中国水产科学 2024 年 5 月, 31(5): 602–612
Journal of Fishery Sciences of China

研究论文

DOI: 10.12264/JFSC2023-0349

南海北部渔业生物声学密度的底表层间差异及与多类非生物因子的相关性分析

孙铭帅[1], 蔡研聪[1], 张魁[1], 许友伟[1], 杨玉滔[1], 陈作志[1, 2]

1. 中国水产科学研究院南海水产研究所, 农业农村部外海渔业可持续利用重点实验室, 广东 广州 510300
2. 广东省渔业生态环境重点实验室, 广东 广州 510300

摘要: 本研究旨在分析南海北部渔业生物在不同水层(表层混合层和底层冷水层)间声学密度的差异, 并探讨这种差异与 41 种非生物因子的相关关系, 以期为南海北部渔业资源的有效管理和保护提供科学依据。研究采用渔业声学方法, 使用 Simrad EY60 分裂波束科学探鱼仪在南海北部进行声学数据采集。通过 Echoview 渔业声学数据处理系统分析声学数据, 计算表层和底层的声学密度(NASC)。采用极限梯度提升算法(XGBoost)和随机森林算法(random forest)建模分析 41 种非生物因子对声学密度差异的影响, 并评估因子的重要性。研究发现, 底层渔业生物声学密度明显高于表层, 底层平均值为 106.00 m^2/nmi^2, 表层为 43.39 m^2/nmi^2。极限梯度提升算法和随机森林算法的建模效果相似, 重要性分析显示, 温度因素(底层 2 m 温度、表–底温度差、表层 2 m 温度)和水深是影响声学密度差异的最关键因素。南海北部渔业资源的表层多、底层少的负值区域主要分布在海南岛周边。温度和水深是影响渔业生物分布差异的主要因素, 而人类活动对磷酸盐、叶绿素等因子的调节也可能对声学密度差异产生影响。这些发现为南海北部渔业资源的管理和保护提供了重要科学依据。

关键词: 非生物因子; 极限梯度提升算法(XGBoost); 随机森林(Random Forest); 渔业声学; 南海北部
中图分类号: S931 **文献标志码:** A **文章编号:** 1005−8737−(2024)05−0602−11

南海北部近海作为中国重要的传统渔业生产作业海域, 同时也是海洋鱼类的重要产卵场与育肥场。近年来, 该区域渔业生物在各栖息水层(尤其是表层混合层和底层冷水层)呈现不同程度的低龄化、小型化、低质化现象, 引起了学术界和渔业管理部门的高度关注[1-2]。南海北部近海渔业资源的分布特征环境影响因素已有类似研究开展[1-3], 而水层间分布特征差异化与环境因子的关系则是在此类研究基础上的深入探索。

渔业声学方法是一种海洋生物资源调查与评估的重要技术手段, 与传统的底拖网采样相比, 渔业声学方法具有直观、快捷、连续、取样率大和时空数据丰富等特点, 通过声学数据后处理获得的海面散射系数(NASC)能够直观反映调查海域物种资源分布特征[4-9]。尤其在本研究中, 该方法能够分层分析渔业生物的声学密度, 并进一步对不同水层间的渔业生物声学密度差异进行深入分析。底表层渔业生物声学密度的层间差异有助于反映不同区域渔业生物分布的水层宜居情况以及表层和底层种类的资源量差异, 而引起这种层间差异的原因尚缺少详细的研究信息。

极限梯度提升算法(XGBoost)[10]是近年来兴起的高效集成算法, 在分类和回归上表现出超高性能。随机森林(random forest)算法则是装袋法(bagging)的代表[11]。这两种算法在图像分类[12-13]、数据分析[14-16]、信息分类[17-18]以及特征变量重要性分析[11,19-20]等多个领域均有广泛应用。为深入挖掘引起渔业资源表-底层间差异的限制因子和

收稿日期: 2023-12-29; **修订日期:** 2024-01-30.
基金项目: 广东省重点研发计划项目(2020B11111030001); 中国水产科学研究院中央公益性科研机构基础研究基金项目(2023TD05); 中国水产科学研究院中央级公益性科研院所基本科研业务费专项资金资助项目(2021SD01).
作者简介: 孙铭帅(1986−), 男, 博士, 助理研究员, 研究方向为渔业声学. E-mail: sunmingshuai@scsfri.ac.cn
通信作者: 陈作志, 研究员, 研究方向为海洋渔业资源评估和管理. E-mail: chenzuozhi@scsfri.com

潜在因素, 本研究分析了表层混合层(表层 20 m)和底层冷水层(底层 20 m)渔业生物声学密度层间差异的短期空间分布, 并分析其与 41 种非生物因子之间的相关关系, 以期对南海北部渔业资源的有效管理和保护提供科学依据。

1　材料与方法

1.1　调查区域及采样

2014 年 7—8 月, 由“北渔 60011”单船底拖网渔船(渔船主机功率 441 kW, 总吨位 242 t, 船体长度 36.8 m, 宽度 6.8 m)搭载双频 Simrad EY60 便携式分裂波束科学探鱼仪(70 kHz 和 120 kHz, 设备参数见表 1)对南海北部近海区域进行声学数据采集, 声学设备使用固定装置固定于左舷中部, 换能器入水深度为 1.5 m。调查开始前按照规范对声学设备进行校准。走航路线见图 1。

渔获物样品由单船底拖网采集。网具类型为 404 型底拖网, 网口周长 80.80 m, 网衣全长 60.54 m, 上纲长度 37.7 m, 网口网目尺寸 20 cm, 网囊网目尺寸 39 mm。共 99 个拖网站位, 每站作业一网, 拖网时间为约 60 min。对每种渔获物进行体重、体长测量, 并统计每种的总重和总数。

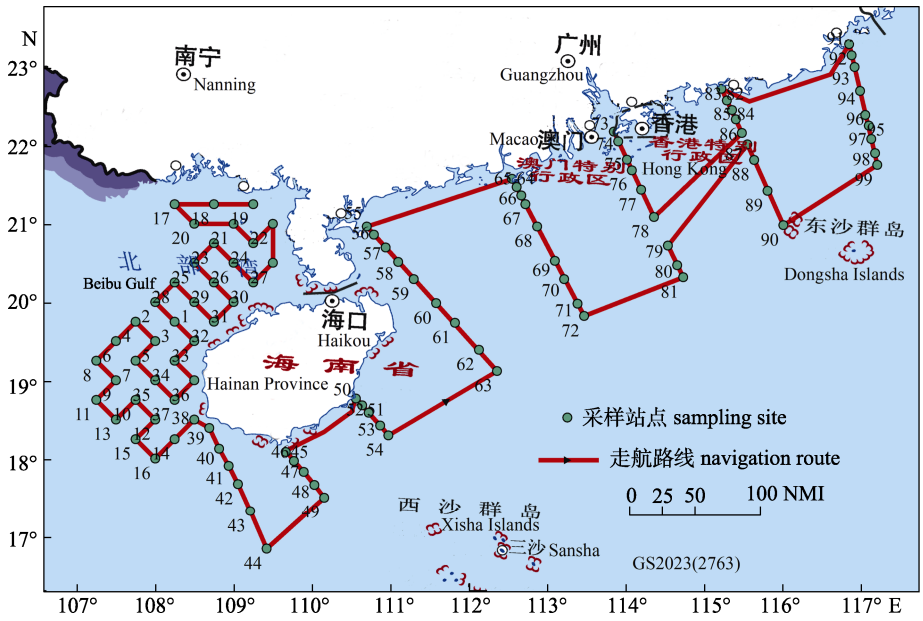

图 1　走航路线及采样站点示意图

Fig. 1　Map of navigation route and sampling site

表 1　Simrad EY60 便携式分裂波束科学探鱼仪相关参数

Tab. 1　Main technical parameters setting of Simrad EY60

技术参数 technical parameter	参数设置/ kHz value	
	70	120
发射功率/W transmitting power	800	500
脉冲宽度/ms pulse width	0.512	0.512
探测范围/m detection range	1000	800
换能器增益/dB gain of the converter	27.00	27
纵向波束角度/(°) longitudinal beam angle	7.00	7.00
横向波束角度/(°) horizontal beam angle	7.00	7.00
吸收系数/(dB/m^2) absorption coefficient	0.018	0.045
声速/(m/s) sound velocity	1535	1535
波束等效立体角/dB beam equivalent solid angle	21	21

水质环境特征中水温、盐度、水深使用 AML Plus X 多参数剖面仪探测获得, 其余特征如营养盐、透明度等均按照《海洋调查规范》中的规范方法采样, 其中每站位营养盐采样水深为 0 m、10 m、20 m。

1.2　表层、底层声学密度

研究过程中发现调查海域温跃层垂直范围多在水面以下 20 m 至水底以上 20 m 之间, 温跃层以外的表层混合层和底层冷水层是温度稳定区, 因此, 为研究温跃层外上下水层间的渔业生物声学密度与特征间的重要性关系, 设置分析水层深度为 20 m, 包括表层 20 m (即表层线以下 20 m 范围, 不包含声学盲区)、底层 20 m (即水底线以上

20 m 范围, 不包含声学盲区)。基本积分航程单元设为 5 nmi。利用 Echoview 渔业声学数据处理系统(Version 6.1)进行声学数据分析。所有声学数据都被逐一仔细检查, 非走航时段内的数据不用于资源密度计算。积分阈值设置为−80 dB。对于背景噪声较多的区域, 采用背景噪声移除模块(background noise removal)对背景噪声进行后期消除, 分别获得表层、底层 NASC 积分值(nautical area scattering coefficient, m^2/nmi^2)。因所取表层和底层的 NASC 值均为 20 m 范围, 所以表层 NASC 和底层 NASC 即为相同水体体积的渔业生物声学密度积分值, 进一步计算底层声学密度与表层声学密度差异(即底层 NASC 减去表层 NASC 的差值)。

1.3 非生物因子

对渔获物信息按照常居水层分为两类: 表层类表层和近底层类。其中头足类白天常居近底层, 夜间往往上移至表层, 所以将所有头足类渔业生物按照 0.5 : 0.5 的比例分别计入表层类和近底层类。

水质环境因素包括初级特征及衍生特征(表 2), 初级特征包括表层 2 m 处盐度(SS)、表层 2 m 处温度(ST)等 21 项因子; 由初级特征间计算得到的衍生特征包括底-表盐度差(DS)、表-底温度差(DT)、NO_2^- 0 m 与 10 m 浓度差(N2-d010)等 17 项非生物因子; 还有地理信息特征包括水深(WD, m)、经度(Lon)、纬度(Lat) 3 个因子; 初级特征共有 21 种; 衍生特征共 17 个。同时, 尝试将初级特征中 0 m、10 m 以及透明度(TRA)、叶绿素(CHL)的水质环境因素定义为表层初级非生物因子, 将 20 m 的水质环境因素定义为近底层初级非生物因子, 其余带有“表层”字样的初级非生物因子均

表 2 因子列表及分类

Tab. 2 List and grouping of all factors

分类 group	因子缩写 factor abbreviation	因子全称 name	单位 unit	备注 note
空间因子 spatial factor	water depth	水深 water depth	m	
	X	经度 longitude	°	
	Y	纬度 latitude	°	
初级因子 primary factor	SS	海表盐度 surface salinity		表层 surface
	BS	海底盐度 bottom salinity		底层 bottom
	ST	海表温度 surface temperature	℃	表层 surface
	BT	海底温度 bottom temperature	℃	底层 bottom
	TRA	透明度浓度 transparency	m	表层 surface
	CHL	叶绿素浓度 chlorophyll concentration	mg/m^3	表层 surface
	N2M0	亚硝酸盐在 0 m 的浓度 NO_2^- 0 m concentration	mg/L	表层 surface
	N2M10	亚硝酸盐在 10 m 的浓度 NO_2^- 10 m concentration	mg/L	表层 surface
	N2M20	亚硝酸盐在 20 m 的浓度 NO_2^- 20 m concentration	mg/L	底层 bottom
	N3M0	硝酸盐在 0 m 的浓度 NO_3^- 0 m concentration	mg/L	表层 surface
	N3M10	硝酸盐在 10 m 的浓度 NO_3^- 10 m concentration	mg/L	表层 surface
	N3M20	硝酸盐在 20 m 的浓度 NO_3^- 20 m concentration	mg/L	底层 bottom
	N4M0	铵盐在 0 m 的浓度 NH_4^+ 0 m concentration	mg/L	表层 surface
	N4M10	铵盐在 10 m 的浓度 NH_4^+ 10 m concentration	mg/L	表层 surface
	N4M20	铵盐在 20 m 的浓度 NH_4^+ 20 m concentration	mg/L	底层 bottom
	P0	磷酸盐在 0 m 的浓度 PO_4^{3-} 0 m concentration	mg/L	表层 surface
	P10	磷酸盐在 10 m 的浓度 PO_4^{3-} 10 m concentration	mg/L	表层 surface
	P20	磷酸盐在 20 m 的浓度 PO_4^{3-} 20 m concentration	mg/L	底层 bottom
	Si0	硅酸盐在 0 m 的浓度 SiO_3^{2-} 0 m concentration	mg/L	表层 surface
	Si10	硅酸盐在 10 m 的浓度 SiO_3^{2-} 10 m concentration	mg/L	表层 surface
	Si20	硅酸盐在 20 m 的浓度 SiO_3^{2-} 20 m concentration	mg/L	底层 bottom
衍生因子 derived factor	DS	表底盐度差异 salinity difference between surface and bottom layers	ppt	
	DT	表底温度差异 temperature difference between surface and bottom layers	℃	
	N2D010	亚硝酸盐在 0 m 和 10 m 的浓度差 concentration difference between NO_2^- 0 m and 10 m	mg/L	

(待续 to be continued)

(续表 2　Tab. 2 continued)

分类 group	因子缩写 factor abbreviation	因子全称 name	单位 unit	备注 note
衍生因子 derived factors	N2D020	亚硝酸盐在 0 m 和 20 m 的浓度差 concentration difference between NO_2^- 0 m and 20 m	mg/L	
	N2D1020	亚硝酸盐在 10 m 和 20 m 的浓度差 concentration difference between NO_2^- 10 m and 20 m	mg/L	
	N3D010	硝酸盐在 0 m 和 10 m 的浓度差 concentration difference between NO_3^- 0 m and 10 m	mg/L	
	N3D020	硝酸盐在 0 m 和 20 m 的浓度差 concentration difference between NO_3^- 0 m and 20 m	mg/L	
	N3D1020	硝酸盐在 10 m 和 20 m 的浓度差 concentration difference between NO_3^- 10 m and 20 m	mg/L	
	N4D010	铵盐在 0 m 和 10 m 的浓度差 concentration difference between NH_4^+ 0 m and 10 m	mg/L	
	N4D020	铵盐在 0 m 和 20 m 的浓度差 concentration difference between NH_4^+ 0 m and 20 m	mg/L	
	N4D1020	铵盐在 10 m 和 20 m 的浓度差 concentration difference between NH_4^+ 10 m and 20 m	mg/L	
	P010	磷酸盐在 0 m 和 10 m 的浓度差 concentration difference between PO_4^{3-} 0 m and 10 m	mg/L	
	P020	磷酸盐在 0 m 和 20 m 的浓度差 concentration difference between PO_4^{3-} 0 m and 20 m	mg/L	
	P1020	磷酸盐在 10 m 和 20 m 的浓度差 concentration difference between PO_4^{3-} 10 m and 20 m	mg/L	
	Si010	硅酸盐在 0 m 和 10 m 的浓度差 concentration difference between SiO_3^{2-} 0 m and 10 m	mg/L	
	Si020	硅酸盐在 0 m 和 20 m 的浓度差 concentration difference between SiO_3^{2-} 0 m and 20 m	mg/L	
	Si1020	硅酸盐在 10 m 和 20 m 的浓度差 concentration difference between SiO_3^{2-} 10 m and 20 m	mg/L	

为表层初级特征, 带有“底层”字样的初级非生物因子均为底层初级特征。

1.4　数据随机重采样

受采样站位数量限制, 南海北部近海区域可用于分析使用的每种水质环境因素的样本数量不足 100 个, 不利于获得优异的数据分析质量。但各样本均有详细的经纬度坐标数据, 因此可利用空间插值法, 对有限的样本信息进行由点到面的数据信息扩展, 再将扩展后的面数据信息进行随机取样, 进而进行不同取样数的模型效果分析。其中, 使用的空间插值方法包括迭代优化后的普通克里格插值以及反距离权重插值, 选用插值方法的原则是交叉验证中以最高拟合优度(R^2)、最小均方误差(MSE)为首选, 另外随机取样数设定为 100、200、300、400、500、600, 且每轮取样设定取样点间最小间距为 5 nmi。汇总后的 2100 个点没有最小间距限制。

1.5　数据建模、验证及非生物因子重要性计算

本研究旨在探究 41 种水质环境类非生物因子对底层与表层间差异的空间分布特征的影响, 为此采用了极限梯度提升算法和随机森林这两种机器学习算法, 并以严格的交叉验证方法来评估模型性能, 采用 R^2 分值和均方误差(MSE)作为评价标准。研究中, 训练数据集与测试数据集的分配比例为 70%和 30%。

在方法论上, 随机森林和极限梯度提升算法的主要区别在于模型构建的过程。随机森林通过同时建立多棵独立的决策树来进行学习, 而极限梯度提升算法则在每次迭代中, 将注意力集中在前一棵树未能有效模型化的损失上, 即第 N+1 棵树专注于第 N 棵树的误差改进。在本研究中对几个重要参数进行了调优, 包括学习率(learning_rate)、评估

器数量(n_estimators)和子样本(subsample)等[10]。

通过这些方法，利用极限梯度提升算法和随机森林算法分别计算了 41 种非生物因子对底层-表层层间 NASC 积分值差的影响，并得出了这些因子的权重和重要性分值(即贡献率)。此外，本研究还采用了嵌入法进行因子筛选。嵌入法是基于在算法训练中得到的特征权值系数来选择特征，相对于过滤法和包装法，这种方法在特征选择上更为高效和精确。

所用软件为 Python3.7，代码为 Scikit-learn 中相关封装包[21]及独立的极限梯度提升算法库。

1.6 非生物因子重要性等级

为避免因算法差异对非生物因子重要性的偏重，将极限梯度提升算法和随机森林两种高效算法所获的重要性分值进行对应并计算均值，以此判定影响底-表渔业生物声学密度差异的重要水质环境因素。以累计总贡献率 50%、80%、95%为分界点，划分 4 个重要性等级，即第一级至第四级，其中第一级为重要性最高等级，第四级为重要性最低等级。

2 结果与分析

2.1 底层-表层渔业生物声学密度差异的空间分布特征及渔获物组成

图 2 中，分别采用几何间隔和正负值进行间隔分类，平均值 62.62 m^2/nmi^2，标准差 78.01 m^2/nmi^2，其中表层均值 43.39 m^2/nmi^2，标准差 54.73 m^2/nmi^2，底层均值 106.00 m^2/nmi^2，标准差 91.02 m^2/nmi^2。负值区域(表层多，底层少)表示偏重表层鱼类适宜区域，正值区域表示偏重近底层鱼类适宜区域。负值区域(绿色)多分布于沿岸海域，正值(粉色)范围分布广泛。底层平均值是表层平均值的 2 倍以上，底层资源密度高于表层密度。另据表 3、表 4 可知，渔获物种类较丰富，且由表 4 中近底层种类与表层种类的重量和尾数数据相除可知，底表种类的重量比和尾数比分别为 2.13、1.94，与底层、表层 NASC 平均值的比值相吻合。

2.2 样本数量与极限梯度提升算法模型表现

图 3 中显示，在未对极限梯度提升算法进行调参时(默认参数状态)，样本点 100 时获得的拟合度最低，随着样本点增加，拟合度有波动地缓慢提高，至 2100 点时缓慢上升为最高值。MSE 变化跟 R^2 的变化特征有一些相似，存在波动性。R^2 和 MSE 变化表明，样本点不足会严重影响模型效果，必须达到足够样本点才能有稳定效果。

2.3 最优样本量中各算法比较

图 4 显示，随机森林和极限梯度提升算法模型效果相似，其中 R^2 显示极限梯度提升算法略优于随机森林(0.864>0.859)，而 *MSE* 显示随机森林略优于极限梯度提升算法(819<907)，说明两种方法各有优势。线性回归对数据的拟合效果远差于前两者，表现为较低的 R^2 和较高的 *MSE*。

2.4 非生物因子重要性预排序及因子筛选

据图 5，极限梯度提升算法和随机森林的主要特征排序有异同之处。相同之处在于第一级重要特征(前 3 项累计贡献率 50%)完全相同，不同之处在于其他分级特征的排序存在不同程度的差异，但非生物因子重要性分值相差不大。

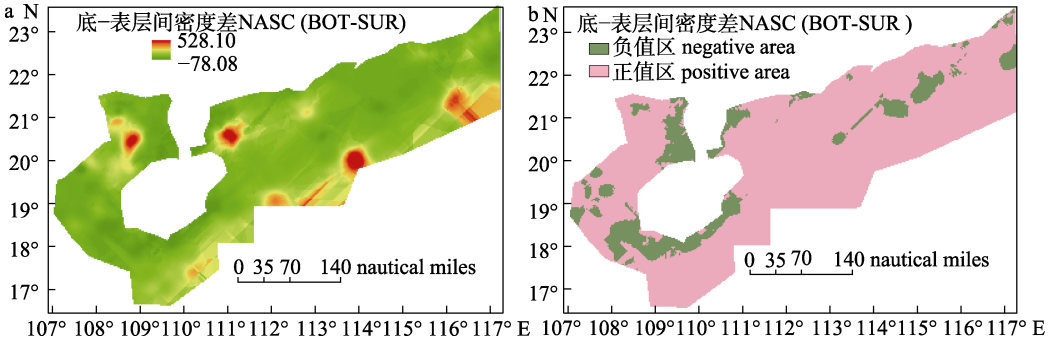

图 2 近底层、近表层渔业生物资源密度差异的分布

Fig. 2 The distribution of fishery resource density difference between bottom layer and surface layer

表 3　渔获物生物量和个体数分类统计

Tab. 3　Classification and statistics of catch weight and quantity

种类 fishery species	栖息水层 habitat layer	重量/kg weight	数量 number
鲅类 (*Pneumatophorus japonicus, Rastrelliger kanagurta*)	表层 surface layer	17.166	165
鲳类 (*Ariomma indica, Psenopsis anomala*)		603.075	8292
蓝圆鲹 (*Decapterus maruadsi*)		774.339	23597
竹筴鱼 (*Trachurus japonicus*)		937.429	25117
沙丁鱼类 (*Sardinella aurita, Sardinella jussieu*)		290.454	23600
带鱼类 (*Tentoriceps cristatus, Trichiurus haumela, Trichiurus nanhaiensis, Trichiurus brevis*)	近底层 near-bottom layer	237.824	3399
黄鳍马面鲀(*Navodon xanthopterus*)		276.389	13139
白姑鱼类 (*Argyrosomus aneus, Argyrosomus macrocephalus, Argyrosomus pawak, Argyrosomus argentatus*)		151.949	9170
蛇鲻类 (*Saurida undosquamis, Saurida tumbil, Saurida elongata*)		206.032	2786
二长棘犁齿鲷 (*Evynnis cardinalis*)		784.218	26980
大眼鲷类 (*Priacanthus macracanthus, Priacanthus tayenus*)		269.555	8107
金线鱼类 (*Nemipterus virgatus, Nemipterus bathybius, Nemipterus oveni, Nemipterus japonicus, Nemipterus nemurus*)		412.087	9362
绯鲤类 (*Upeneus bensasi, Upeneus sulphureus, Upeneus moluccensis, Parupeneus chrysopleuron*)		116.327	4250
篮子鱼类 (*Siganus oramin, Siganus fuscescens*)		66.004	7241
发光鲷类 (*Acropoma japonicum, Acropoma hanedai*)		414.008	8343
剑尖枪乌贼 (*Loligo edulis*)	白天近底层, 夜间表层 near-bottom layer during the day, surface layer at night	233.187	6234
中国枪乌贼 (*Loligo chinensis*)		43.991	1339
其他头足类 (other cephalopods)		1280.103	269370
其他评估种类 (other species of assessment)	大部分近底层 near-bottom layer (almost)	3524.761	195051

表 4　渔获物分层统计

Tab. 4　Stratified statistics of catch

栖息水层 habitat layer	声学密度/(m^2/nmi^2) NASC	重量/kg weight	数量 number	重量百分比/% weight percentage	数量百分比% number percentage
表层 surface	43.39	3401.10	219243	0.32	0.34
近底层 near-bottom	106.00	7237.79	426299	0.68	0.66

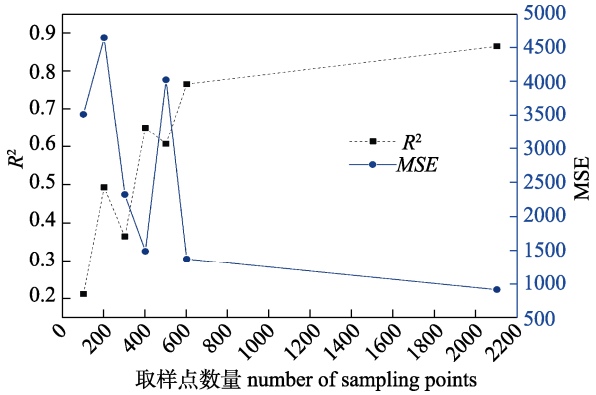

图 3　不同样本点数时利用极限梯度提升算法(XGBoost)进行的底表渔业生物声学密度差异与 41 种非生物因子的 R^2 (黑色方块虚线)及 MSE (蓝色圆点实线)变化

Fig. 3　At different sample points, the changes in R^2 (dotted black square) and MSE (solid blue dot) between the Bottom-Surface fishery resource density and 41 abiotic factors using XGBoost

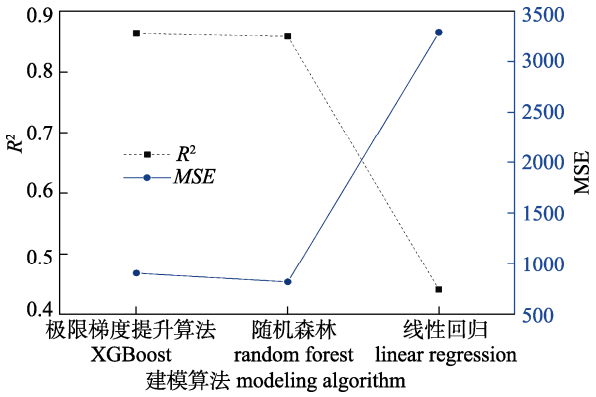

图 4　最优样本量时(2100 个点), 极限梯度提升算法、随机森林、线性回归建模并调参后, 底表渔业生物声学密度差异与 41 种非生物因子的 R^2 (黑色方块虚线)及 MSE (蓝色圆点实线)变化

Fig. 4　When the optimal sample size is 2100 points, the changes in R^2 (dotted black square) and MSE (solid blue dot) between the bottom-surface fishery resource density and 41 abiotic factors, after modeling and parameter tuning by xgboost, random forest and linear regression

从图 6 来看，随着阈值越来越高，全数据集情况下模型的效果逐渐变差，被删除的特征越来越多，信息损失也逐渐变大。阈值为 0.024 (41 项因子的贡献率均值)时，极限梯度提升算法模型因子仅剩 9 项(BT, DT, WD, CHL, P-20 m, P-d1020, N4-d010, ST, N4-0 m)，但 R^2 依然维持在 85%以上，而 RF 模型因子仅剩 6 项(BT, DT, WD, P-20 m, ST, N4-d020)，R^2 维持在 87%以上。

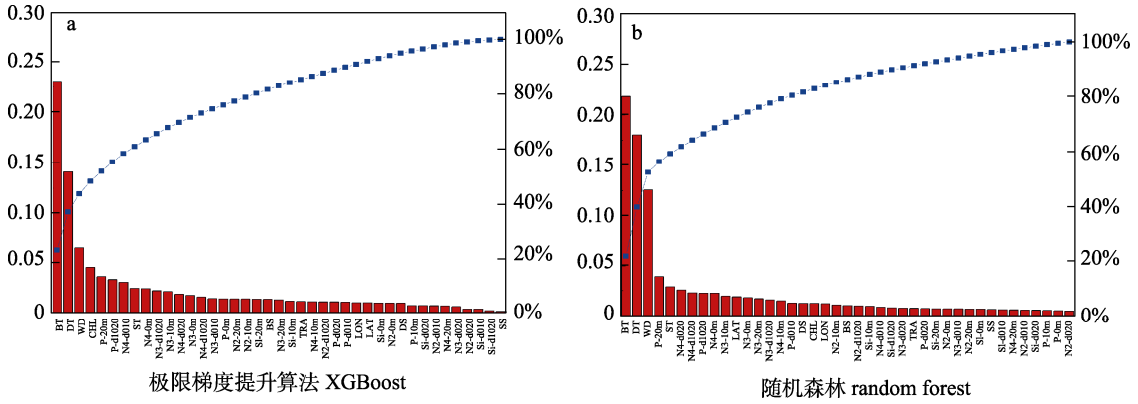

图 5 不同算法评估的底表渔业生物声学密度差与 41 种非生物因子的相关性

Fig. 5 Correlation between acousitic density difference of bottom-surface fishert and 41 abiotic factors based on XGBoost or random forest

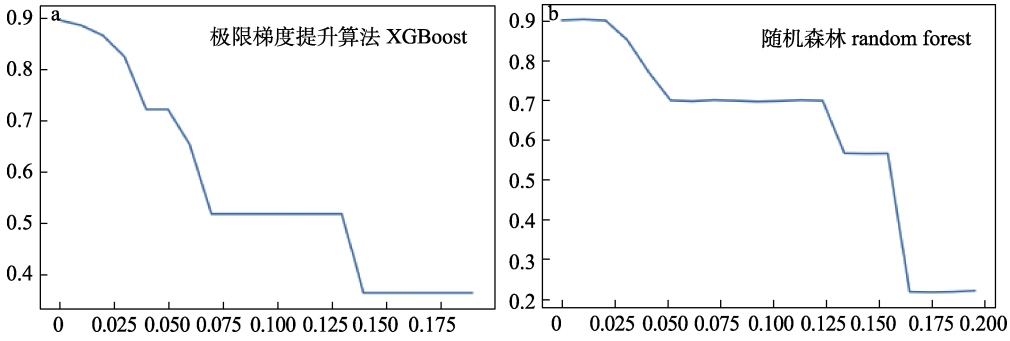

图 6 嵌入法分析非生物因子贡献率阈值与全数据集情况下模型效果(R^2)的关系

Fig. 6 Relationship between abiotic factor contribution rate threshold and model effect R^2 with full data set

2.5 非生物因子重要性等级

据图 7，影响底表渔业生物声学密度分布的第一级非生物因子总分值约 0.5 (特征贡献率 50%)，包括底层 2 m 处温度(BT)、表-底温度差(DT)、水深(WD)，占总分值的一半，与其他 39 个非生物因子重要性分值之和相同。第一级因子可被认定为重要环境因子，其中底层 2 m 处温度(BT)重要性分值 0.22 (特征贡献率 22%)>0.20，是极大概率影响因素，表-底温度差(DT)重要性分值 0.16 (特征贡献率 16%)>0.15，是大概率影响因素，水深(WD)重要性分值 0.10 (特征贡献率 10%)介于 0.05 和 0.15 之间，是较大概率影响因素。第二级非生物因子总分值约 0.3 (特征贡献率 30%)，主要包括 15 种非生物因子，如 PO_4^{3-} 20 m 浓度(P-20 m)、叶绿素(CHL)等。第三级非生物因子总分值约 0.15 (特征贡献率 15%)，包括 14 种非生物因子，如底层 2 m 处盐度(BS)、PO_4^{3-} 0 m 与 10 m 浓度差(P-d010)等。第四级无关非生物因子总分值约 0.05 (特征贡献率 5%)，包括 9 种非生物因子，是重要性分值最低的一组，主要包括 NO_3^- 0 m 与 20 m 浓度差(N3-d020)、NH_4^+ 20 m 浓度(N4-20 m)等。

3 讨论

3.1 极限梯度提升算法和随机森林建模中的特征选择与排序分析

通过不同样本点数量与建模效果比较发现，

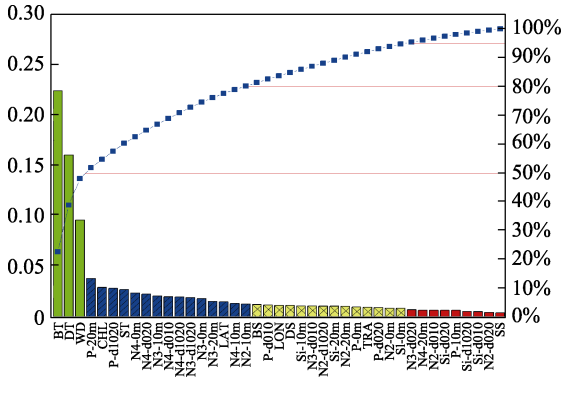

图 7　41 种非生物因子的重要性综合排序

绿色、蓝色、黄色、红色分别表示第一级至第四级特征，累计贡献率区间分别为 0%–50%、50%–80%、80%–95%、95%–100%.

Fig. 7　Comprehensive ranking of the importance of 41 abiotic factors

Green, blue, yellow and red represent the first level to the fourth level respectively, and the cumulative contribution rate ranges are 0%–50%, 50%–80%, 80%–95%, 95%–100%, respectively.

过少的样本点很难取得良好建模效果，并且在样本点数量增加的过程中，建模效果会出现波动，随着样本点数量不断增加直至足够多时，才会获得稳定良好的建模效果。值得一提的是，对于具有精确坐标位置的各类数据，通过空间插值法可有效将源数据进行样本数据模拟、样本数量扩充[22-27]，间接提升极限梯度提升算法的建模效果。除样本点数量外，极限梯度提升算法和随机森林均有多个重要参数可被用于调整以期获得最优建模效果，如对周期性时间序列的多特征样本数据进行建模分析时，可利用极限梯度提升算法和随机森林算法通过调参获得最优回归模型或最优分类模型。调优后的模型可用于已知非生物因子条件下的未知表层、底层资源密度的预测。

由于极限梯度提升算法和随机森林的运算机制差异[10-11]，其得到的各特征的重要性排序不完全相同。尽管如此，最终得到的重要特征的排序却大致相同，尤其是一些对目标数据影响较大的特征，说明相关性高的因素即使在不同的算法之间也是可以被有效识别的；但另一方面，相关性相对较弱的因素往往因为算法差异而表现不同。除了这两种集成学习算法进行的特征排序，还有主成分分析(principal component analysis, PCA)、支持向量机(SVM)[28]、逻辑回归[29-31]等多种机器学习算法可用于特征选择及降维，以及用于挖掘数据中暗含的各种关系的典型相关分析(canonical correalation analysis, CCA)。

3.2　地理因子、初级环境因子、衍生环境因子对底表层间分布差异的重要性

底表渔业生物声学密度分布差异对应地理信息特征、初级特征、衍生特征的重要性分值的和分别为 0.1216、0.5193、0.3591，平均值分别为 0.0405、0.0247、0.0211。由于各类特征的数量不等，所以从平均值来看，底表渔业生物声学密度分布差异，其相关影响因素特征更倾向于地理信息特征，而以衍生特征最弱。部分初级非生物因子重要性较高，如底层 2 m 处温度、表层 2 m 处温度，有部分衍生的衍生特征也具有较高的重要性分值，比如第一级特征中的表-底温度差(DT, ℃)和第二级特征中的磷酸盐(P-d1020)、铵盐(N4-d020、N4-d010、N4-d1020)、硝酸盐(N3-d1020)等。初级特征和衍生非生物因子重要性表现出两极分化现象：部分初级非生物因子重要性较高，如底层 2 m 处温度(BT)、表层 2 m 处温度(ST)，部分衍生的衍生特征也具有较高的重要性分值，比如第一级特征中的表-底温度差(DT)和第二级特征中的磷酸盐(P-d1020)、铵盐(N4-d020、N4-d010、N4-d1020)、硝酸盐(N3-d1020)等。而低重要性特征的分值往往接近 0，且数量较多。

这种重要性分值的两极分化现象表明对初级特征和衍生特征的相关性研究是有意义的，各因素间有显著的重要性差异。

3.3　重要环境因子对底表层间分布差异的影响

温度是常见的环境因素之一，包括底层 2 m 处温度(BT)、表-底温度差(DT)以及表层 2 m 处温度(ST), 3 项特征总贡献率达到了 41.22%，说明温度因素是影响底表渔业生物声学密度差异的最关键因素。温度因素中不仅包括初级的表层和底层的温度，还包括了衍生的衍生特征——层间温差。水温对海洋生物资源行为及分布的影响是显著的，如海洋生物(水母)对温度、温差的趋向行为[32]，温度对鱼类行为的影响[33]，间接影响如温度对鱼类寄生虫的影响[34]以及对群落结构的影响[35]。

适宜的温度和光照是饵料生物大量繁殖的基

本条件, 饵料生物的密集程度也影响着较高级生态位生物的群聚程度, 但不同的鱼类以及不同的年龄段对温度差异的敏感度和反应程度并不相同[36]。

水深(WD)和纬度(LAT)是对表层和底层均有较大影响的最重要的地理信息特征, 水深可能在部分特殊地形环境中(如上升流区域、底质变化)具有重要意义。而且, 水深的重要性也与栖息环境中的种类组成有关, 尤其是对陆架区底栖和近底栖海洋生物。有研究表明, 南海北部 200 m 水深范围内近海海域各类海洋生物占比随水深的增加而变化, 甲壳类在 10~20 m 浅水区最高, 头足类占比在 40~100 m 水深较高, 鱼类占比在 100~200 m 水深较高[37]。

近岸底层鱼类对环境因素变化(如温度、盐度等)的适应范围一般较广, 其分布的区域相对稳定, 无明显的季节变化; 而表层鱼类对温盐的适应范围较窄, 具有较明显的季节变化[38]。研究中的负值区域多在浅水区域, 更适宜近表面海洋生物栖息, 正值区域表示偏重近底层鱼类适宜区域。

另外, 研究中还发现, 受人类活动影响较直接的磷酸盐(P-20 m、P-d1020)、叶绿素(CHL)、铵盐(N4-0 m、N4-d020、N4-d010、N4-d1020、N4-10 m)、硝酸盐(N3-10 m、N3-d1020、N3-0 m、N3-20 m)、亚硝酸盐(N2-10 m)等都属于第二级相关非生物因子, 重要性总贡献率约为 28%, 相反的, 人类生活中接触较少的硅酸盐表现出极低的重要性, 说明海洋生物资源的底表结构组成在一定程度上受到了人为因子(anthropogenic factors)的影响, 但其影响程度暂时无法评定。

不同的时间尺度可能会表现出不同的重要性, 但由于单航次调查时间约 35 d, 因此研究的最小时间尺度为单季节, 即通过单个航次的调查研究来分析当前季节的特征规律。

4 结论

南海北部渔业资源的负值区域(表层多, 底层少)分布于海南岛周边, 其余区域的底部资源密度均高于表层资源密度, 其分布的 41 种潜在影响因素中第一级因子以温度(包括底层 2 m 处温度、表-底温度差以及表层 2 m 处温度)、水深为主, 其中 3 种温度因素占 41.22%; 第二级特征以磷酸盐(P-20 m、P-d1020)等水质环境因素为主。换言之, 自然条件中的温度因素是影响南海北部渔业资源底表差异的最主要相关因素, 水深因素次之, 而人类活动也可能通过调节、改变磷酸盐(P-20 m、P-d1020)、叶绿素(CHL)等因素的浓度而参与了底层-表层密度差异的分布。

参考文献:

[1] Chen Z Z, Qiu Y S, Jia X P, et al. Using an ecosystem modeling approach to explore possible ecosystem impacts of fishing in the Beibu Gulf, northern South China Sea[J]. Ecosystems, 2008, 11(8): 1318-1334.

[2] Chen Z Z, Qiu Y S, Xu S N, et al. Evolution of biological characteristics of *Saurida undosquamis* (Richardson) in the Beibu Gulf, South China Sea[J]. Journal of Fishery Sciences of China, 2012, 19(2): 321-328. [陈作志, 邱永松, 徐姗楠, 等. 北部湾花斑蛇鲻生物学特征的演化[J]. 中国水产科学, 2012, 19(2): 321-328.]

[3] Qiu Y S, Lin Z J, Wang Y Z. Responses of fish production to fishing and climate variability in the northern South China Sea[J]. Progress in Oceanography, 2010, 85(3-4): 197-212.

[4] D'Elia M, Patti B, Bonanno A, et al. Analysis of backscatter properties and application of classification procedures for the identification of small pelagic fish species in the Central Mediterranean[J]. Fisheries Research, 2014, 149: 33-42.

[5] Chen G B, Zhang J, Yu J, et al. Hydroacoustic scattering characteristics and biomass assessment of the purpleback flying squid [*Sthenoteuthis oualaniensis*, (Lesson, 1830)] from the deepwater area of the South China Sea[J]. Journal of Applied Ichthyology, 2013, 29(6): 1447-1452.

[6] Gimona A, Fernandes P G. A conditional simulation of acoustic survey data: Advantages and potential pitfalls[J]. Aquatic Living Resources, 2003, 16(3): 123-129.

[7] Laidre K L, Heide-Jørgensen M P, Heagerty P, et al. Spatial associations between large baleen whales and their prey in West Greenland[J]. Marine Ecology Progress Series, 2010, 402: 269-284.

[8] Simmonds J, MacLennan D N. Fisheries Acoustics: Theory and Practice[M]. New York: John Wiley & Sons, 2008.

[9] Takahashi M, McCormick M I, Munday P L, et al. Influence of seasonal and latitudinal temperature variation on early life-history traits of a coral reef fish[J]. Marine and Freshwater Research, 2012, 63(10): 856-864.

[10] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining. New York: ACM, 2016: 785-794.
[11] Svetnik V, Liaw A, Tong C, et al. Random forest: A classification and regression tool for compound classification and QSAR modeling[J]. Journal of Chemical Information & Computer Sciences, 2003, 43(6): 1947-1958.
[12] Ren X, Guo H, Li S, et al. A novel image classification method with CNN-XGBoost model[C]// Proceedings of the 16th International Workshop on Digital Forensics and Watermarking. Cham: Springer, 2017: 378-390.
[13] Bosch A, Zisserman A, Munoz X. Image classification using random forests and ferns[C]//Proceedings of the 2007 IEEE 11th International Conference on Computer Vision. Piscataway: IEEE, 2007: 1-8.
[14] Marmion M, Parviainen M, Luoto M, et al. Evaluation of consensus methods in predictive species distribution modelling[J]. Diversity & Distributions, 2009, 15(1): 59-69.
[15] Lu M, Sadiq S, Feaster D J, et al. Estimating individual treatment effect in observational data using random forest methods[J]. Journal of Computational and Graphical Statistics, 2018, 27(1): 209-219.
[16] Fitriah N, Wijaya S K, Fanany M I, et al. EEG channels reduction using PCA to increase XGBoost's accuracy for stroke detection[J]. AIP Conference Proceedings, 2017, 1862(1): 030128.
[17] Chen W B, Fu K, Zuo J W, et al. Radar emitter classification for large data set based on weighted-xgboost[J]. IET Radar, Sonar & Navigation, 2017, 11(8): 1203-1207.
[18] Torlay L, Perrone-Bertolotti M, Thomas E, et al. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Informatics, 2017, 4(3): 159-169.
[19] Menze B H, Kelm B M, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data[J]. BMC Bioinformatics, 2009, 10(1): Article No.213.
[20] Stijven S, Minnebo W, Vladislavleva K. Separating the wheat from the chaff: On feature selection and feature importance in regression random forests and symbolic regression[C]//Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation. New York: ACM, 2011: 623-630.
[21] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python[J]. Journal of Machine Learning Research, 2013, 12(10): 2825-2830.
[22] Manson S M, Burrough P A, McDonnell R A. Principles of geographical information systems: Spatial information systems and geostatistics[J]. Geofisica Internacional, 1988, 42(4): 357-358.
[23] Pereira P, Oliva M, Misiune I. Spatial interpolation of precipitation indexes in Sierra Nevada (Spain): Comparing the performance of some interpolation methods[J]. Theoretical and Applied Climatology, 2016, 126(3-4): 683-698.
[24] Sales M H, Souza C M, Kyriakidis P C, et al. Improving spatial distribution estimation of forest biomass with geostatistics: A case study for Rondônia, Brazil[J]. Ecological Modelling, 2007, 205(1-2): 221-230.
[25] Webster R, Oliver M A. Geostatistics for Environmental Scientists[M]. Chichester: John Wiley & Sons Ltd., 2007.
[26] Freeman E A, Moisen G G. Evaluating Kriging as a tool to improve moderate resolution maps of forest biomass[J]. Environmental Monitoring and Assessment, 2007, 128(1-3): 395-410.
[27] Sun M S, Chen Z Z, Cai Y C, et al. Application of a spatial interpolation method for the assessment of fishery resources in the Beibu Gulf[J]. Journal of Fishery Sciences of China, 2017, 24(4): 853-861. [孙铭帅, 陈作志, 蔡研聪, 等. 空间插值法在北部湾渔业生物声学密度评估中的应用[J]. 中国水产科学, 2017, 24(4): 853-861.]
[28] Huang M L, Hung Y H, Lee W M, et al. SVM-RFE based feature selection and *Taguchi parameters* optimization for multiclass SVM classifier[J]. The Scientific World Journal, 2014: 795624.
[29] Cheng Q, Varshney P K, Arora M K. Logistic regression for feature selection and soft classification of remote sensing data[J]. IEEE Geoscience & Remote Sensing Letters, 2006, 3(4): 491-494.
[30] Pal M. Multinomial logistic regression-based feature selection for hyperspectral data[J]. International Journal of Applied Earth Observation & Geoinformation, 2012, 14(1): 214-220.
[31] Talenti L, Luck M, Yartseva A, et al. L1 logistic regression as a feature selection step for training stable classification trees for the prediction of severity criteria in imported malaria[J]. Journal of Bone & Mineral Research, 2015, 24(6): 1055-1065.
[32] Purcell J E, Uye S, Lo W T. Anthropogenic causes of jellyfish blooms and their direct consequences for humans: A review[J]. Marine Ecology Progress Series, 2007, 350: 153-174.
[33] Stegmann P M, Yoder J A. Variability of sea-surface temperature in the South Atlantic bight as observed from satellite: Implications for offshore-spawning fish[J]. Continental Shelf Research, 1996, 16(7): 843-861.
[34] Franke F, Armitage S A O, Kutzer M A M, et al. Environmental temperature variation influences fitness trade-offs and tolerance in a fish-tapeworm association[J]. Parasites & Vectors, 2017, 10(1): Article No.252.
[35] Riegl B. Effects of the 1996 and 1998 positive sea-surface temperature anomalies on corals, coral diseases and fish in the Arabian Gulf (Dubai, UAE)[J]. Marine Biology, 2002, 140(1): 29-40.
[36] Elliott J M, Hurley M A. Variation in the temperature preference and growth rate of individual fish reconciles differences between two growth models[J]. Freshwater Biology, 2003, 48(10): 1793-1798.
[37] Liu W D, Lin Z J, Jiang Y E, et al. Spatial distribution of demersal fishery resources in the continental shelf of the northern South China Sea[J]. Journal of Tropical Oceano-

graphy, 2011, 30(5): 95-103. [刘维达, 林昭进, 江艳娥, 等. 南海北部陆架区底层渔业资源的空间分布特征[J]. 热带海洋学报, 2011, 30(5): 95-103.]

[38] Qiu Y S. The regional changes of fish community on the northern continental shelf of South China Sea[J]. Journal of Fisheries of China, 1988, 12(4): 303-313. [邱永松. 南海北部大陆架鱼类群落的区域性变化[J]. 水产学报, 1988, 12(4): 303-313.]

Difference in the fishery resource density between the bottom and surface layers and an analysis of multiple types of related factor importance in the Northern South China Sea

SUN Mingshuai[1], CAI Yancong[1], ZHANG Kui[1], XU Youwei[1], YANG Yutao[1], CHEN Zuozhi[1, 2]

1. South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences; Key Laboratory for Sustainable Utilization of Open-Sea Fishery, Ministry of Agriculture and Rural Affairs, Guangzhou 510300, China
2. Guangdong Provincial Key Laboratory of Fishery Ecology and Environment, Guangzhou 510300, China

Abstract: This study aims to analyze the differences in acoustic density of fishery resources between surface and bottom layers (surface mixed layer and bottom cold water layer) in the northern South China Sea and to explore the relationship between these differences and 41 abiotic factors. This research provides a scientific basis for the effective management and conservation of fishery resources in the northern South China Sea. The northern offshore area of the South China Sea is a crucial traditional fishing ground and an important spawning and feeding ground for marine fish. In recent years, fishery resources in this region have shown significant declines in age, size, and quality, attracting significant attention from both academia and fishery management authorities. Fishery acoustic methods were employed, using a Simrad EY60 split-beam scientific echosounder to collect acoustic data in the northern South China Sea. Acoustic data were analyzed using the Echoview fishery acoustic data processing system to calculate the acoustic density (NASC) of surface and bottom layers. Extreme Gradient Boosting (XGBoost) and Random Forest algorithms were utilized to model the influence of 41 abiotic factors on the differences in acoustic density and to assess the importance of these factors. Results indicated that the bottom layer had significantly higher acoustic density than the surface layer, with mean values of 106.00 m²/nmi² and 43.39 m²/nmi², respectively. Both XGBoost and Random Forest models performed similarly, with temperature factors (bottom 2 m temperature, surface-bottom temperature difference, and surface 2 m temperature) and water depth identified as the most critical factors affecting acoustic density differences. The negative value region, where surface density exceeds bottom density, is primarily distributed around Hainan Island. The study concluded that temperature and water depth are the primary factors influencing the distribution differences of fishery resources, while human activities may also contribute by altering the concentrations of factors such as phosphate and chlorophyll. Additionally, the discussion highlights the implications of these findings for fisheries management, suggesting that targeted measures to monitor and regulate temperature and nutrient levels could significantly improve resource sustainability. The analysis underscores the importance of incorporating advanced machine learning algorithms in marine resource assessment to enhance the accuracy and reliability of environmental impact evaluations. These findings provide vital scientific insights for the management and conservation of fishery resources in the northern South China Sea, offering a comprehensive understanding of the environmental factors that drive spatial distribution patterns in marine ecosystems. This research thus lays a foundation for future studies aiming to mitigate the impacts of climate change and human activities on marine biodiversity and resource availability.

Key words: abiotic factors, XGBoost, Random Forest, fishery acoustics; the northem South China Sea

Corresponding author: CHEN Zuozhi. E-mail: chenzuozhi@scsfri.com