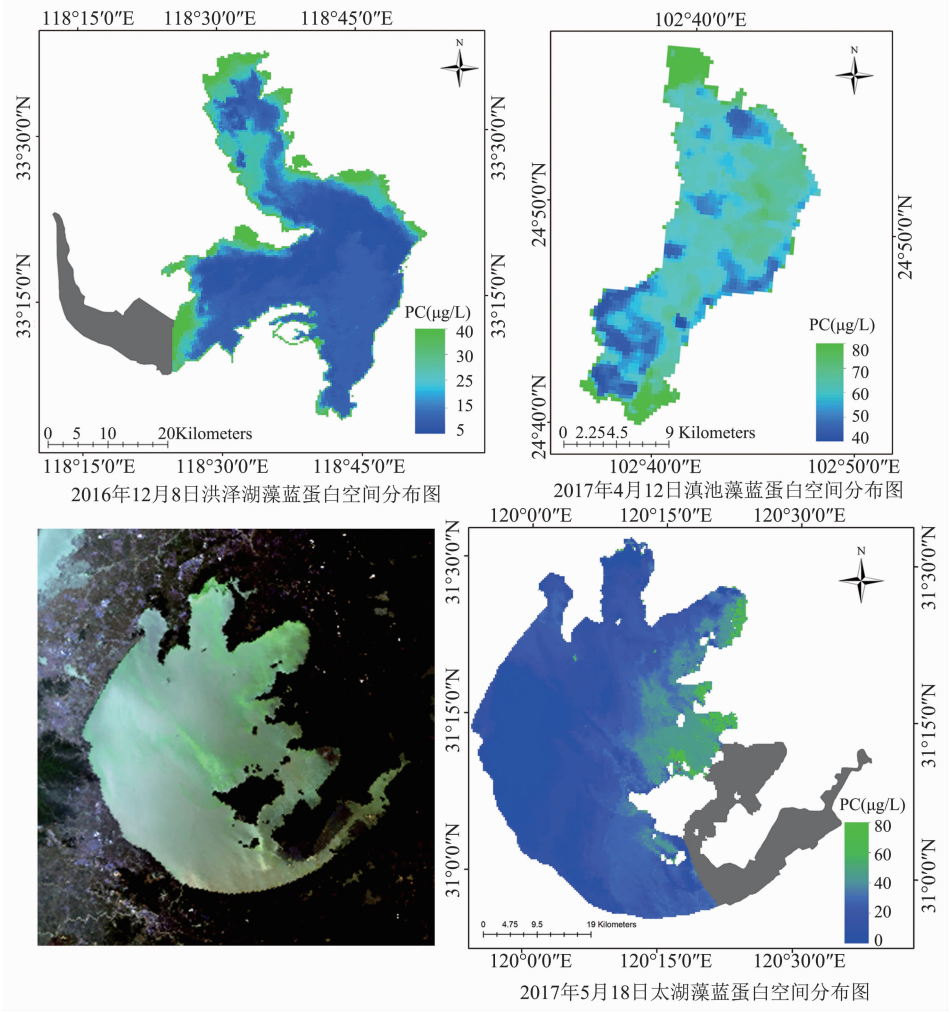

图 8　哨兵 3A-OLCI 影像太湖真彩色显示图和洪泽湖、滇池、太湖藻蓝蛋白分布图

Fig. 8　The true color display map and distribution map of phycocyanin of the sentinel 3A-OLCI image in Hongzehu Lake, Dianchi Lake and Taihu Lake

for eutrophic waters by high resolution remote sensing[J]. *Amsterdam Vrije Universiteit*, 1993.

[9] Schalles J F, Yacobi Y Z. Remote detection and seasonal patterns of phycocyanin, carotenoid and chlorophyll pigments in eutrophic waters[J]. *Ergebnisse Der Limnologie*, 2000, **55**:153 - 168.

[10] ZKStefan G H. Simis. Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water[J]. *Limnology & amp; Oceanography*, 2005 , **50**(1): 237 - 245.

[11] Stefan G. H. Simis, Antonio Ruiz-Verdú, Ramón Peña-Martinez, Herman J. Gons. Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass[J]. *Remote Sensing of Environment*, 2007, **106**(4):414 - 427.

[12] YIN Bin, Lyu Heng, Li Yun-mei, *et al.* Retrieve Phycocyanin Concentrations Based on Semi-analytical Model in the Dianchi Lake, China[J]. *Environment Science*, (尹斌, 吕恒, 李云梅, 等. 基于半分析模型的滇池藻蓝蛋白浓度反演[J]. **环境科学**), 2011, **32**(2):472 - 478.

[13] Qi L, Hu C, Duan H, *et al.* A novel MERIS algorithm to derive cyanobacterial phycocyanin pigment concentrations in a eutrophic lake: Theoretical basis and practical considerations[J]. *Remote Sensing of Environment*, 2014, **154**: 298 - 317.

[14] Liaw A, Wiener M. Classification and regression by randomForest[J]. *R news*, 2002, **2**(3):18 - 22.

[15] Leo Breiman. Bagging Predictors[J]. *Machine Learning*, 1996, **24**(2):123 - 140.

[16] Ho T K. The random subspace method for constructing decision forests[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1998, **20**(8):832 - 844.

[17] Grimm R, Behrens T, Märker M, *et al.* Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis[J]. *Geoderma*, 2008, **146**(1): 102 - 113.

[18] BAI Lin, XU Yong-ming, HE Miao, *et al.* Remote Sensing Inversion of Near Surface Air Temperature Based on Random Forest[J]. *Journal of Geo-information Science*, (白琳, 徐永明, 何苗, 等. 基于随机森林算法的近地表气温遥感反演研究. **地球信息科学学报**), 2017, **19**(3): 390 - 397.

[19] Zhang Ying, Gao Qian-qian. Water quality evaluation of

Chaohu Lake based on random forest method[J]. *Chinese Journal of Environment Engineering*,(张颖, 高倩倩. 基于随机森林分类算法的巢湖水质评价. **环境工程学报**), 2016, **10**(2):992-998.

[20] SHI Hao, LI Xuwen, NIU Zhichun, *et al*. Remote sensing information extraction of aquatic vegetation in Lake Taihu based on Random Forest Model[J]. *Lake Science*,(侍昊, 李旭文, 牛志春, 等. 基于随机森林模型的太湖水生植被遥感信息提取. **湖泊科学**), 2016, **28**(3): 635-644.

[21] JIANG jia-le, LIU Xiang-nan, LIU Mei-qing, *et al*. Remote sensing retrieval model of sea surface salinity in Hong Kong waters based on the random forest[J]. *Marine Science Bulletin*,(江佳乐, 刘湘南, 刘美玲, 等. 基于随机森林的香港海域海表盐度遥感反演模型. **海洋通报**), 2014, **33**(3):333-341.

[22] Lyu H, Li X, Wang Y, *et al*. Evaluation of chlorophyll-a retrieval algorithms based on MERIS bands for optically varying eutrophic inland lakes.[J]. *Science of the Total Environment*, 2015, s530-531:373-382.

[23] Lyu H, Wang Q, Wu C, *et al*. Variations in optical scattering and backscattering by organic and inorganic particulates in Chinese lakes of Taihu, Chaohu and Dianchi[J]. *Chinese Geographical Science*, 2015, **25**(1):26-38.

[24] Shi K, Li Y, Li L, *et al*. Absorption characteristics of optically complex inland waters: Implications for water optical classification[J]. *Journal of Geophysical Research Biogeosciences*, 2013, **118**(2):860-874.

[25] TANG Jun-wu, TIAN Guo-liang, WANG Xiao-yong, *et al*. The Methods of Water Spectral Measurement and Analysis Ⅰ:Above-water Method [J]. *Journal of Remote Sensing*,(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.

[26] Team, Sentinel-3. Sentinel-3 User Handbook[M]. *ESA Standard Document*, 2013.

[27] Ruddick K G, Ovidio F, Rijkeboer M. Atmospheric correction of SeaWiFS imagery for turbid coastal and inland waters[J]. *Applied optics*, 2000, **39**(6):897-912.

[28] FANG Kuang-nan. Random Forest Portfolio Forecasting Theory and Its Applicaton in Finance[M]. Xiamen University Press,(方匡南. 随机森林组合预测理论及其在金融中的应用. 厦门大学出版社), 2012.

[29] LI Xu-qing, LIU Xiangnan, LIU Meiling, *et al*. Random forest algorithm and regional applications of spectral inversion model for estimating canopy nitrogen concentration in rice[J]. *Journal of Remote Sensing*,(李旭青, 刘湘南, 刘美玲, 等. 水稻冠层氮素含量光谱反演的随机森林算法及区域应用. **遥感学报**), 2014, **18**(4):923-945.

[30] LIU Tian-le. Study of spectral angle classification based on Hyperspectral Remote Sensing Images[A]. *Chinese Society for Geodesy, Photogrammetry and Cartography*,(刘天乐. 基于高光谱的遥感图像的光谱角度分类方法的研究. 中国测绘学会),2006:5.

[31] LIU Ge, LI Yun-mei, LYU Heng ,*et al*. Remote Sensing of Chlorophyll-a Concentration in Lake Hongze Using Long Time Series MERIS Observations[J]. *Environment Science*,(刘阁,李云梅,吕恒, 等. 基于 MERIS 影像的洪泽湖叶绿素 a 浓度时空变化规律分析. **环境科学**), 2017, **38**(09): 3645-3656.

[32] Ren Y, Pei H, Hu W, *et al*. Spatiotemporal distribution pattern of cyanobacteria community and its relationship with the environmental factors in Hongze Lake, China.[J]. *Environmental Monitoring & Assessment*, 2014, **186**(10):6919.

[33] KONG Wei-juan, MA Rong-hua, DUAN Hong-tao, *et al*. Monitoring Cyanobacterial Blooms Using MODIS Images in Taihu Lake,China[J]. *Remote Sensing Information*,(孔维娟, 马荣华, 段洪涛, 等. 太湖秋冬季蓝藻水华 MODIS 卫星遥感监测. **遥感信息**), 2009, **4**: 80-84.

[34] YIN Bin. Estimation of Cyanobacteria in Taihu lake Based on MERIS Image[D]. *Nanjing Normal University*,(尹斌. 基于 MERIS 数据的太湖蓝藻估算研究. 南京师范大学), 2011.

[35] JIANG Da-lin. Spatio temporal change of algal bloom and its driving factors in Dianchi based on GIS/RS[D]. *Southwest University*,(蒋大林. 基于 GIS/RS 的滇池藻类水华时空变化.驱动因子分析. 西南大学),2015.

[36] ZHANG Hu-cai, CHANG Feng-qin, *et al*. Water Quality Characteristics and Variations of Lake Dianchi[J]. *Advances in Earth Science*,(张虎才, 常凤琴, 等. 滇池水质特征及变化. **地球科学进展**), 2017, **32**(06):651-659.