

基于群组特征与随机森林算法的大庆地区湖泊健康评价

邢贞相^{1,2,3}, 侯泉滢¹, 王轶男¹, 姜 皓⁴, 李瀚宇¹, 关 越¹, 张 涵^{1,2}

(1. 东北农业大学水利与土木工程学院; 2. 黑龙江省寒区水资源与水利工程重点实验室;
3. 松花江流域生态环境保护研究中心; 4. 清华大学未央书院)

摘要:为探究黑龙江省大庆地区湖泊健康状况及其主导影响因子的差异,采用 *K*-means 机器学习分类算法,依据大庆地区 60 个典型湖泊的属性特征对湖泊进行群组划分,并探究其群组内共性典型特征及其组间差异;构建评价指标体系并利用随机森林算法的高维非线性映射能力,结合湖泊群组划分结果,分别构建群组湖泊健康评价模型对各群组内的湖泊健康进行评价,并识别各群组内湖泊健康的主导影响因子。结果表明:研究区典型湖泊按属性特征可分为 5 个群组,分组结果可用于提升随机森林模型健康评价精度,其群组典型特征分别为高浮游植物多样性指数、低植被覆盖率、高水体自净能力、高鱼类保有指数、优水质类别;基于随机森林算法的湖泊健康评价模型能够准确地评价大庆地区典型湖泊的健康状况,且能避免传统评价方法主观性较强的不足;构建的评价指标体系能够更全面、综合地考虑湖泊的健康影响因子,新增的湖泊面积萎缩比例(0.0544)、天然湿地保留率(0.0705)、植被覆盖率(0.0489)等评价指标的重要度均值均高于所有评价指标的重要度均值(0.0445),进一步完善了湖泊健康评价的指标体系。

关键词:湖泊健康评价;聚类分析;群组典型特征;随机森林算法;大庆地区

Health assessment of lakes in Daqing region based on cluster characteristics and random forest algorithm//Xing Zhenxiang^{1,2,3}, Hou Quanying¹, Wang Yinan¹, Jiang Hao⁴, Li Hanyu¹, Guan Yue¹, Zhang Han^{1,2}(1. College of Water Resources and Civil Engineering, Northeast Agricultural University; 2. Heilongjiang Provincial Key Laboratory of Water Resources and Hydraulic Engineering in Cold Regions; 3. Songhua River Basin Ecological Environment Protection Research Center; 4. Weiyang College, Tsinghua University)

Abstract: To investigate the health status of lakes in the Daqing region of Heilongjiang Province and the differences in their dominant influencing factors, the *K*-means clustering machine learning classification algorithm was employed to group 60 typical lakes in the Daqing region based on their attribute characteristics. This approach explored the common typical features within each group and the differences between groups. An evaluation indicator system was established. Leveraging the high-dimensional nonlinear mapping capability of the random forest algorithm, lake health evaluation models were constructed for each cluster based on the grouping results. These models assessed lake health within each cluster and identified the dominant influencing factors for lake health within each group. The research findings indicate: The typical lakes in the study area can be classified into five clusters based on their attribute characteristics. This grouping can be used to enhance the accuracy of the random forest model for health assessment. The typical features of each cluster are: high phytoplankton diversity index, low vegetation coverage, high water self-purification capacity, high fish stock index, and excellent water quality category. The lake health assessment model based on the random forest algorithm can accurately evaluate the health status of typical lakes in the Daqing region while avoiding the shortcomings of traditional evaluation methods, which are prone to significant subjectivity. The constructed evaluation indicator system comprehensively considers the factors influencing lake health. The newly added evaluation indicators, such as lake area shrinkage ratio (0.0544), natural wetland retention rate (0.0705), and vegetation coverage rate (0.0489), exhibit higher importance values than the average importance value (0.0445) of all evaluation indicators. This further refines the indicator system for lake health evaluation.

Key words: lake health assessment; cluster analysis; cluster typical characteristics; random forest algorithm; Daqing region

基金项目:国家自然科学基金项目(51979038)
作者简介:邢贞相(1976—),男,教授,博士,主要从事水文分析与计算及河湖健康评价研究。E-mail:zxing@neau.edu.cn
通信作者:张涵(1992—),男,讲师,博士,主要从事河湖健康评价和水资源优化调度研究。E-mail:hanzhang@neau.edu.cn

人类能够直接利用的淡水资源极为有限^[1],其中约 90% 的液态地表淡水资源集中分布于湖泊之中^[2]。湖泊生态系统在区域水文循环、气候调节以及生物多样性维持等方面发挥着不可替代的重要作用^[3]。然而,近年来随着我国经济的快速发展,水资源开发利用强度显著增加,湖泊水生态系统退化等问题日益严峻^[4]。系统探究湖泊间的属性特征差异,构建科学、系统的生态健康评价指标体系并对湖泊健康进行准确评估,已成为湖泊生态环境保护与治理的关键环节。

当前湖泊健康评价研究普遍选择综合指标法^[5-6],差异主要体现在各指标权重赋值的不同。这类方法虽然能够在一定程度上评估湖泊健康状况,但是对不同湖泊环境的异质性考虑不足,且在评估结果上多依赖于指标赋权^[7],难以表征指标与湖泊健康的非线性关系^[8],从而影响湖泊健康评价结果的准确性。聚类分析可将同一地区众多湖泊划分成若干组内相似度高而组间相似度较低的群组,针对不同群组湖泊采用不同的评价标准可以充分考虑相同地区不同湖泊间的属性差异,以及湖泊环境的异质性对健康评价的影响。 K 均值聚类 (K -means) 作为机器学习分类中最常用的方法,在计算效率方面有着巨大的优势^[9]。随机森林模型能够帮助建模者识别模型因变量的主导因素^[10],揭示主导因素与因变量之间的关系^[11-12],且能够解决指标赋权时主观性较强的问题^[13]。

本文以大庆市 60 个典型湖泊作为研究对象,在对湖泊进行群组划分的基础上,从水文水资源、物理结构、水质、水生生物和社会服务功能 5 个维度,构建湖泊健康评价指标体系并构建了基于 RF 的湖泊

健康评价模型同时将其评价结果与黑龙江省发布的 DB23/T 3788—2024《河湖健康评价技术规范》(以下简称“《规范》”)中推荐方法的评价结果进行对比,探究湖泊群组间的差异特性,检验所构建模型的适用性并实现基于湖泊群组的湖泊健康评估,识别各湖泊群组健康主要影响因子。研究成果旨在为实现湖泊的分类治理和幸福河湖建设提供决策依据。

1 研究区概况与数据来源

1.1 研究区概况

大庆市位于黑龙江省西部嫩江下游与松花江上游交汇处(图 1),是黑龙江省重要的经济、文化与交通中心之一,同时也是全国著名的石油生产基地^[14]。大庆地区地势相对较低的区域往往有因季节性积水而形成的洼地和沼泽,大量湖泊、泡沼分布其中,大庆素有“百湖之城”之称^[15]。由于大庆境内无天然河流且缺少外流河,加之地势起伏整体较小,导致大量湖泊和水泡排水不畅通,湖泊水体自净能力较弱,富营养化现象时有发生。

1.2 数据来源

研究数据主要包括统计数据、现场采样与实验室样本检测数据。统计数据多数来源于《大庆统计年鉴》,其中降水量数据来源于中国气象数据网(<https://data.cma.cn>),土地利用情况及湿地分布来源于中国遥感应应用中心(<http://www.crsa.cn>)。现场采样与实验室样本检测数据根据《河湖健康评估技术导则》中的要求进行水质样本、浮游植物检测样本与底泥的采集,并在实验室进行水质检测,浮游植物样本、大型底栖无脊椎动物(底泥)样本均在

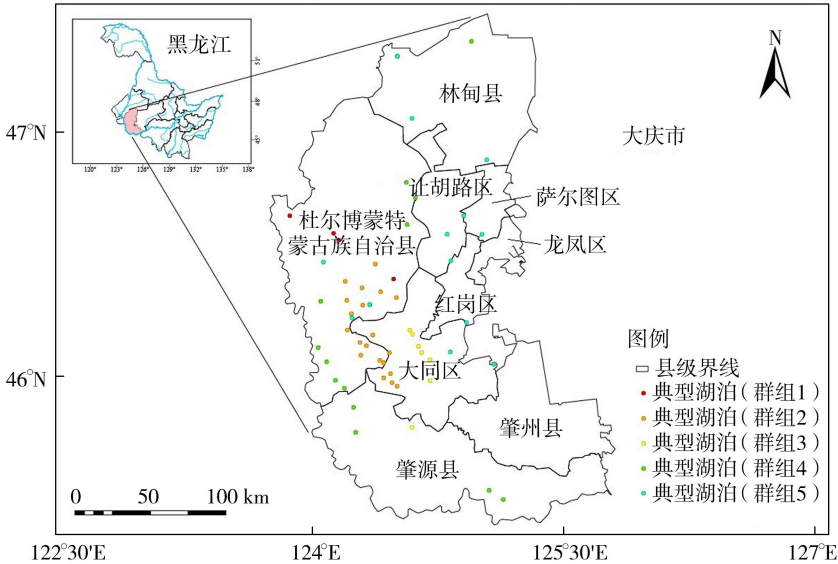


图 1 研究区位置与典型湖泊分布

Fig. 1 Location of study area and representative lakes

实验室显微镜下检测,透明度、水温均为现场检测数据;岸带状况、鱼类数据来源于地方资料和现场调查;公众满意度数据来源于现场调查问卷的统计。

2 研究方法

2.1 K-means 算法

本文采用 *K-means* 算法对研究区 60 个湖泊进行分类(即湖泊的群组划分),为后续分类构建相应的健康评价模型提供依据。

聚类分析是一种数据归约技术,旨在揭示某个数据集中观测值的子集,其中最常用的方法为 *K-means* 算法^[9]。选取划分聚类中的 *K-means* 算法对大庆地区的 60 个典型湖泊进行聚类分析。*K-means* 算法大致可以分为 4 步,步骤如下:

第 1 步 标准化数据,决定聚类的个数 *K*,即 *K* 个位置随机的初始聚类中心点。基于本文研究对象的湖泊健康各项指标的数据特点、本文聚类分析目标和计算成本,选择肘部法^[16]、DBI (Davies-Bouldin index) 法^[17] 以及 CHI (Calinski-Harabasz index) 法^[18] 相结合来确定研究样本的最优聚类类别数。

第 2 步 将每个观测点关联到其最近的聚类中心点上得到 *K* 个簇。

第 3 步 对于每一个簇找到其所有关联点的聚类中心点,将其作为新的聚类中心。

第 4 步 重复第二步和第三步,直到全部的观测点不再重新被分配,不再产生新的聚类结果。

2.2 随机森林模型

以目标湖泊各项健康评价指标为自变量,利用随机森林模型对大庆地区典型湖泊健康状况进行评价,并识别各群组内湖泊健康的主导因子。随机森林算法是一种基于决策树的集成学习算法,广泛应用于分类与回归任务^[19-20]。随机森林模型的优点在于能够有效应对随机变量的干扰,通过随机抽取的样本或特征,显著缓解高维数据中的多重共线性问题。此外,随机森林模型便于分析输入变量与输出变量之间的关联,从而有助于阐释各因子对湖泊健康的驱动机制^[21-22]。

随机森林模型运行步骤:①对自变量进行相关性分析,以保证输入的自变量间不存在显著多重共线性;②输入包含自变量即健康评价指标和分类变量即健康等级的结构化数据集;③随机森林模型超参数优化,确定决策树棵树与决策树最大深度。模型构建完成后还需进行模型准确率和有效性检验。随机森林算法流程如图 2 所示。

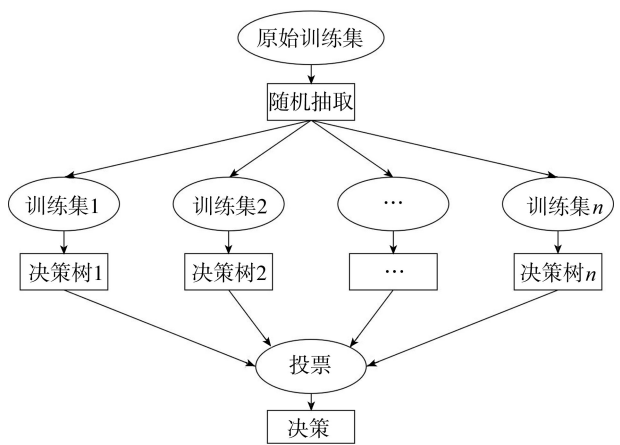


图 2 随机森林算法流程

Fig. 2 Random forest algorithm flowchart

3 结果与分析

3.1 湖泊群组划分

选取湖泊的经度、纬度、流域面积、历史水面面积、水面面积、正常蓄水位、蓄水量、湖泊岸线长度、管理范围面积、多年平均降水量、多年平均蒸发量、年径流深、封冻天数、最大冰厚、生物要求最小平均水深、湖底平均高程、生态水位限值、湖岸带平均植被覆盖率、湖岸带最大植被覆盖率、pH 值、高锰酸盐指数、化学需氧量、五日生化需氧量、总磷、总氮、叶绿素 α 、溶解氧、粪大肠菌群、透明度、浮游植物种属、浮游植物多样性指数、历史鱼类种数、鱼类种数、公众满意度、湖泊成因、泄水出路、取水口、天然湿地、防洪工程和入湖排污口数共 40 个湖泊属性指标,利用 *K-means* 算法对研究区 60 个湖泊进行群组划分。

基于 *K-means* 算法对湖泊群组进行划分的关键环节就是确定最优聚类类别数 *K*。肘部法评价 *K* 值好坏的标准是各观测点到其相应中心点的距离平方和,样本划分随着聚类数的增大更加精细、每个簇的聚合程度逐渐提高、距离平方和自然变小,当 *K* 值小于最佳聚类数时,*K* 的增大会大幅增加每个簇的聚合程度,故距离平方和的下降幅度会很大,而当 *K* 到达最佳聚类数后,距离平方和的下降幅度会骤减,这个转折点对应的 *K* 值就是样本数据的最佳聚类数;DBI 的核心思想是计算每个类和与其最相似的类之间相似程度的均值,该指标越小表明聚类效果越好;而 CHI 由类间分离度与类内紧密度的比值得到,该指标越大表明聚类效果越好。图 3 给出了不同聚类个数对应的指标曲线。

图 3(a)表明距离平方和随聚类数 *K* 的增加而减少,当 $K \geq 5$ 时,距离平方和趋于稳定,因此,确定 *K* 的取值应大于等于 5。为进一步分析聚类个数的合理性,分别计算了不同聚类数对应的 DBI 和 CHI,

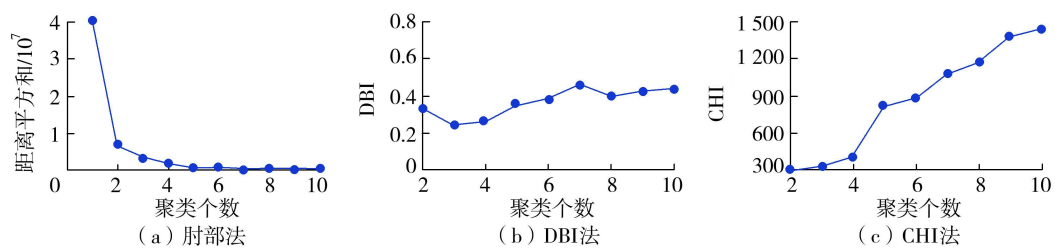


图3 基于不同方法的湖泊聚类数指标曲线

Fig. 3 Curves of lake clustering number indicators based on different methods

并分别绘制 DBI 和 CHI 不同聚类数对应指标曲线图(图 3(b)、图 3(c)),CHI 曲线(图 3(c))整体上呈上升趋势,在 $K < 4$ 时增速缓慢,位于 4 和 5 之间时增速较大,之后缓慢上升,再结合图 3(b)中不同聚类数对应的 DBI 数值,最终选定 5 作为聚类类别数。基于以上分析,大庆地区 60 个目标湖泊划分为 5 个群组,空间分布如图 1 所示。通过统计各群组湖泊的各项健康评价指标,确定群组 1~5 的湖泊典型特征分别为高浮游植物多样性指数、低植被覆盖率、高水体自净能力、高鱼类保有指数和优水质类别,该结果将用于后续各群组湖泊样本抽取过程。从图 1 可以看出各个群组内湖泊分布较为集中,群组间边界较为明显,聚类结果显著;群组 1 湖泊分布在杜尔伯特蒙古族自治县中部;群组 2 湖泊主要分布在大同区西部和杜尔伯特蒙古族自治县中东部;群组 3 湖泊主要集中在大庆市区南部;群组 4 湖泊分布较为分散,遍布林甸县、杜尔伯特蒙古族自治县和肇源县境内;群组 5 湖泊分布同样较为分散,主要分布在大庆地区中部及北部。

3.2 结构化数据集构建

训练和验证随机森林模型都需要输入包含自变量及其对应的健康等级变量在内的结构化数据集,本文分别构建用于训练不同随机森林模型的 6 类模拟湖泊结构化数据集(1 类是基于《规范》中的评价指标体系随机模拟的,其余 5 类是基于本文构建的评价指标体系随机模拟的)。

a. 构建湖泊健康评价指标体系。为了使评价结果能够更全面考虑目标湖泊各项特征现状,从而更好地体现出其健康状况,本文以《规范》的评价指标体系为基础,在采用该规范中同样的准则层和全部指标的同时,在参考相关文献^[23-28]的基础上对其部分准则层中的指标进行补充,得到本文构建的湖泊健康评价指标体系,如表 1 所示。

b. 确定评价指标分级标准。为保证健康评价结果的规范化和可比性,确定各湖泊健康评价指标的分级标准,对《规范》中曾出现的指标沿用其分级标准,对本文补充的各项指标通过查阅相关文献^[23-28]对其评价标准进行划分,并结合大庆地区最

表 1 评价指标体系

Table 1 Evaluation indicator system

目标层	准则层	指标符号	指标层
湖泊健康状况	水文水资源	H_1	生态水位满足程度*
		H_2	湖泊连通指数*
		H_3	年平均降水量
	物理结构	P_1	湖泊面积萎缩比例
		P_2	岸带状况*
		P_3	天然湿地保留率
		P_4	堤防工程防洪达标率*
		P_5	植被覆盖率
	水质	W_1	水质优劣程度*
		W_2	营养状态*
		W_3	水体自净能力
		W_4	湖泊饮用水水源地水质达标率
	水生生物	A_1	浮游植物多样性指数*
		A_2	大型底栖无脊椎动物生物完整性指数*
		A_3	鱼类保有指数*
	社会服务功能	S_1	人口密度
		S_2	GDP 年增长率
		S_3	洪灾频率
		S_4	洪灾强度
		S_5	供水水量保证程度
		S_6	公众满意度*
		S_7	公众环保意识
		S_8	政策法规贯彻力度

注: * 表示黑龙江省《规范》中已有指标,其他指标为本文补充的评价指标。下表同。

近 10 年各项指标数值确定分级标准上、下限值,见表 2。

c. 构建模拟湖泊结构化数据集。根据不同湖泊群组的特征和各评价指标的分级标准,模拟生成一个结构化数据集,作为一个模拟湖泊。该数据集包含了各项湖泊健康评价指标的数值,以及这一湖泊的健康等级。多个模拟湖泊的集合将用于后续随机森林模型的训练及检验环节。

构建以《规范》评价指标体系为基础的模拟湖泊结构化数据集:以《规范》中的评价指标和健康分级标准为基础,在各项指标分级均为非常健康的范围内分别随机抽取 5 000 个湖泊健康评价指标的模拟数值作为随机森林模型训练所需的自变量,并将其对应的健康等级变量设定为非常健康,即随机模拟 5 000 个健康等级为非常健康对应的所有评价指

表 2 湖泊健康评价指标分级标准
Table 2 Lake health assessment indicator grading criteria

健康等级	$H_1/\%$	H_2	H_3/mm	$P_1/\%$	P_2	$P_3/\%$	$P_4/\%$	$P_5/\%$	W_1	W_2	$W_3/(\text{mg/L})$	$W_4/\%$
非常健康	80~100	9~10	1 000~<1 200	0~<5	0~<300	90~100	90~100	75~100	1~<2	0~<20	7.5~<9	95~100
健康	60~<80	7~<9	800~<1 000	5~<10	300~<500	66~<90	70~<90	50~<75	2~<3	20~<36	6~<7.5	85~<95
亚健康	40~<60	5~<7	600~<800	10~<20	500~<800	40~<66	50~<70	30~<50	3~<4	36~<52	5~<6	70~<85
不健康	20~<40	3~<5	400~<600	20~<30	800~<1 000	16~<40	30~<50	20~<30	4~<5	52~<70	3~<5	50~<70
病态	0~<20	0~<3	0~<400	30~100	1 000~<1 300	0~<16	0~<30	0~<20	5~<6	70~100	0~<3	0~<50

健康等级	A_1	A_2	$A_3/\%$	$S_1/(\text{人}/\text{km}^2)$	$S_2/\%$	$S_3/\text{次}$	$S_4/\text{万元}$	$S_5/\%$	S_6	$S_7/\%$	S_8
非常健康	0~<0.4	1.62~<2	80~100	0~<100	-2~<1	0~<0.05	0~<3	95~100	9~10	90~100	9~10
健康	0.4~<1	1.03~<1.62	60~<80	100~<200	1~<3	0.05~<0.1	3~<5	85~<95	7~<9	80~<90	7~<9
亚健康	1~<2	0.31~<1.03	40~<60	200~<300	3~<5	0.1~<0.2	5~<7	75~<85	5~<7	70~<80	5~<7
不健康	2~<2.6	0.1~<0.31	20~<40	300~<400	5~<6	0.2~<0.3	7~<10	50~<75	3~<5	60~<70	3~<5
病态	2.6~<3	0~<0.1	0~<20	400~<500	6~<8	0.3~<0.4	10~<12	0~<50	0~<3	0~<60	0~<3

标数据,它们对应的结构化数据集作为随机森林模型的输入数据。以此类推,分别抽取其余四种健康等级的模拟湖泊结构化数据集,组成结构化数据集用于训练以《规范》中评价指标体系为基础的基于随机森林算法的湖泊健康评价模型,这一模型将用于进行随机森林模型在大庆地区湖泊健康领域的适用性检验。其余 5 类模拟湖泊结构化数据集分别用于训练以更全面的评价指标体系为基础的基于随机森林算法的大庆地区不同群组湖泊健康评价模型。

构建以本文构建的评价指标体系为基础的模拟湖泊结构化数据集:以群组 1 湖泊中评价等级为非常健康中的湖泊群的模拟指标数据集为例,由于该组湖泊典型特征为高浮游植物多样性指数,故浮游植物多样性指数的模拟值是从其该健康等级对应的上限值和下限值的均值右侧范围内随机抽取,以体现该组湖泊高浮游植物多样性指数的属性特征,其余指标均从其非常健康对应的上、下限值之间的范围内随机抽取。以此类推,穷举评价等级(非常健康、健康、亚健康、不健康、病态)和湖泊所属群组(群组 1、群组 2、群组 3、群组 4、群组 5)的 25 种组合,分别采用上述方式抽取模拟湖泊结构化数据集各 5 000 个,共得到 125 000 个模拟湖泊结构化样本数据,形成用于训练不同群组湖泊随机森林健康评价模型的 5 类模拟湖泊结构化数据集。

3.3 基于随机森林算法的湖泊健康评价模型构建

随机森林模型的构建主要分为 3 步:第一步自变量相关性分析,即计算湖泊各健康指标间的皮尔逊相关系数矩阵,并绘制相应热力图,相关性分析结果如图 4 所示(图中更深的颜色表示更强的相关性)。由于不同自变量之间的皮尔逊相关系数绝对

值均未超过 0.7(图 4),未达到显著多重共线性的关注点阈值^[29],表明该结构化数据集满足随机森林模型对自变量独立性的基本要求,可用于后续研究。第二步输入结构化数据集,即将模拟湖泊健康指标数据作为输入自变量,将其对应的健康评价等级作为输出变量,用于随机森林模型的训练和验证。本文构建的 6 类结构化数据集中随机抽取 80%的模拟数据作为模型的训练数据集,其余 20%的数据作为验证数据集。第三步模型超参数优化,通过网格搜索交叉验证优化模型参数,穷举决策树棵数(200、300、400、500、600)与决策树最大深度(5、10、15、20、None)的各种组合,以模型总体准确率、召回率、精确率及 F1 分数为评估标准进行交叉验证^[30],结果表明对于 6 个模型均为决策树棵数 600 与不限制最大深度这一组合的随机森林模型性能最优。

3.4 健康评价模型的性能分析

将 6 类模拟数据集中剩余的 20%样本用于已构建随机森林模型的验证,以检验对应模型的测度准确率和有效性,采用总体准确率、召回率、精确率、和 F1 分数来评估各健康评价模型的性能^[31],结果如表 3 所示,6 个随机森林模型各项指标数值均高于 97%且彼此间十分接近,这一现象的根本原因是在输入训练数据集时,各健康等级的湖泊数量是相同的,准确率本身就比较可靠。准确率这一指标表明本次构建的 6 个基于随机森林算法的湖泊健康评价模型性能可靠,评价结果较为准确;召回率、精确率及其综合而成的 F1 分数则表明在每个细分维度(每个类别的识别能力)上也表现出色,能够用于大庆地区湖泊的健康评价,并能够较好地考虑不同湖泊群组的典型特征。

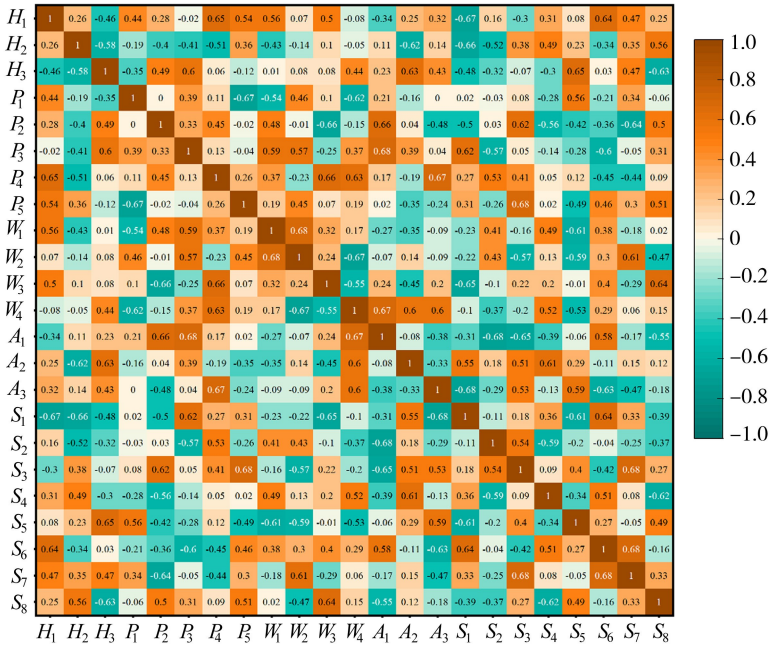


图 4 自变量相关系数矩阵热力图

Fig. 4 Heatmap of independent variable correlation coefficient matrix

表 3 模型评估结果

Table 3 Model evaluation results

模型类别	数据集	准确率	召回率	精确率	F1 分数
适用性检验模型	训练集	1	1	1	1
	验证集	0.9710	0.9710	0.9711	0.9711
群组 1 模型	训练集	1	1	1	1
	验证集	0.9750	0.9750	0.9750	0.9750
群组 2 模型	训练集	1	1	1	1
	验证集	0.9722	0.9723	0.9722	0.9722
群组 3 模型	训练集	1	1	1	1
	验证集	0.9746	0.9746	0.9746	0.9746
群组 4 模型	训练集	1	1	1	1
	验证集	0.9750	0.9750	0.9751	0.9750
群组 5 模型	训练集	1	1	1	1
	验证集	0.9734	0.9734	0.9735	0.9734

表 4 湖泊健康评价指标重要度

Table 4 Importance of lake health evaluation indicators

指标符号	《规范》	适用性检验模型	群组随机森林模型
H_1	0.1500	0.1419	0.0710
H_2	0.1000	0.0552	0.0273
H_3	0	0	0.0380
P_1	0	0	0.0544
P_2	0.1000	0.1196	0.0598
P_3	0	0	0.0705
P_4	0.1500	0.1406	0.0108
P_5	0	0	0.0489
W_1	0.0750	0.0972	0.0489
W_2	0.0750	0.1183	0.0590
W_3	0	0	0.0380
W_4	0	0	0.0490
A_1	0.0670	0.0985	0.0495
A_2	0.0670	0.0539	0.0272
A_3	0.0670	0.0762	0.0379
S_1	0	0	0.0271
S_2	0	0	0.0494
S_3	0	0	0.0109
S_4	0	0	0.0221
S_5	0	0	0.0491
S_6	0.1500	0.0986	0.0540
S_7	0	0	0.0489
S_8	0	0	0.0483

3.5 湖泊健康影响因子分析

随机森林模型能够给出不同群组中影响湖泊健康的指标重要度。为便于与《规范》中给出的指标权重进行对比,对模型输出的重要度进行归一化处理。归一化后的指标重要度(以下简称“重要度”)在意义上与《规范》中的权重具有可比性,各群组湖泊指标重要度如表 4 和图 5 所示。结合表 4 和图 5 发现,生态水位满足程度(H_1)在群组 2~5 中为对湖泊健康影响最大的指标,重要度分别为 0.0710、0.0740、0.0720 和 0.0710,其在群组 1 中为第二重要的指标,重要度为 0.0670;天然湿地保留率(P_3)在群组 1 中为对湖泊健康影响最大的指标,重要度为 0.0725,其在群组 2~5 中为第二重要的指标,重要度分别为 0.0705、0.0695、0.0705 和 0.0695;岸带状况(P_2)在群组 1 和群组 2 中为第三重要的指

标,在群组 3 和群组 5 中为第四重要的指标,其重要度分别为 0.0598、0.0618、0.0598 和 0.0598;营养状态(W_2)在群组 1 和群组 2 中为第四重要的指标,在群组 3~5 中为第三重要的指标,其重要度分别为 0.0570、0.0590、0.0600、0.0590 和 0.0600;湖泊面积萎缩比例(P_1)在群组 4 中为第四重要的指标,其重要度为 0.0584。基于上述分析结果可知,五类群

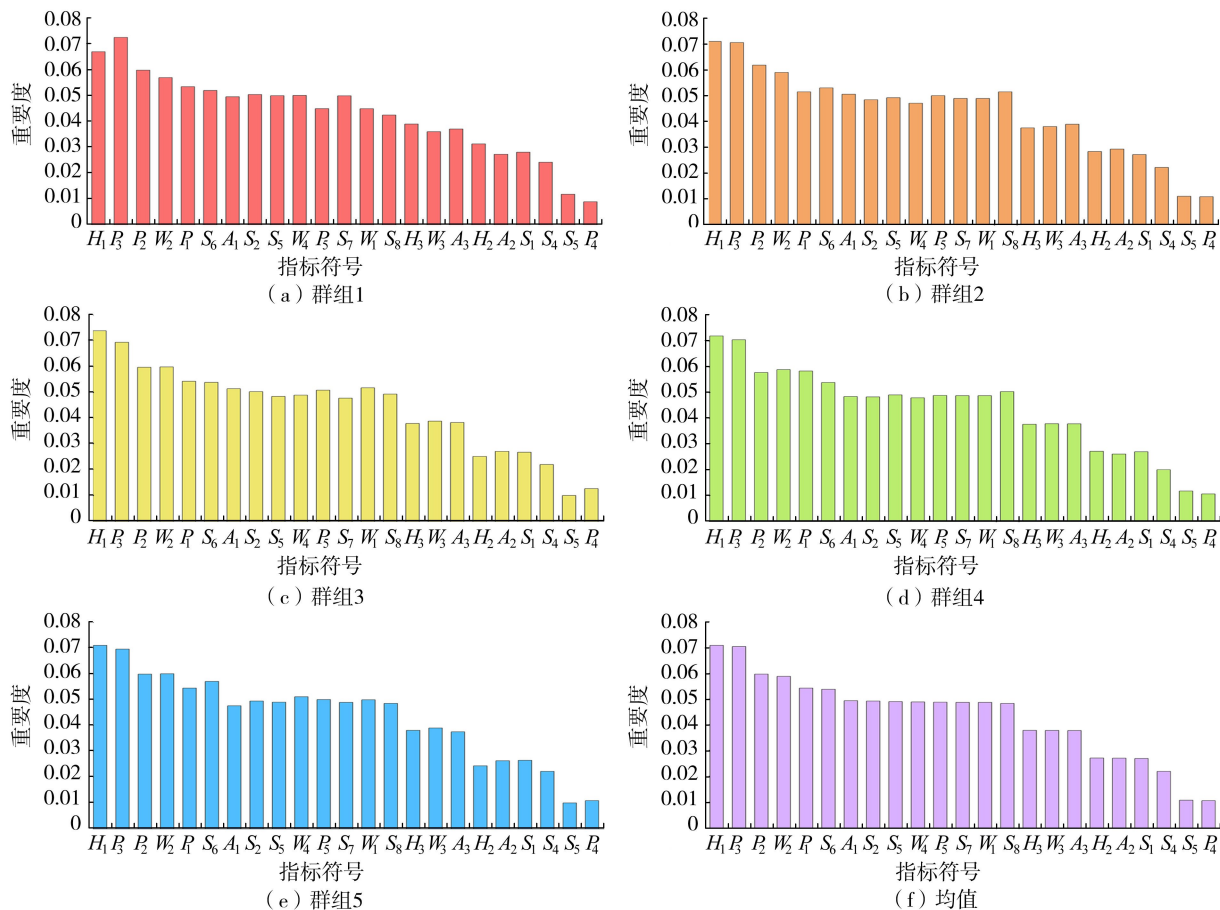


图 5 各群组湖泊健康评价指标重要度

Fig. 5 Importance of lake health evaluation indicators across different groups

组湖泊各指标重要度在数值上略有差异,但在整体排序上基本类似,其均值与《规范》中建议的指标权重及适用性检验模型中指标重要度对比见表 4。由表 4 可以看出,本文增加的评价指标如湖泊面积萎缩比例(0.054 4)、天然湿地保留率(0.070 5)、植被覆盖率(0.048 9)等的重要度均高于《规范》中所有评价指标的重要度均值(0.044 5),这表明本文增加的评价指标能够更全面地考虑湖泊的健康影响因素。

基于各群组湖泊指标重要度的分析结果,可以看出新增指标的平均重要度均高于《规范》中原有评价指标的重要度均值。如湖泊面积萎缩比例、天然湿地保留率、植被覆盖率,因此,应在现行措施基础上进一步扩大管理维度。加强生态水位调控以减缓湖泊面积萎缩,通过优化水资源配置、完善生态调度等措施维持湖泊合理水位;严格保护天然湿地,落实湿地保护红线,推进连通性恢复与退化湿地修复,提升生态系统稳定性;通过流域绿化、岸带植被重建和生态缓冲带建设等措施提高植被覆盖率,增强湖泊生态功能^[32]。可见,未来湖泊健康保护应在延续现有治理体系的基础上,同时兼顾其他重要指标的协同提升,以构建更加全面、系统和精细化的湖泊健

康管理体系。

3.6 湖泊健康评价结果

在对各群组随机森林模型进行最优参数配置的情况下,模型通过 Bootstrap 采样生成多棵具有差异性的学习器,基于大庆地区 60 个典型湖泊的各评价指标的实测值(调查值),分别对其湖泊健康进行评价后,采用投票机制形成最终大庆地区典型湖泊健康评价结果。分析结果可知,《规范》和适用性检验模型评价结果一致的湖泊共有 50 个,占总评价湖泊数 60 的 83.3%,这一结果表明随机森林模型能够较好表征评价指标与相应健康等级间的复杂映射关系,可应用于湖泊的健康评价中,这在一定程度上可规避《规范》中推荐的基于线性内插赋分的加权评价方法可能带来的权重确定过程中主观性较强的不足。

为进一步对比分析本文增加评价指标与群组划分前后两类随机森林模型得到的湖泊健康评价结果与基于线性内插赋分的加权评价方法所得评价结果的差异,现分别绘制了《规范》与适用性检验模型和群组随机森林模型评价结果对比的混淆矩阵热力图(图 6)。在热力图中,位于对角线的方格颜色越深表示两种方法得到相同评价结果的样本(湖泊)数

量越多,评价结果的相似性更高;非对角线上的方块则代表得到不同结果的湖泊数量,即不同方块中的数字表示研究区湖泊被评为不同等级的个数。例如,在图 6(b)中位于主对角线上的数值“29”,代表采用《规范》中推荐方法与指标得出的评价结果为亚健康的 43(2+7+29+5)个湖泊中有 29 个仍被评为亚健康;而未处于主对角线上的数值“7”,则表示采用《规范》中推荐方法与指标得出的评价结果为亚健康的 43 个湖泊中有 7 个被评为健康,这反映了群组随机森林模型评价结果与《规范》推荐的基于线性内插赋分的加权评价方法的评价结果的差异。由此可知,本文构建的基于随机森林算法的群组湖泊健康评价模型所得湖泊健康评价结果与《规范》推荐的基于线性内插赋分的加权评价方法所得结果相同的湖泊数为 45,占总体的 75%,可见所构建的模型性能在总体上与《规范》中推荐的加权评价方法较符合,但部分湖泊的评价结果也有所差异。

15 个湖泊中。15 个湖泊中的 11 个湖泊评价结果不同,是因评价指标体系不变、评价方法改变,其中 3 个湖泊评价结果不同是因评价指标体系、评价方法均改变。现以一二二泡和老江身泡为例进行具体分析。一二二泡在评价方法改变时,其健康等级由亚健康变为健康,但在后续改变评价指标体系时其评价结果不再发生变化;这是因为一二二泡的营养状态和浮游植物多样性指数两项指标评价结果均为非常健康,这两项指标的重要度(>0.3)在适用性检验模型和群组随机森林模型中均远高于《规范》推荐标准,对一二二泡的最终评价结果产生了积极影响。而老江身泡评价等级由亚健康变为健康后又变为非常健康,这是因为老江身泡在营养状态和浮游植物多样性指数这两个指标上表现优秀,使其评价结果由亚健康提升为健康,而本文新增指标湖泊面积萎缩比例、天然湿地保留率、植被覆盖率和水体自净能力的评价结果均处于健康及以上水平,所以其健康等级在群组随机森林模型评价结果中再次提升为非常健康。其余 4 个湖泊的评价结果不同,是因为评价指标体系和评价方法同时发生改变。现以马勒盖泡为例进行具体分析。马勒盖泡在评价方法发生改变时健康等级不变,在评价指标体系和评价方法同时发生改变时其健康等级由亚健康变为健康;这是因为马勒盖泡在本文新增指标年平均降水量和人口密度的评价结果为健康,植被覆盖率评价结果为非常健康,所以在采用群组随机森林模型进行评价时其健康等级得到了一定程度的提升。

4 结 论

a. 湖泊群组划分除了与湖泊特性密切相关外,在地理分布上亦能体现出显著的聚类结果,可用于群组随机森林湖泊健康评价模型的构建,以期提高湖泊健康评价模型的性能。基于选取的湖泊属性指标因子,应用 K-means 算法将研究区湖泊划分为 5 个群组,其典型特征分别为高浮游植物多样性指数、低植被覆盖率、高水体自净能力、高鱼类保有指数和优水质类别。

b. 构建的随机森林模型在大庆地区湖泊健康评价中具有较好的适用性,拓展了湖泊健康评价的研究途径。基于随机森林模型的湖泊健康评价结果与《规范》中推荐的加权评价方法的评价结果具有较好的符合性,能够较为准确地获取高维评价指标与健康状况的非线性映射关系,并克服《规范》中评价指标权重线性内插的不足,使评价结果更趋合理;模型计算的指标重要度可用于各群组湖泊健康主导影响因素的识别。

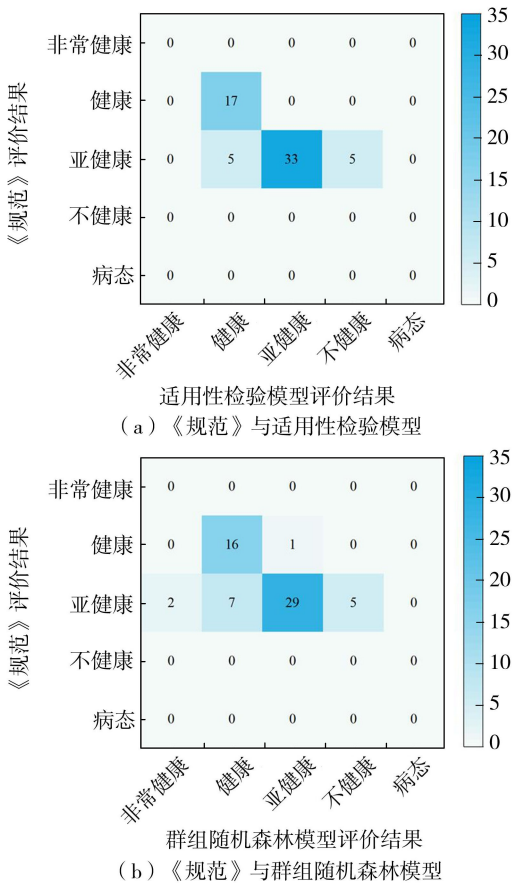


图 6 评价结果对比热力图

Fig. 6 Heatmap for evaluation result comparison

本文所构建的湖泊健康评价模型对各等级湖泊健康的评价在一定程度上克服了权重分配的主观性,且考虑的指标更全面。与《规范》推荐的基于线性内插赋分的加权评价方法所得评价结果相比,本文中评价方法对应的评价结果与其差异性体现在

c. 本文增加的评价指标能够更全面、综合地考虑湖泊的健康影响因子,完善了现有湖泊健康评价指标体系。5 类群组湖泊各指标重要度在数值上略有差异,但在整体排序上基本类似,本文增加的评价指标如湖泊面积萎缩比例、天然湿地保留率、植被覆盖率等的重要度均值均高于《规范》中所有评价指标的重要度均值,建议在现有治理体系基础上强化对新增高重要度指标的关注。

参考文献:

[1] 李茱,徐敏,王璐晨,等. 基于数据融合的高时空分辨率作物蒸散发反演与高效精细化灌溉决策[J]. 农业机械学报, 2025, 56 (8): 62-73. (Li Mo, Xu Min, Wang Luchen, et al. Data fusion-based high spatiotemporal resolution retrieval of crop evapotranspiration and efficient refined irrigation decision-making for irrigation areas[J]. Transactions of the Chinese Society for Agricultural Machinery,2025,56(8):62-73. (in Chinese))

[2] 姜翠玲,许睿亭. 水文要素对河湖生态环境的调控效应研究进展[J]. 水利水电科技进展, 2025, 45 (5): 119-128. (Jiang Cuiling, Xu Ruiting. Research progress on regulation effects of hydrological factors on ecological environment of rivers and lakes[J]. Advances in Science and Technology of Water Resources, 2025, 45 (5): 119-128. (in Chinese))

[3] Inacio M, Barcelo D, Zhao Wenwu, et al. Mapping lake ecosystem services:a systematic review[J]. Science of the Total Environment,2022,847:157561.

[4] 刘涛. 生态文明建设背景下水资源管理的重要性及实践举措分析[J]. 水资源保护, 2025, 41 (4): 268-269. (Liu Tao. Analysis of the importance and practical measures of water resource management in the context of ecological civilization development[J]. Water Resources Protection,2025,41(4):268-269. (in Chinese))

[5] Zhao Zhixuan, Wei Fangfei, Wu Haotian, et al. A framework to comprehensively assess lake health from a perspective of ecosystem integrity and services [J]. Ecological Indicators,2025,171:113169.

[6] 李冰,杨桂山,万荣荣. 湖泊生态系统健康评价方法研究进展[J]. 水利水电科技进展, 2014, 34 (6): 98-106. (Li Bing, Yang Guishan, Wan Rongrong. Progress on evaluation methods of lake ecosystem health[J]. Advances in Science and Technology of Water Resources, 2014, 34 (6): 98-106. (in Chinese))

[7] Padrón R S, Gudmundsson L, Greve P, et al. Large-scale controls of the surface water balance over land: insights from a systematic review and meta-analysis [J]. Water Resources Research,2017,53(11):9659-9678.

[8] Zutta B R, Salinas N, Cosio E G, et al. Satellite-derived

forest canopy greenness shows differential drought vulnerability of secondary forests compared to primary forests in Peru [J]. Environmental Research Letters, 2023, 18 (6): 064004.

[9] Zhang Zitong, Chen Xiaojun, Wang Chen, et al. Structured multi-view *k*-means clustering [J]. Pattern Recognition, 2025, 160: 111113.

[10] Bai Peng, Liu Xiaomang, Zhang Dan, et al. Estimation of the Budyko model parameter for small basins in China [J]. Hydrological Processes, 2020, 34 (1): 125-138.

[11] Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests [J]. Statistics and Computing, 2017, 27 (3): 659-678.

[12] Haddad K, Rahman A. Dimensionality reduction for regional flood frequency analysis: linear versus nonlinear methods [J]. Hydrological Processes, 2023, 37 (4): e14864.

[13] Wu Ying, Wang Yigang, Liu Hongbing, et al. Risk assessment of bridge construction investigated using random forest algorithm [J]. Scientific Reports, 2024, 14 (1): 20964.

[14] 韩铠合,许书晨,马克. 中国石油大庆油田当好保障国家能源安全“顶梁柱” [J]. 中国人大, 2024 (19): 56. (Han Kaihe, Xu Shuchen, Ma Ke. China National Petroleum Corporation ’ s Daqing Oilfield: a pillar supporting national energy security [J]. National People ’ s Congress of China, 2024 (19): 56. (in Chinese))

[15] 杨明杰. 统筹规划,推进大庆百湖之城建设 [J]. 大庆社会科学, 2008 (2): 71-74. (Yang Mingjie. Coordinate planning to advance the development of Daqing as the City of a Hundred Lakes [J]. DaQing Social Sciences, 2008 (2): 71-74. (in Chinese))

[16] Shi Congming, Wei Bingtao, Wei Shoulin, et al. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm [J]. EURASIP Journal on Wireless Communications and Networking, 2021 (1): 31.

[17] Ros F, Riad R, Guillaume S. PDBI: a partitioning Davies-Bouldin index for clustering evaluation [J]. Neurocomputing, 2023, 528: 178-199.

[18] Zheng Yulan. Customer segmentation research in marketing through clustering algorithm analysis [J]. Journal of Intelligent & Fuzzy Systems, 2023, 45 (4): 5441-5448.

[19] Wu Hao, Lin Anqi, Xing Xudong, et al. Identifying core driving factors of urban land use change from global land cover products and POI data using the random forest method [J]. International Journal of Applied Earth Observation and Geoinformation, 2021, 103: 102475.

(下转第 271 页)

Three Gorges Reservoir area [J]. Journal of Water Resources & Water Engineering, 2024, 35(2): 48-59. (in Chinese))

[45] 杨强强,徐光来,杨先成,等. 青弋江流域土地利用/景观格局对水质的影响[J]. 生态学报, 2024, 40(24): 9048-9058. (Yang Qiangqiang, Xu Guanglai, Yang Xiancheng, et al. Responses of water quality to land use & landscape pattern in the Qingyijiang River Watershed[J]. Acta Ecologica Sinica, 2024, 40(24): 9048-9058 (in Chinese))

+++++

(上接第 250 页)

[20] Rodriguez-Galiano V F, Ghimire B, Rogan J, et al. An assessment of the effectiveness of a random forest classifier for land-cover classification [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2012, 67: 93-104.

[21] Abedi R, Costache R, Shafizadeh-Moghadam H, et al. Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees[J]. Geocarto International, 2022, 37(19): 5479-5496.

[22] Tan Xuezhi, Li Yi, Wu Xinxin, et al. Identification of the key driving factors of flash flood based on different feature selection techniques coupled with random forest method [J]. Journal of Hydrology: Regional Studies, 2024, 51: 101624.

[23] 潘伯宁, 李文轩, 徐力刚, 等. 苏州河湖系统浮游植物完整性指数构建及水生态健康评价[J]. 河海大学学报 (自然科学版), 2025, 53(05): 73-80. (Pan Boning, Li Wenxuan, Xu Ligang, et al. Construction of phytoplankton index of biotic integrity and evaluation of water ecological health of lake and river systems in Suzhou City [J]. Journal of Hohai University (Natural Sciences), 2025, 53(05): 73-80. (in Chinese))

[24] 李晓瑛, 王华, 王屹晴, 等. 基于机器学习的长江口溶解氧预测模型与评估[J]. 环境科学, 2024, 45(12): 7123-7133. (Li Xiaoying, Wang Hua, Wang Yiqing, et al. Machine learning-based dissolved oxygen prediction modeling and evaluation in the yangtze river estuary[J]. Environmental Science, 2024, 45(12): 7123-7133. (in Chinese))

[25] 杜云彬, 陈求稳, 王智源, 等. 江苏省典型湖泊饮用水源地安全综合评价[J]. 水资源保护, 2020, 36(5): 71-78, 92. (Du Yunbin, Chen Qiwen, Wang Zhiyuan, et al. Safety evaluation of typical lake drinking water sources in Jiangsu Province[J]. Water Resources Protection, 2020, 36(5): 71-78, 92. (in Chinese))

[46] 刘可暄, 王冬梅, 常国梁, 等. 多空间尺度景观格局与地表水质响应关系研究[J]. 环境科学学报, 2022, 42(2): 23-31. (Liu Kexuan, Wang Dongmei, Chang Guoliang, et al. Study on the relationship between landscape pattern and surface water quality at multiple spatial scales[J]. Acta Scientiae Circumstantiae, 2022, 42(2): 23-31. (in Chinese))

(收稿日期: 2025-09-29 编辑: 胡新宇)

+++++

[26] 樊贤璐, 徐国宾. 基于生态-社会服务功能协调发展度的湖泊健康评价方法[J]. 湖泊科学, 2018, 30(5): 1225-1234. (Fan Xianlu, Xu Guobin. Lake health assessment method based on the coordinated development degree of ecology and social service function[J]. Journal of Lake Sciences, 2018, 30(5): 1225-1234. (in Chinese))

[27] 薛联青, 朱祎真, 王思琪, 等. 基于投影寻踪的湖泊生态健康评价[J]. 水资源保护, 2016, 32(2): 67-72. (Xue Lianqing, Zhu Yizhen, Wang Siqi, et al. Lake ecosystem health assessment based on projection pursuit[J]. Water Resources Protection, 2016, 32(2): 67-72. (in Chinese))

[28] 陈星, 许钦, 何新玥, 等. 城市浅水湖泊生态系统健康与保护研究[J]. 水资源保护, 2016, 32(2): 77-81. (Chen Xing, Xu Qin, He Xinyue, et al. Research on health evaluation and protection of urban shallow lake ecosystem [J]. Water Resources Protection, 2016, 32(2): 77-81. (in Chinese))

[29] Murtagh F, André H. Multivariate data analysis [J]. Technometrics, 2006, 19(3): 549-563.

[30] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[31] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks [J]. Information Processing & Management, 2009, 45(4): 427-437.

[32] 陈文龙, 吴琼, 李一平. 城市河流低水位运行生态修复机制与实践效果[J]. 水资源保护, 2024, 40(2): 16-22. (Chen Wenlong, Wu Qiong, Li Yiping. Ecological restoration mechanism and practical effects of low water level operation in urban rivers [J]. Water Resources Protection, 2024, 40(2): 16-22. (in Chinese))

(收稿日期: 2025-11-06 编辑: 彭桃英)