基于随机森林算法的卫星监测太湖蓝藻数据集（2019）

杨 子[1,2]，潘 鑫[2,3*]，袁 洁[1,2]，宋 昊[1,2]，
许 坤[1,2]，吴宇航[1,2]，杨英宝[2,3*]

1. 河海大学地球科学与工程学院，南京 211100；
2. 河海大学江苏省水资源环境遥感监测评估工程研究中心，南京 211100；
3. 河海大学地理与遥感学院，南京 210098

摘 要：太湖蓝藻数据是太湖水资源治理必不可少的重要数据。作者选取 2019 年太湖的高分六号卫星的红外、近红外、绿波段数据，采用基于多遥感因子（归一化植被指数和归一化水体指数）的随机森林算法对蓝藻信息进行提取，得到太湖西部蓝藻数据。采用总体分类精度、Kappa 系数、生产者精度、用户精度、错分误差和漏分误差对数据集进行验证，验证结果表明，此数据集总体分类精度和 Kappa 系数的均值达到 0.97 和 0.95。数据集内容包括：2019 年 5–12 月 6 个时期的蓝藻分布数据。数据集的空间分辨率为 20 m。数据集存储为.tif 格式，由 6 个数据文件组成，数据量为 0.98 MB（压缩为 1 个文件，601 KB）。

关键词：太湖；蓝藻；随机森林；2019 年
DOI: https://doi.org/10.3974/geodp.2023.03.11
CSTR: https://cstr.escience.org.cn/CSTR:20146.14.2023.03.11
数据可用性声明：
本文关联实体数据集已在《全球变化数据仓储电子杂志（中英文）》出版，可获取：
https://doi.org/10.3974/geodb.2023.12.01.V1 或 https://cstr.escience.org.cn/CSTR:20146.11.2023.12.01.V1.

1 前言

太湖是我国第二大淡水湖，随着经济的快速发展和工业的扩张，太湖的浮游植物大量繁殖造成蓝藻爆发。这破坏了太湖的生态环境，影响了周边城市的生活用水。因此对太湖的蓝藻爆发监测具有重要意义[1,2]。

蓝藻传统的监测方法是实地采集水样，但这种方法对人力物力的需求高，且能够采集的样点数量有限。而遥感技术有着覆盖面积大，低成本，快速及时的特点，因此近年来基于遥感影像进行蓝藻监测是国内外学者关注的研究方向。高分六号是我国首颗具有红边波

收稿日期：2023-07-15；**修订日期：**2023-09-21；**出版日期：**2023-09-25
基金项目：国家自然科学基金（41701487，42071346，42371397）
***通讯作者：**潘鑫，河海大学地理与遥感学院，px1013@hhu.edu.cn
杨英宝，河海大学地理与遥感学院，yyb@hhu.edu.cn
数据引用方式：[1] 杨子，潘鑫，袁洁等. 基于随机森林算法的卫星监测太湖蓝藻数据集（2019）[J]. 全球变化数据学报，2023，7(3)：321–326. https://doi.org/10.3974/geodp.2023.03.11. https://cstr.escience.org.cn/CSTR:20146.14.2023.03.11.
[2] 杨子，潘鑫，袁洁等. 基于随机森林算法的卫星监测太湖蓝藻数据集（2019）[J/DB/OL]. 全球变化数据仓储电子杂志，2023. https://doi.org/10.3974/geodb.2023.12.01.V1. https://cstr.escience.org.cn/CSTR:20146.11.2023.12.01.V1.

段的光学成像卫星，能够较好地监测植被，而蓝藻和植被具有相似的光谱信息，因此基于高分六号卫星数据生成太湖蓝藻数据集具有重要意义。

基于遥感数据提取蓝藻的方法主要有经验模型法、阈值法和机器学习法[3–7]。其中经验模型法主要是建立指数和实测叶绿素浓度的非线性或线性的模型，再通过叶绿素浓度来监测蓝藻，但此方法具有区域局限性[8, 9]。阈值法是通过设置单波段或者指数的阈值来对蓝藻进行提取，但阈值的准确选择是一个难点[10, 11]。机器学习方法是基于特征指标来对蓝藻进行分类提取，具有挖掘大数据的规律特点，其包括支持向量机和随机森林法等[12]。随机森林法是机器学习中较为流行的方法，已经成功应用于参数的反演。

因此基于高分六号数据采用随机森林的方法对蓝藻进行提取更能满足未来长时序、高精度地进行太湖蓝藻的动态监测。本数据集基于 2019 年高分六号影像质量较好的太湖影像，采用随机森林法实现了太湖蓝藻的提取，生成了 2019 年太湖蓝藻的数据产品。此数据集可以为后续应用提供相应的基础支撑。

2 数据集元数据简介

《基于随机森林算法的卫星监测太湖蓝藻数据集（2019）》[13]的名称、作者、地理区域、数据年代、时间分辨率、空间分辨率、数据集组成、数据出版与共享服务平台、数据共享政策等信息见表 1。

3 数据研发方法

3.1 算法原理

数据集主要采用的是随机森林（Random Forest，RF）算法，是一种特殊的套袋算法，是由 Leo[15]提出的一种可以进行分类或回归的算法。RF 方法与原始的套袋算法的不同之处在于它使用决策树作为模型，本文采用的是 RF 的分类算法，通过决策树进行投票，获得投票最多的类为当前对象所属的类。RF 模型的建立需要输入变量和分类结果隶属的代表数值，最后基于 RF 模型对结果进行预测[16]。

此数据集选取的输入变量是归一化植被指数和归一化水体指数，因为理想地认为太湖只存在水体和蓝藻两种地物，NDVI（Normalized Differential Vegetation Index，NDVI）和 NDWI（Normalized Differential Water Index，NDWI）能够较好地分别识别植被和水体。因此将 NDVI 和 NDWI 作为 RF 的输入变量。

$$NDVI = \frac{\rho_{Nir} - \rho_{Red}}{\rho_{Nir} + \rho_{Red}} \tag{1}$$

$$NDWI = \frac{\rho_{Green} - \rho_{Nir}}{\rho_{Green} + \rho_{Nir}} \tag{2}$$

式中，ρ_{Red} ρ_{Nir} ρ_{Green} 分别代表的是红、近红和绿波段的反射率。根据 NDVI 指数的波段要求，高分六号卫星选择第 3 波段和第 4 波段进行 NDVI 指数的计算。根据 NDWI 指数的波段要求，高分六号卫星选择第 2 波段和第 4 波段进行 NDWI 指数的计算。

表 1　《基于随机森林算法的卫星监测太湖蓝藻数据集（2019）》元数据简表

条　目	描　述
数据集名称	基于随机森林算法的卫星监测太湖蓝藻数据集（2019）
数据集短名	Taihu_Cyanobacteria
作者信息	杨子，河海大学地球科学与工程学院，河海大学江苏省水资源环境遥感监测评估工程研究中心，18339161755@163.com 潘鑫，河海大学地理与遥感学院，河海大学江苏省水资源环境遥感监测评估工程研究中心，px1013@hhu.edu.cn 袁洁，河海大学地球科学与工程学院，河海大学江苏省水资源环境遥感监测评估工程研究中心，yj000801@163.com 许坤，河海大学地球科学与工程学院，河海大学江苏省水资源环境遥感监测评估工程研究中心，919505610@qq.com 吴宇航，河海大学地球科学与工程学院，河海大学江苏省水资源环境遥感监测评估工程研究中心，yuhangwu2022@163.com 杨英宝，河海大学地理与遥感学院，河海大学江苏省水资源环境遥感监测评估工程研究中心，yyb@hhu.edu.cn
地理区域	太湖，地理范围包括 30°55′40″N–31°32′58″N，119°52′32″E–120°36′10″E
数据年代	2019 年
时间分辨率	2 天
空间分辨率	20 m
数据格式	短整型格式（数值为 0，1，Nodata;1 代表蓝藻，0 代表水体，Nodata 为非研究区）
数据量	0.98MB
数据集组成	数据集包括 6 个文件，分别是 2019 年 6 期（5 月 5 日、7 月 29 日、9 月 13 日、10 月 29 日、11 月 5 日、12 月 12 日）的太湖蓝藻影像
基金项目	国家自然科学基金（41701487，42071346，42371397）
数据计算环境	ENVI、Matlab
出版与共享服务平台	全球变化科学研究数据出版系统 http://www.geodoi.ac.cn
地址	北京市朝阳区大屯路甲 11 号 100101，中国科学院地理科学与资源研究所
数据共享政策	（1）“数据”以最便利的方式通过互联网系统免费向全社会开放，用户免费浏览、免费下载；（2）最终用户使用“数据”需要按照引用格式在参考文献或适当的位置标注数据来源；（3）增值服务用户或以任何形式散发和传播（包括通过计算机服务器）“数据”的用户需要与《全球变化数据学报（中英文）》编辑部签署书面协议，获得许可；（4）摘取“数据”中的部分记录创作新数据的作者需要遵循 10%引用原则，即从本数据集中摘取的数据记录少于新数据集总记录量的 10%，同时需要对摘取的数据记录标注数据来源[14]
数据和论文检索系统	DOI，CSTR，Crossref，DCI，CSCD，CNKI，SciEngine，WDS/ISC，GEOSS

3.2　技术路线

此数据集的技术路线图如图 1 所示，主要分为三个步骤分别是数据准备、数据预处理和 RF 模型的建立和预测。其中数据准备主要是是筛选云量较少的 2019 年高质量的高分六号数据。数据预处理主要是对原始的高分六号数据进行辐射校正、大气校正和几何校正，因为原始的高分六号数据是卫星获取的 DN 值，需要通过预处理转化成地表的反射率值。RF 模型的建立和预测过程主要是基于预处理后的高分六号数据计算 NDVI 和 NDWI 的值，将两个因子作为 RF 的训练数据建立模型，最后对太湖蓝藻进行预测得到太湖蓝藻数据集。

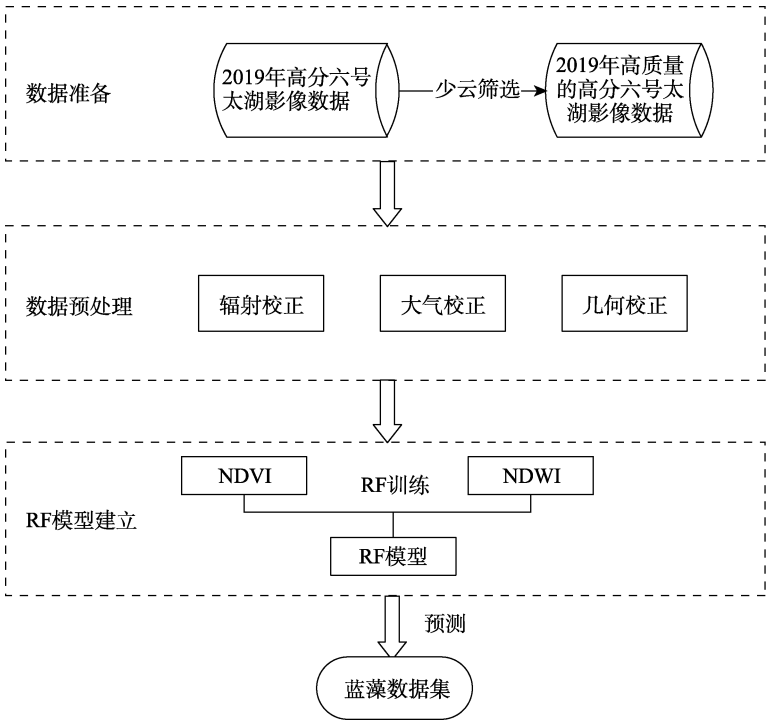

图 1 卫星监测太湖蓝藻数据集研发技术路线图

4 数据结果与验证

4.1 数据集组成

《基于随机森林算法的卫星监测太湖蓝藻数据集（2019）》，共 1 个文件夹。文件夹中包含 6 期蓝藻数据，数据格式为标签图像文件格式（.tif）。

4.2 数据结果

图 2 是 2019 年太湖蓝藻结果的空间分布图，其中蓝色和绿色分别代表水体和蓝藻。由于东太湖存在水生植被，因此东太湖的空间分布不予以展示。在 2019 年 5 月 5 日，太湖蓝藻主要集中在梅梁湖的南部和西北沿岸区。7 月 29 日太湖蓝藻的爆发情况明显减弱，蓝藻爆发集中在竺山湖和梅梁湖，贡湖存在少许的蓝藻。相比于 7 月 29 日，9 月 13 日太湖的西北、西南和南部沿岸区蓝藻开始爆发，并且梅梁湖蓝藻持续爆发。10 月 19 日，太湖蓝藻全面爆发，主要集中在太湖的中部和梅梁湖。11 月 5 日，太湖蓝藻爆发急速减弱，贡湖基本上没有蓝藻爆发，爆发的蓝藻主要在梅梁湖和西北沿海岸。12 月 12 日，太湖几乎很少蓝藻，零星的蓝藻主要集中在西南沿岸区。这说明 2019 年太湖蓝藻主要在秋季爆发，春季和夏季蓝藻爆发情况一般，冬季蓝藻几乎没有爆发。

4.3 数据结果验证

4.3.1 验证方法

引入混淆矩阵来计算总体分类精度、Kappa 系数、生产者精度、用户精度、错分误差和漏分误差[17]六种指标来对提取的结果进行评定。此数据集验证由近红、红和绿波段合成的假彩色影像作为参考，通过目视解译的方式均匀选取 300 个样本点（150 个蓝藻样点和

150 个水体样点）作为验证。

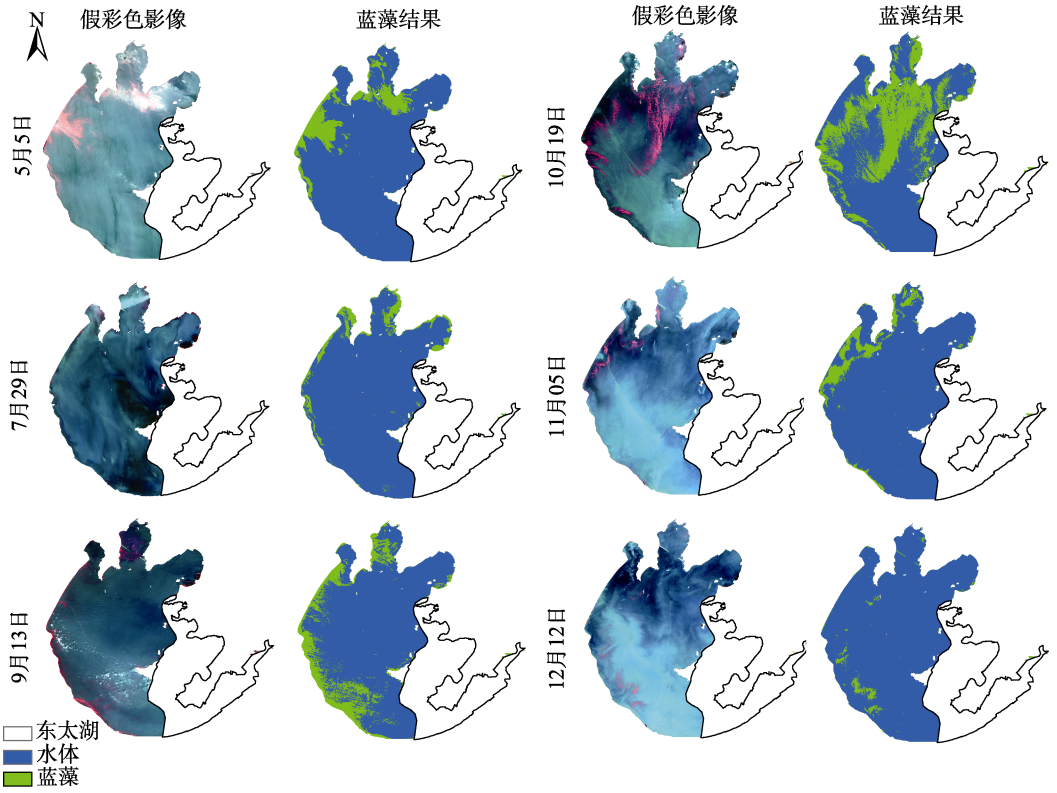

图 2　2019 年部分太湖蓝藻空间分布图

4.3.2　验证结果

本数据集的精度评定表 2 所示，可以看出此数据集的总体分类精度均值达到 0.97，Kappa 系数均值也达到了 0.95。这说明此数据集的精度较高。尤其是 2019 年 12 月 12 日的太湖蓝藻结果总体分类精度、生产者精度和用户精度都达到了 0.99，2019 年 12 月 12 日的太湖蓝藻结果精度是 6 期太湖影像中最高的。5 月 5 日太湖蓝藻结果的精度相对略低，总体分类精度、Kappa 系数和生产者精度分别达到了 0.95、0.91 和 1.00，但 5 月 5 日太湖蓝藻结果有将水体错分为蓝藻的情况，这可能是由于 5 月 5 日的太湖影像有薄云的影响。总之，从表 2 可以看出基于高分六号数据采用 NDVI 和 NDWI 的 RF 方法能够较好地提取出太湖的蓝藻。

表 2　基于高分六号卫星太湖蓝藻数据集（2019）的精度评价结果统计表

日期	总体分类精度(%)	Kappa	生产者精度	用户精度	错分误差	漏分误差
5 月 5 日	0.95	0.91	1.00	0.92	0.08	0.00
7 月 29 日	0.98	0.97	1.00	0.97	0.03	0.00
9 月 13 日	0.97	0.93	1.00	0.94	0.06	0.00
10 月 19 日	0.96	0.98	1.00	0.96	0.04	0.00
11 月 5 日	0.97	0.95	0.98	0.97	0.03	0.02
12 月 12 日	0.99	0.98	0.99	0.99	0.01	0.01
均值	0.97	0.95	0.99	0.95	0.04	0.01

5 讨论和总结

由于以往对太湖蓝藻提取的研究中，较少将国内高分系列卫星影像作为数据源，故本设计采用 RF 法基于高分六号数据得到高质量的 2019 年太湖蓝藻数据集。本数据集生产经过原始高分六号数据的辐射校正、大气校正、几何校正及 RF 训练和预测。该数据产品包含原始影像质量相对较好的 2019 年太湖蓝藻的结果，空间分辨率为 20 m，数据格式为标签图像文件格式（.tif）。

本数据集对太湖的环境治理提供了数据支撑，对太湖的蓝藻动态监测具有重要意义。相比于传统的阈值法提取蓝藻，本产品的算法避免了阈值取值的不确定性，且省时省力。但本产品算法的精度与训练样本的数量和分布有较强的联系，且本产品的算法未来要考虑云对结果的影响。

作者分工：潘鑫和杨英宝对数据集的开发做了总体设计；杨子和袁洁采集和处理了高分六号的原始数据；杨子设计了模型和算法；宋昊、许坤和吴宇航做了数据验证；杨子撰写了数据论文。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献

[1] 秦伯强, 高光, 朱广伟等. 湖泊富营养化及其生态系统响应[J]. 科学通报, 2013, 58(10): 855–864.
[2] 祝令亚. 湖泊水质遥感监测与评价方法研究[D]. 北京: 中国科学院遥感应用研究所, 2006.
[3] 杨运, 航王琳, 谢洪波等. 基于 Landsat8 的含蓝藻湖泊水体信息提取方法研究[J]. 国土资源遥感, 2020, 32(4): 130–136.
[4] 王萌, 郑伟, 刘诚. 利用 Himawari-8 高频次监测太湖蓝藻水华动态[J]. 湖泊科学, 2017, 29(5): 1043–1053.
[5] 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[J]. 湖泊科学, 2016, 28(3): 635–644.
[6] 夏晓瑞, 韦玉春, 徐宁等. 基于决策树的 Landsat TM/ETM+图像中太湖蓝藻水华信息提取[J]. 湖泊科学, 2014, 26(6): 907–915.
[7] 黄家柱, 赵锐. 卫星遥感监测太湖水域蓝藻暴发[J]. 遥感信息, 1999(4): 43–44.
[8] 李旭文, 侍昊, 张悦等. 基于欧洲航天局“哨兵-2A”卫星的太湖蓝藻遥感监测[J]. 中国环境监测, 2018, 34(4): 169–176.
[9] 苗松, 王睿, 李建超等. 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[J]. 红外与毫米波学报, 2018, 37(5): 621–630.
[10] 李晓俊, 吕恒, 李云梅等. 基于 MODIS 与 GOCI 数据蓝藻水华提取空间尺度差异分析[C]. 江苏省海洋湖沼学会, 2013.
[11] 李亚春, 孙佳丽, 谢志清等. 基于 MODIS 植被指数的太湖蓝藻信息提取方法研究[J]. 气象科学, 2011, 31(6): 737–741.
[12] Yang, Z., Pan, X., You, C., *et a*l. Spatio-temporal variation of fractional vegetation coverage in the Aydingkol Lake Basin [J]. *Journal of Applied Remote Sensing*, 2022, 16(1): 1–23. DOI: http://dx.doi.org/10.1117/1.JRS.16.014506.
[13] 杨子, 潘鑫, 袁洁等. 基于随机森林算法的卫星监测太湖蓝藻数据集(2019)[J/DB/OL]. 全球变化数据仓储电子杂志, 2023. https://doi.org/10.3974/geodb.2023.12.01.V1.
[14] 全球变化科学研究数据出版系统. 全球变化科学研究数据共享政策[OL]. https://doi.org/10.3974/dp.policy.2014.05 (2017 年更新).
[15] Breiman, L., Cutler, R. A., Random Forests Machine Learning [J]. *Journal of Clinical Microbiology*, 2001, 45(1): 5–32. DOI: https://doi.org/10.1023/A:1010933404324.
[16] 潘鑫, 杨子, 杨英宝等. 基于高分六号数据提取太湖蓝藻方法的对比及适用性析[J]. 湖泊科学, 2022, 34(6): 1866–1876.
[17] 尹靖, 朱煜峰. OLI 影像的不同区域水体提取方法对比研究[J]. 江西科学, 2020, 38(5): 743–747.