

基于随机森林模型的太湖水生植被遥感信息提取*

侍昊¹, 李旭文¹, 牛志春¹, 李继影², 李杨³, 李宁^{4**}

(1: 江苏省环境监测中心, 南京 210036)

(2: 苏州市环境监测中心站, 苏州 215004)

(3: 南京师范大学虚拟地理环境教育部重点实验室, 南京 210023)

(4: 南京大学生命科学院, 南京 210046)

摘要: 水生植被作为太湖湿地的重要组成部分, 其数量和范围变化影响着湖泊生态系统的平衡, 故利用遥感技术对水生植被的空间分布开展研究有助于太湖湿地生态系统的保护。以 Landsat 8 多光谱遥感影像为主要数据源, 利用光谱指数和图像变换方法构建多个特征变量, 结合随机森林(RF)模型, 提取太湖水生植被的空间分布。结果表明: (1) 通过对比分析训练样本特征值的平均值、标准差和变异系数, *NDVI*、*NDWIF*、*SR* 等指数更易于区分开敞水域和沉水植被、浮叶植被和挺水植被; (2) 当设置 1000 棵分类树和 4 个分割节点的随机变量时, RF 分类模型的袋外误分率小于 6%, 误分主要受 *SR*、*MNDVI* 和 *NDVI* 等特征变量影响; (3) 通过验证分析, 基于 RF 模型获得的 2014 年 7 月太湖水生植被覆盖面积约为 306.0 km², 分类精度为 88.56% (Kappa 系数为 0.88), 主要分布在湖体的东部和南部, 以沉水和浮叶植被为主, 两者占水生植被覆盖总面积的 84.9%。

关键词: 水生植被; 随机森林模型; 特征变量; Landsat 8; 太湖

Remote sensing information extraction of aquatic vegetation in Lake Taihu based on Random Forest Model

SHI Hao¹, LI Xuwen¹, NIU Zhichun¹, LI Jiying², LI Yang³ & LI Ning^{4**}

(1: *Jiangsu Environmental Monitoring Center, Nanjing 210036, P.R.China*)

(2: *Suzhou Environmental Monitoring Center, Suzhou 215004, P.R.China*)

(3: *Key Laboratory of VGE, Ministry of Education, Nanjing Normal University, Nanjing 210023, P.R.China*)

(4: *School of Life Science, Nanjing University, Nanjing 210046, P.R.China*)

Abstract: Aquatic vegetation is a vital component of the ecosystem of Lake Taihu. Assessment of its distribution and abundance by using remote sensing techniques is important for the protection of Lake Taihu as the information serves as an excellent indicator of aquatic environmental quality. In this research, to extract spatial distribution of the different aquatic vegetation types in Lake Taihu, we took Landsat 8 multi-spectral images as the main data source and applied them with Random Forest Model on the basis of multiple characteristic variables, which were constructed by the method of spectral index and image transformation. The results show: (a) By analyzing and comparing statistics mean, standard deviation and variable coefficient obtained from different training samples, we found *NDVI*, *NDWIF* and *SR* were better characteristic variables for distinguishing open water and floating-leaf vegetation, submerged vegetation and emergent vegetation than others. (b) Under the condition of 1000 classification trees with 4 random variables in split node, the out-of-bag error of Random Forest Model was below 6%. The error of model was mainly affected by *SR*, *MNDVI* and *NDVI*. (c) According to validation analysis, the overall accuracy classification of image based on Random Forest Model was about 88.56% with a high Kappa index of 0.88. The total area of aquatic vegetation was about 306.0 km² in July of 2014,

* 高分辨率对地观测系统重大专项 (05-Y30B02-9001-13/15-WX2)、国家水体污染控制与治理科技重大专项 (2012ZX07506003)、江苏省环境监测科研基金项目 (1315、1416)、环保公益性行业科研专项 (201309008) 和江苏省普通高校研究生科研创新计划资助项目 (KYLX15_0745) 联合资助。2015-05-24 收稿; 2015-09-09 收修改稿。侍昊 (1986~), 男, 博士; E-mail: shihao_752@163.com.

** 通信作者; E-mail: lining196@126.com.

in which floating-leaf vegetation and emergent vegetation accounted for 84.9% and were mainly distributed in the eastern and southern parts of Lake Taihu.

Keywords: Aquatic vegetation; Random Forest Model; characteristic variables; Landsat 8; Lake Taihu

水生植被具有净化水质、保护环境等生态功能,是湖泊生态系统演化和平衡的重要调控者,对陆域水体的物理和化学环境具有显著改善作用^[1-2]. 但近几十年来,随着我国部分湖泊藻型富营养化严重,造成水质急剧恶化,进而导致水生高等植物的消失. 同时由于湖泊渔业结构的调整,使得浅水草型湖泊植物群落结构发生改变,沼生植物比例不断增加,湖泊沼泽化现象也有所加剧^[3-5]. 为此,为了准确及时了解和分析水生植被的空间分布和变化趋势,保护和恢复水生植被资源和生态环境,国内外研究人员应用遥感技术开展了一系列水生植被的研究工作.

与陆地植被遥感相比,水生植被遥感信息提取易受到多种因素影响,如水体透明度对沉水植被信息提取的干扰,蓝藻水华对浮叶植被提取精度的影响^[6-8]. 因此,除了利用影像的原始波段外,研究人员还常借助光谱指数和图像变换等方法建构多个变量增强水生植被的光谱特征,然后将多组特征变量结合决策树模型进行水生植被遥感信息提取^[6,9-14]. 目前决策树方法在水生植被信息提取中,拥有训练复杂度较低、预测过程快速、模型易于表达等优点. 但随着类别和特征变量个数的增加,造成树形过于复杂,分类错误数量不断上升,后期需要对树进行必要的修剪,避免过度拟合现象发生;同时在修剪的过程中,也常会丢失一些看似无用但实际却包含重要信息的变量,限制了决策树分类精度的提高^[12]. 针对这些问题,有学者在决策树算法的基础上,提出了随机森林(Random Forest, RF)算法,该算法不仅具有数据处理量大、运行速度较快、抗噪声能力强和抑制过度拟合等优势,同时也不需要顾虑多元共线性问题,可以在不做变量选择的情况下,对特征变量重要性进行估计,有效地提高了模型的可靠性和可解释性^[15-17]. 尽管与决策树方法相比,RF 具有众多优势,但将该方法应用到水生植被遥感信息提取上的研究鲜有报道.

因此,本文以太湖地区 Landsat 8 多光谱遥感影像为主要数据源,利用光谱指数和图像变换方法建构多组特征变量,分析其对不同类型水生植被的识别能力;在此基础上,以多组特征变量为输入量结合 RF 模型提取水生植被的空间分布,并对模型的分类精度和特征变量的重要性进行分析和评价,以期为大型湖泊水生植被遥感监测与信息提取提供有效的方法补充.

1 研究区概况

太湖(30°55'40"~31°32'58"N, 119°52'32"~120°36'10"E)是我国第三大淡水湖泊,位于长江三角洲南翼坦荡的太湖平原上. 按平均水位 3 m(吴淞基面)计算,其湖泊面积为 2427.8 km²,除去湖中 51 个岛屿面积(89.7 km²),实际水域面积为 2338 km²,太湖南北长 68.5 km,东西平均宽 34 km,最宽处 56 km,湖泊平均水深 1.9 m,是一个典型的浅水湖泊. 随着太湖流域经济快速发展和湖泊资源高强度利用,太湖水生植物从 1960s 至今出现了明显的衰退,除东太湖及局部区域有少量分布外,湖区内水生高等植物几乎绝迹,“水下森林”现象逐步消失^[18]. 近年来整个湖区水生植物的多样性也出现了明显的空间差异,东太湖表现出 9 个植物类型,而西太湖面积虽为东太湖的 16.8 倍,但仅有 4 个植物类型^[3,19].

1.1 野外调查

太湖水生植被野外调查在 2013 和 2014 年的 5—9 月份展开,调查区域遍及整个湖区,调查内容包括植被类型、经纬度信息、植被覆盖度、群丛组成和优势种等各类专题信息. 根据实际调查情况和相关文献资料^[3,7,9,19],湖区共有 116 个调查点(其中 2013 年 79 个,2014 年 37 个),东部和南部沿岸等水生植被较为密集区域设置的调查点相对较多(图 1). 调查点要求植被覆盖面积一般大于 60 m×60 m(利用 GPS 结合遥感影像进行量测),同时植被覆盖度大于 50%(随机选择 3~5 处区域,在其上方 1.2 m 处拍摄照片,利用 CAN_EYE 软件提取覆盖度参数).

1.2 遥感数据

遥感数据以 Landsat 8 的多光谱影像为主,OLI 陆地成像仪包括 9 个波段,不仅拥有 ETM+传感器所有的波段,还有两个新增的波段:主要应用于海岸带观测的深蓝波段(band 1:0.433~0.453 μm)和可用于云检测

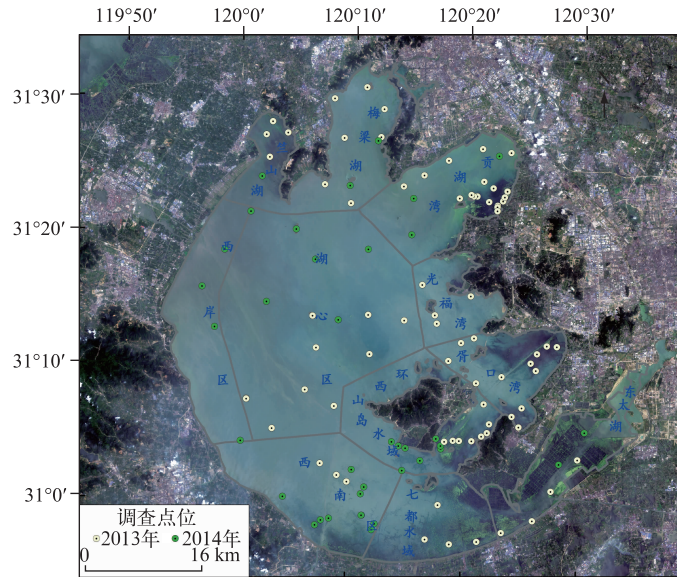


图1 太湖水生植被调查点的空间分布

Fig.1 Distribution of field survey points of aquatic vegetation in Lake Taihu

的卷云波段 (band 9: 1.360~1.390 μm)^[20]. 本文共选取 6 幅太湖地区的 Landsat 8 遥感影像 (覆盖区域云量小于 5%) 开展水生植被遥感信息提取工作, 成像时间分别为 2013 年 4 月 14 日、2013 年 7 月 19 日、2013 年 11 月 8 日、2013 年 12 月 10 日、2014 年 3 月 16 日和 2014 年 7 月 22 日, 数据下载来自美国地质勘探局 (United States Geological Survey, <http://glovis.usgs.gov/>).

2 研究方法

2.1 光谱特征变量选择

植被一般对近红外和红外波段较为敏感, 而蓝、绿和红波段在提取水体上具有较好的效果, 故多数研究主要基于上述波段构建提取水生植被的特征变量, 如归一化植被指数、归一化水体指数、均值指数和比值指数等^[6-9]. 也有研究表明主成分变换和缨帽变换能够增强水生植被光谱特征信息^[10-11], 其中主成分变换可以分离主要信息和噪音, 减小数据冗余度及波段之间的相关性; 缨帽变换中前 3 个分量可以表征土壤光谱、植被光谱和地物湿度信息, 利于植被和水体信息的分离. 因此, 本文基于 Landsat 8 影像, 构建上述常用的特征变量 (表 1), 利用均值 (Mean)、标准差 (SD) 和变异系数 ($CV = SD/Mean \times 100\%$) 对各变量识别水生植被能力进行对比分析.

2.2 随机森林模型构建

RF 模型是一个以分类回归树 (Classification and Regression Tree, CART) 作为元分类器的组合分类器算法, 它由多个决策树 ($h(x, \theta_k)$, $k = 1, 2, \dots, l$) 组成, 其中 θ_k 是相互独立且同分布的随机向量, 每个决策树对于输入给出独立的分类结果, 最终根据所有决策树的分类结果通过多数投票决定最终输出结果 (图 2)^[15-17]. 在运算过程中, RF 主要通过 Bootstrap 自助抽样法生成各不相同的训练集来构建各个元分类器; 当利用抽样方法生成子集时, 原始样本中近 37% 的袋外数据 (Out of Bag, OOB) 不会出现在新的子集中, 这些数据被用于估计 RF 的泛化误差 (泛化误差即分类器对训练集之外数据的误分率), 以此来表征模型的稳定性和准确性. RF 也可以根据 OOB 数据进行特征变量重要性的估计, 如利用所有树计算出某一特征变量的重要性度量值取平均, 获得这一特征变量的原始重要性度量值, 采用同样的方法, 可以得到其它特征变量的原始重要性度量值^[21, 22]. 因此, 本文以 NDVI 等 11 个特征变量为输入量, 结合 RF 模型进行水生植被信息提取, 并对模型精度和各变量的重要性进行评价.

表 1 遥感影像特征变量
Tab.1 The list of remote sensing image characteristic variables

类型	指标名称	简称	公式
归一化指数	Normalized Difference Vegetation Index	NDVI	$(R_{NIR} - R_{Red}) / (R_{NIR} + R_{Red})$
	Normalized Difference Water Index	NDWI	$(R_{NIR} - R_{SWIR}) / (R_{NIR} + R_{SWIR})$
	Normalized Difference Water Index of McFeeters	NDWIF	$(R_{Green} - R_{NIR}) / (R_{Green} + R_{NIR})$
	Modified Normalized Difference Water Index	MNDWI	$(R_{Green} - R_{SWIR}) / (R_{Green} + R_{SWIR})$
均值指数	Average Reflectance of Band2, Band3 and Band4	AVE234	$(R_{Blue} + R_{Green} + R_{Red}) / 3$
比值指数	Simple Ratio	SR	R_{Red} / R_{NIR}
	Simple Ratio for Water Clarity	SRWC	R_{Red} / R_{Blue}
主成分变换	Principal Component1	PC1	
缨帽变换	Tasseled Cap(Brightness)	TC(B)	
	Tasseled Cap(Greenness)	TC(G)	
	Tasseled Cap(Wetness)	TC(W)	

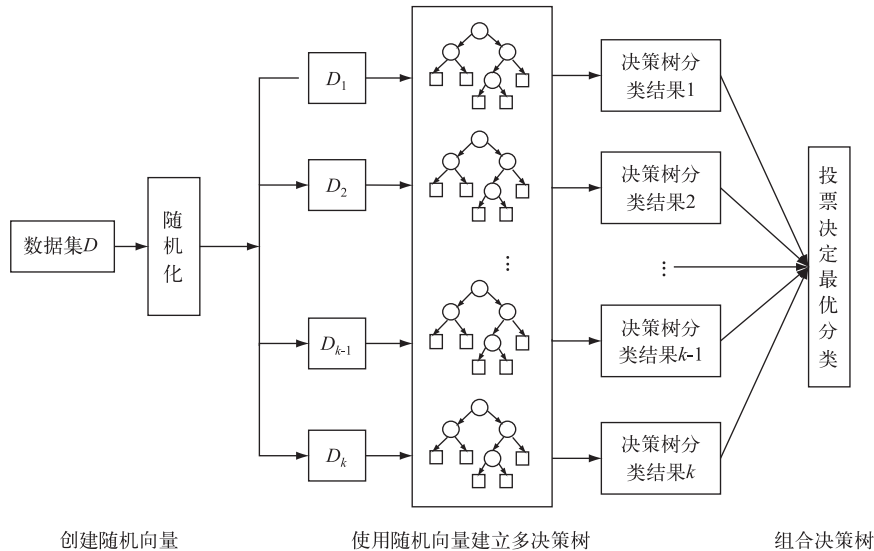


图 2 随机森林模型结构示意图

Fig.2 The structure scheme of Random Forest Model

2.3 影像训练样本的建立

根据水生植被在遥感影像上的光谱特征, 一般将水生植被划分为开敞水域(无明显水生植被聚集区域)、挺水植被、浮水植被和沉水植被 4 种类型^[2,10], 但实际调查点位中常出现不同类型植被间并无明显的边界、多种类型植被共生现象, 如浮叶-沉水植被群落(以浮叶植被为主或以沉水植被为主), 同时一般水生植被野外调查时间跨度较长, 植物生长状态及分布区域也会受到季节等各种因素影响而发生变化, 给水生植被信息提取带来一定难度^[7,9]. 为此, 本文首先利用野外调查数据结合相近时间的 6 景 Landsat 8 影像, 以调查中区域最大覆盖面积的植被类型为主, 从影像色调、颜色、大小、形状等要素上建立水生植被的遥感影像判别标志(如影像上某一区域所对应的相近时间实地调查中的浮叶植被覆盖面积最大, 则该解译标志被判别为浮叶植被); 其次参考相关研究成果^[7,9,19], 并结合 6 景影像发现, 太湖水生植被在 6 月份之前生长较为缓慢, 一般 8-9 月份植被覆盖范围最大, 并且保持相对稳定的状态, 故选择与该时间段最为相近的 2014 年 7 月 22 日影像数据; 最后利用建立的水生植被解译标志, 在待分类影像上建立开敞水域、挺水植被、浮水

植被和沉水植被 4 类训练样本,并进行水生植被信息提取.同时为了保证样本数据的可靠性,根据实际调查中水生植被分布,并利用 *NDVI* 值的平均值+三倍标准差方法^[23],对训练样本中离异值进行剔除,最终获得每类的训练样本不小于 2000 个像元(样本中 50% 像元用于构建水生植被分类模型,50% 像元用于遥感影像分类后精度评价).

遥感数据的预处理工作主要基于 ENVI(The Environment for Visualizing Images) 平台开展:首先利用 FLAASH(Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes) 模块对影像进行大气校正;然后利用 Landsat 系列卫星影像数据的 GLS2005 控制点库^[24],在每期影像上均匀选取 20~25 个参考点进行几何校正,要求校正后影像的均方根误差小于 1 个像元;最后将太湖矢量图层(WGS_1984_UTM 投影坐标系)与影像叠加裁剪出太湖水体区域.

3 结果与讨论

3.1 水生植被光谱特征变量比较

利用光谱指数和图像变换方法获得的 11 个光谱特征值的空间分布(图 3)表明,*NDVI*、*NDWIF*、*SR*、*PC1*、*TC(G)* 和 *TC(W)* 等能够较好地识别开敞水域和沉水植被,但对不同类型的水生植被分离能力差异较大.通过比较 4 种类型训练样本特征值的统计量(表 2)发现,在归一化指数中,*NDVI* 和 *NDWIF* 对不同类型的分离效果较好,*NDVI* 均值和标准差分别在 -0.27~0.69 和 0.07~0.09 之间,*NDWIF* 均值和标准差分别在 -0.56~0.39 和 0.06~0.11 之间;而 *NDWI* 和 *MNDWI* 分离透明度较低的水域和沉水植被的能力较差,两种类型的标准差变化范围明显较大.在均值和比值指数中,*SR* 在浮叶植被、挺水植被和沉水植被、开敞水域上具有较好的辨识能力;而 *AVE234*、*SRWC* 能够较好地分离开敞水域和沉水植被,但是很难区分浮叶植被和挺水植被,浮叶植被和挺水植被的 *AVE234* 均值分别为 895.42 和 843.09,*SRWC* 均值分别为 1.12 和 1.14.

从图像变换方法构建的特征变量上看,利用多光谱波段(除了卷云波段和全色波段)进行主成分变换获得第 1 分量的特征值贡献率达到了 80% 以上,对特征信息具有较好的解释能力,训练样本的 *PC1* 均值在 -2632.03~-254.28 之间,但识别浮叶植被和挺水植被能力较差,其标准差波动较大.而利用缨帽变换获得的 *TC(B)*、*TC(G)* 和 *TC(W)* 能够较好地地区分开敞水域-沉水植被和浮叶植被-挺水植被,尤其是 *TC(W)* 对这两大类有较好的区分效果,但进一步细分能力较弱,如 *TC(B)* 很难分离浮叶植被和挺水植被,以及 *TC(G)* 和 *TC(W)* 很难区分开敞水域和沉水植被.

表 2 水生植被特征变量统计参数

Tab.2 The statistical parameters of characteristic variables of aquatic vegetation types

指标	开敞水域			沉水植被			浮叶植被			挺水植被		
	均值	标准差	变异系数	均值	标准差	变异系数	均值	标准差	变异系数	均值	标准差	变异系数
<i>NDVI</i>	-0.27	0.08	-30	-0.14	0.08	-57	0.53	0.07	13	0.69	0.09	13
<i>NDWI</i>	0.21	0.09	43	0.27	0.15	56	0.34	0.04	12	0.35	0.05	14
<i>NDWIF</i>	0.39	0.08	21	0.29	0.11	38	-0.39	0.06	-15	-0.56	0.11	-20
<i>MNDWI</i>	0.55	0.88	160	0.51	0.18	35	-0.05	0.07	-140	-0.27	0.14	-52
<i>AVE234</i>	994.84	96.02	10	616.44	167.85	27	895.42	177.95	20	843.09	281.72	33
<i>SR</i>	1.76	0.23	13	1.34	0.22	16	0.31	0.06	19	0.19	0.07	37
<i>SRWC</i>	0.93	0.07	8	0.79	0.08	10	1.12	0.10	9	1.14	0.19	17
<i>PC1</i>	-948.14	189.17	-20	-254.28	334.52	-132	-1949.57	692.94	-36	-2632.03	805.58	-31
<i>TC(B)</i>	1972.90	179.82	9	1279.91	318.33	25	2434.18	514.24	21	2748.61	633.92	23
<i>TC(G)</i>	-692.75	70.45	-10	-518.26	104.37	-20	-917.59	199.93	-22	-1115.43	214.37	-19
<i>TC(W)</i>	62.55	133.70	214	11.83	121.25	1025	-2257.46	730.30	-32	-3752.56	959.30	-26

与利用 Landsat TM/ETM+ 影像构建的特征变量相比^[6-11],基于 Landsat 8 遥感数据建立的光谱指数具有较好的延续性,如 *NDVI*、*NDWIF*、*SR* 对不同类型的水生植被仍具有较强的敏感性.但在传感器新增深蓝光波段

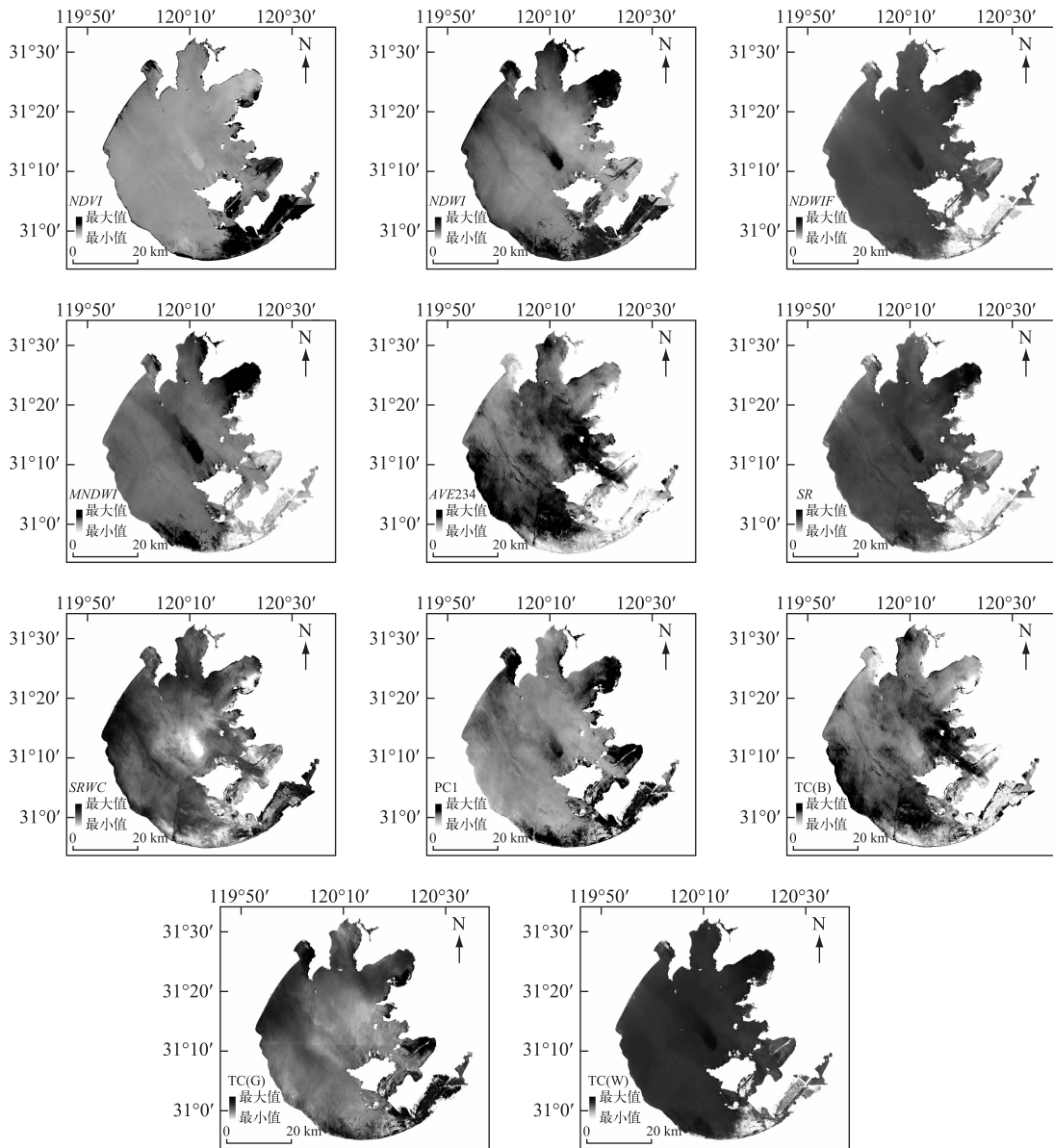


图3 基于光谱指数和影像变换方法的计算结果

Fig.3 The calculation results based on spectral index and image transformation

和卷云波段,并且对部分波段的波长范围进行了调整的情况下^[20,25],主成分变换和缨帽变换方法并没有表现出原有的优势,4种类型样本特征值的变异系数波动性较大,明显高于光谱指数法。因此,目前所采用主成分变换和缨帽变换等增强算法可能需要 Landsat 8 影像的特征进一步改进和优化。

3.2 水生植被信息提取模型评价

利用 RF 模型可以处理大数据集和多特征变量,并且能够在分类结果中给出各变量重要性估计的特点,以上述 11 个特征变量作为输入量构建水生植被信息提取模型。本文将分类树的数量(k)设置为 1000 棵随机树,采用 $m = \sqrt{mall}$ ($mall$ 表示所有变量的个数)方法^[15-16]确定分割节点随机变量的个数(m),即 m 为 4

($\sqrt{11}$). 分类模型共进行的 10 次迭运算结果(图 4)表明,随着分类树个数的增加,模型的 OOB 误分率存在一定波动,误分率在 0.42%~11.6%之间;当分类树个数达到 50 棵时,浮叶植被和挺水植被 OOB 的误分率逐步收敛,开敞水域和沉水植被的 OOB 误差几乎无明显变化;当分类树的个数达到 100 棵以上,浮叶植被和挺水植被的 OOB 误分率分别低于 6%和 3%. 从分类精度上看(1000 棵树),沉水植被的分类精度最低,OOB 误分率约为 6%,数据的错分主要在沉水植被与浮叶植被、沉水植被与开敞水域较难分离的区域,而水体的精度最高,提取精度达到 99%以上. 需要注意的是,目前 RF 模型中设置的 1000 棵分类树能够满足常规错误逐渐收敛的要求,但没有充分考虑到分割节点随机变量的个数(m)设置对分类树的相关关系和分类强度的影响:一方面减少 m 个数能够减弱分类树间的相关性,但降低树的分类强度,增加了森林的误差,另一方面增加 m 个数提高了森林的强度,但同时增大了分类树的相关关系,影响了模型的效率^[26-29].

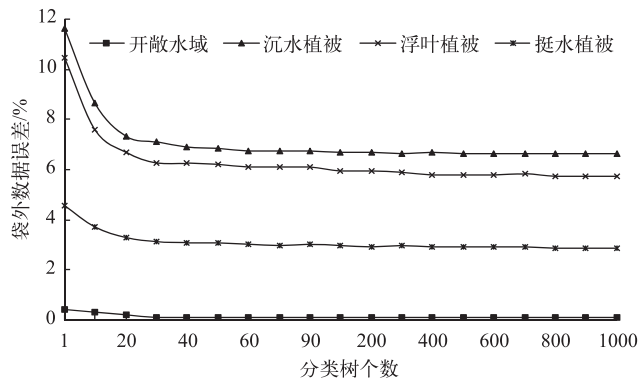


图 4 OOB 数据误分率变化(1~1000 棵)

Fig.4 The variation of OOB errors with the trees from 1 to 1000

与先采用线性判别分析等方法^[30]筛选重要特征变量,然后结合决策树构建分类模型的流程不同,RF 可以把变量的重要性扩展到所有的变量中,通过利用 OOB 误分率对特征变量的重要性变化进行判别,避免剔除重要的变量^[31]. 从特征变量重要性排序(图 5)来看, SR 、 $MNDWI$ 和 $NDVI$ 的重要性值超过了 6,对模型精度具有较高的贡献率,其中 $MNDWI$ 对开敞水域和沉水植被的分离效果较差,但在区分浮叶植被和挺水植被上仍具有一定的优势;而缨帽变换的 $TC(G)$ 、 $TC(B)$ 和主成分变换的 $PC1$ 对模型的重要性较低,其中 $PC1$ 的重要值仅为 1.38. 与单一采用图像变换,以及图像变换结合光谱指数的方法进行水生植被信息提取的研究相比^[10-12],在保持全体特征变量的情况下,图像变换方法并没有表现出信息增强的优势,尤其是主成分变换和缨帽变换($TC(G)$ 、 $TC(B)$)对模型精度的贡献明显低于光谱指数变量. 但需要注意的是,RF 重要性估计有助于对模型中各特征变量进一步了解分析,有效提高模型运算的效率和精度,但 RF 也是一个黑箱模型,无法获得分割节点随机变量的名称及分割阈值大小.

3.3 水生植被空间分布格局分析

利用 RF 模型提取 2014 年 7 月太湖开敞水域、沉水植被、浮叶植被和挺水植被的空间概率分布(概率分布图上的值越高,表明被判别为该类型的可能性越高). 从概率分布(图 6)看,太湖西部和湖区几乎无明显的水生植被集聚现象(图 6a);环西山岛水域、七都水域和东太湖浮叶植被的分布面积较大(图 6b);沉水植被主要分布在贡湖湾、胥口湾和西南区一带(图 6c);挺水植被以东太湖及沿岸两侧的浅水区域分布居多(图 6d).

综合各类型概率分布获得 2014 年 7 月太湖水生植被的空间分布,并利用检验样本对分类结果进行精度验证(表 3),结果表明:水生植被总体分类精度为 88.56% (Kappa 系数为 0.88),沉水植被的用户精度为 84%,在沉水植被和水域之间的误分较为明显,浮叶植被的制图精度为 85%,在沿岸交错带的浮叶植被常错判成挺水植被. 从分类结果统计(图 6、7)上看,提取的水生植被覆盖面积约 306.0 km²,其中沉水植被、浮叶植被和挺水植被面积分别为 132.4、127.5 和 46.1 km²,沉水植被和浮叶植被占水生植被总面积的 84.9%;

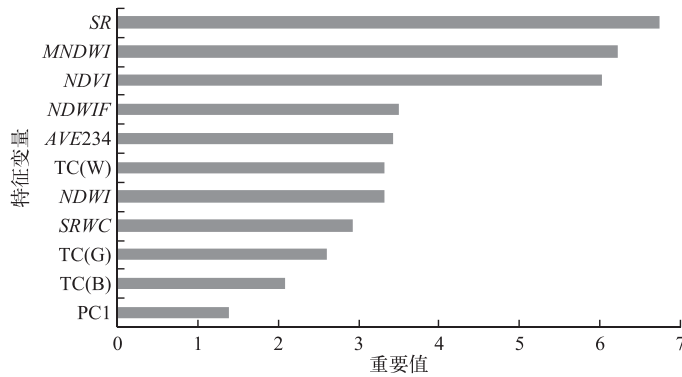


图 5 RF 模型中变量的重要性排序

Fig.5 The rank of variable in order of importance in Random Forest Model

东部和南部湖区的水生植被分布面积相对较大,主要分布在贡湖湾、广福湾、胥口湾、环西山岛水域、七都水域、西南区、东太湖等区域;不同类型的水生植被存在一定的共生现象,以浮叶-沉水植被居多.与近年来太湖水生植被空间信息提取研究成果相比,本文提取的水生植被面积总体略有下降,从2007年的364.1 km²^[9]减少到2014年的306.0 km²;与 Zhao 等^[9]2010年9月获取的面积相近,但由于太湖水生植被生长不仅受到温度、光照和生物等因素影响,而且也受到围网养殖、水体富营养化、水利工程和生态修复等人类活动共同影响^[3-5,19],沉水植被的空间分布上存在一定空间差异,主要表现为环西山岛水域的沉水植被面积略有减少,而太湖南部的沉水植被面积略有扩大,并呈现出集聚状分布趋势.这一空间分布特征也验证了野外调查的结果,2014年夏季太湖南部水域出现大面积沉水植被,以马来眼子菜(*Potamogeton malaianus*)为主,物种结构较为单一;随着优势种生物量急剧增加,分布范围逐渐扩大,由东向西、由沿岸带向湖中心蔓延,影响了部分航道的运行,但目前对湖泊生态系统是否产生影响,还需要结合多年的调查资料进一步分析和讨论.

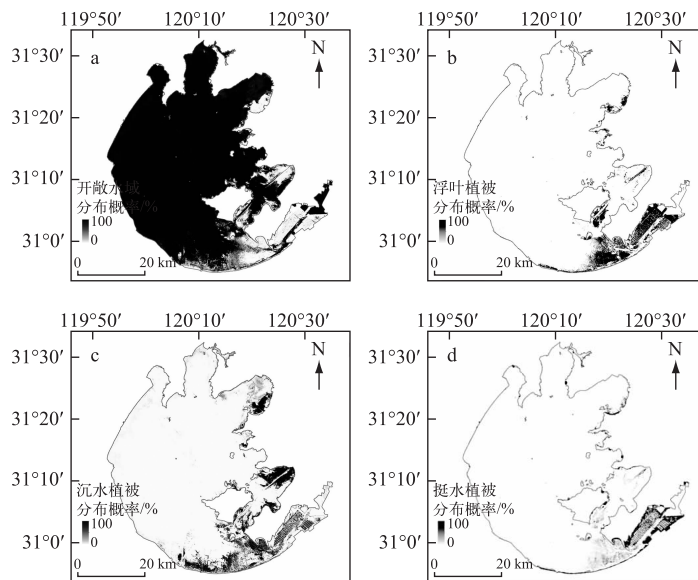


图 6 2014 年 7 月太湖水生植被的概率分布

Fig.6 The probability of aquatic vegetation types in Lake Taihu in July of 2014

表 3 基于 RF 分类精度评价*
Tab.3 Accuracy assessment on RF-based classification

类型	开敞水域	沉水植被	浮叶植被	挺水植被	合计	用户精度
开敞水域	1001	36	0	0	1037	97%
沉水植被	133	1000	54	0	1187	84%
浮叶植被	0	13	1012	151	1176	86%
挺水植被	4	3	125	1008	1140	88%
合计	1138	1052	1191	1159	4021	—
制图精度	88%	95%	85%	87%	—	—

* 总体精度 = 88.56%, Kappa 系数 = 0.88.

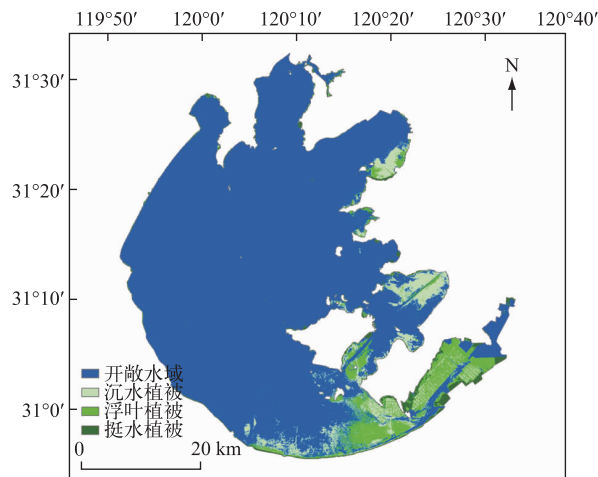


图 7 2014 年 7 月太湖水生植被的空间分布

Fig.7 The distribution of aquatic vegetation types in Lake Taihu in July of 2014

4 结论

本文以 Landsat 8 多光谱遥感影像为主要数据源,将光谱指数和图像变换方法构建的多特征变量作为输入量,利用 RF 模型提取了太湖水生植被空间分布. 结果表明该方法较好地提取了 2014 年 7 月太湖水生植被空间信息,同时清晰地判别了多特征变量的重要性. 在计算过程中,该方法实现简单,具有数据处理量大、运行速度较快、抗噪音能力强和抑制过度拟合等优点,有效地提高了分类结果的可靠性和可解释性,可为水生植被遥感监测体系提供方法补充. 但从目前研究中也发现,在利用 Landsat 8 影像新特征改进探测水生植被信息的特征变量、优化 RF 模型中分类树和分割节点随机变量个数的设置以及利用长时间序列遥感数据对水生植被的空间分布及变化趋势开展深入分析等方面还需要进一步研究探讨.

5 参考文献

- [1] 杨清心. 东太湖水生植被的生态功能及调节机制. 湖泊科学, 1998, **10**(1): 67-72. DOI 10.18307/1998.0111.
- [2] 李俊生, 吴迪, 吴远峰等. 基于实测光谱数据的太湖水华和高等水生植物识别. 湖泊科学, 2009, **21**(2): 215-222. DOI 10.18307/2009.0209.
- [3] 谷孝鸿, 张圣照, 白秀玲等. 东太湖水生植物群落结构的演变及其沼泽化. 生态学报, 2008, **25**(7): 1541-1548.
- [4] Cai YJ, Gong ZJ, Qin BQ. Benthic macroinvertebrate community structure in Lake Taihu, China: Effects of trophic status, wind-induced disturbance and habitat complexity. *Journal of Great Lakes Research*, 2012, **38**: 39-48.

- [5] Dong BL, Qin BQ, Gao G *et al.* Submerged macrophyte communities and the controlling factors in large, shallow Lake Taihu (China): Sediment distribution and water depth. *Journal of Great Lakes Research*, 2014, **40**:646-655.
- [6] 张寿选, 段洪涛, 谷孝鸿. 基于水体透明度反演的太湖水生植被遥感信息提取. *湖泊科学*, 2008, **20**(2): 184-190. DOI 10.18307/2008.0208.
- [7] Ma RH, Duan HT, Gu XH *et al.* Detecting aquatic vegetation changes in Taihu Lake, China using multi-temporal satellite imagery. *Sensors*, 2008, **8**: 3988-4005.
- [8] Oyama Y, Matsushita B, Fukushima T. Distinguishing surface cyanobacterial blooms and aquatic macrophytes using Landsat/TM and ETM+ shortwave infrared bands. *Remote Sensing of Environment*, 2015, **157**: 35-47.
- [9] Zhao DH, Jiang H, Yang TW *et al.* Remote sensing of aquatic vegetation distribution in Taihu Lake using an improved classification tree with modified thresholds. *Journal of Environmental Management*, 2012, **95**(1): 98-107.
- [10] 林 川, 宫兆宁, 赵文吉. 基于中分辨率 TM 数据的湿地水生植被提取. *生态学报*, 2010, **30**(23): 6460-6469.
- [11] 贾永红, 李芳芳. 一种新的湿地信息遥感提取方法研究. *华中师范大学学报: 自然科学版*, 2007, **41**(4): 641-644.
- [12] 那晓东, 张树清, 李晓峰等. 基于 QUEST 决策树兼容多源数据的淡水沼泽湿地信息提取. *生态学杂志*, 2009, **28**(2): 357-365.
- [13] 韦 玮, 崔丽娟, 李胜男等. 基于偏差平均值的乌梁素海湿地变化监测研究. *林业科学研究*, 2012, **25**(6): 719-725.
- [14] 徐德兰, 张东东, 张翠英. 3 个时期骆马湖大型水生植物的分布及变化. *湿地科学*, 2013, **11**(3): 320-325.
- [15] Breiman L. Bagging predictors. *Machine Learning*, 1996, **24**(2): 123-140.
- [16] Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5-32.
- [17] 李欣海. 随机森林模型在分类与回归分析中的应用. *应用昆虫学报*, 2013, **50**(4): 1190-1197.
- [18] 雷泽湘. 太湖大型水生植被及其环境效应研究[学位论文]. 广州: 暨南大学, 2006.
- [19] 刘伟龙, 胡维平, 陈永根等. 西太湖水生植物时空变化. *生态学报*, 2007, **27**(1): 159-170.
- [20] 徐涵秋, 唐 菲. 新一代 Landsat 系列卫星: Landsat 8 遥感影像新增特征及其生态环境意义. *生态学报*, 2013, **33**(11): 3249-3257.
- [21] 雷 震. 随机森林及其在遥感影像处理中应用研究[学位论文]. 上海: 上海交通大学, 2012.
- [22] 张 雷, 王琳琳, 张旭东等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例. *生态学报*, 2014, **34**(3): 650-659.
- [23] Timm BC, McGarigal K. Fine-scale remotely-sensed cover mapping of coastal dune and salt marsh ecosystems at Cape Cod National Seashore using Random Forests. *Remote Sensing of Environment*, 2012, **127**: 106-117.
- [24] 张 浩, 葛小青, 冯旭祥. Landsat 8 影像像元地理坐标计算. *遥感信息*, 2013, **28**(5): 52-58.
- [25] 初庆伟, 张洪群, 吴业炜等. Landsat-8 卫星数据应用探讨. *遥感信息*, 2013, **28**(4): 110-114.
- [26] Rodriguez-Galiano VF, Ghimire B, Rogan J *et al.* An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2011, **67**: 93-104.
- [27] Rodriguez-Galiano VF, Chica-Olmo M, Abarca-Hernandez F *et al.* Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, 2012, **121**: 93-107.
- [28] Beijma SV, Comber A, Lamb A. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. *Remote Sensing of Environment*, 2014, **149**: 118-129.
- [29] Magdon P, Fischer C, Fuchs H *et al.* Translating criteria of international forest definitions into remote sensing image analysis. *Remote Sensing of Environment*, 2014, **149**: 252-262.
- [30] Pu RL, Landry S. A comparative analysis of high spatial resolution IKONOS and World View-2 imagery for mapping urban tree species. *Remote Sensing of Environment*, 2012, **124**: 516-533.
- [31] 刘海娟, 张 婷, 侍 昊等. 基于 RF 模型的高分辨率遥感影像分类评价. *南京林业大学学报: 自然科学版*, 2015, **39**(1): 99-103.