中国水产科学 2018 年 9 月, 25(5): 1082−1090
Journal of Fishery Sciences of China

研究论文

DOI: 10.3724/SP.J.1118.2018.17388

基于 CART 算法的长江口鱼种丰富度预测

戴黎斌[1, 2, 3], 陈锦辉[4], 田思泉[1, 2, 3, 5], 高春霞[1, 2, 3, 5], 王家启[1, 2, 3], 杜晓雪[1, 2, 3], 王学昉[1, 2, 3, 5]

1. 上海海洋大学 海洋科学学院, 上海 201306;
2. 中国远洋渔业数据中心, 上海 201306;
3. 大洋渔业资源可持续开发教育部重点实验室, 上海 201306;
4. 上海市长江口中华鲟自然保护区管理处, 上海 200092;
5. 国家远洋渔业工程技术研究中心, 上海 201306

摘要: 长江口是西太平洋最大的河口生态系统和典型的生态群落交错区, 预测鱼类生物多样性对评价其生态系统有着重要的科学指示意义。结合 2012—2013 年长江口 15 个站点的渔业资源和环境调查数据, 利用分类与回归树(CART)算法中的回归树算法, 构建了长江口鱼种丰富度预测模型。基于 1-SE 准则, 最优决策树的复杂性参数设置为 0.067, 结果表明, 盐度、溶解氧和季节是影响长江口鱼类生物多样性的主要因子。此外, 使用 2014 年的观测数据对回归树模型预测的长江口鱼种丰富度予以验证, 均方根误差(RMSE)、平均相对误差(ARE)和平均绝对误差(AAE)值的统计结果显示, 回归树模型在春、夏季的预测效能优于秋、冬季, 模型总体上呈现出了较好的预测能力, 表明利用 CART 算法对长江口鱼种丰富度进行预测是可行的。

关键词: 分类与回归树; 长江口; 鱼种丰富度; 预测
中图分类号: S931 **文献标志码:** A **文章编号:** 1005−8737−(2018)05−1082−09

河口是连接鱼类所在淡水和海洋生态系统的过渡区域, 具有环境梯度变化大、生产力高的特点[1-2], 其生境对栖息在其中的鱼类有着重要的生态意义[3]。如河口定居性鱼类的全部生活史过程都在河口完成; 部分海洋鱼类则将河口作为育肥场; 而河海洄游性鱼类又视河口为洄游通道[4]。然而, 河口很多重要的理化特性并不具有连续性和稳定性[5], 且极易受到人类活动的影响, 由此产生的过度捕捞、生境退化、水体富营养化等问题日益严重[4, 6]。生物多样性被用以表示某一区域内物种的多元化程度, 通常以物种丰富度(即调查区域内所观察到的物种数量)来进行量化分析。长江口是西太平洋地区最大的河口生态系统和典型的生态群落交错区, 同时也是许多鱼类重要的洄游通道[5, 7]。因此, 分析长江口生境的鱼类生物多样性对评估环境变化和人类扰动对河口生态系统产生的综合影响有着显著的科学指示意义。

目前, 关于长江口鱼类多样性的研究主要集中于鱼类群落的结构[5]、空间分布特征[8]、多样性变动[3, 7]等, 尚未有直接以鱼种丰富度为衡量指标的生物多样性研究报道。为此, 本研究基于分类与回归树(classification and regression trees, CART)算法, 结合 2012—2014 年长江口渔业资源和环境监测数据构建鱼种丰富度与环境、时空因子之间的决策树预测模型, 旨在探究长江口鱼种丰富度现状, 为长江口渔业资源的科学管理提供理论依据。

收稿日期: 2017-10-21; **修订日期:** 2018-01-01.
基金项目: 长江口中华鲟增殖放流跟踪监测和效果评估项目(170062); 上海市科委地方能力建设项目(18050502000).
作者简介: 戴黎斌(1994–), 男, 硕士研究生, 研究方向为渔业生态学. E-mail: 644318716@qq.com
通信作者: 陈锦辉, 副研究员. E-mail: 1114260882@qq.com

1　材料与方法

1.1　数据来源

数据源于 2012—2014 年长江口底拖网渔业资源调查，调查时间为每年的 2 月(冬季)、5 月(春季)、8 月(夏季)和 11 月(冬季)，调查方式为定点采样，共设 15 个计划站点(图 1)，调查区域为长江口中华鲟自然保护区及附近水域，详细站点位置见表 1。调查船为沪崇渔 1511 号，渔具为双囊底拖网，网口宽 6 m，网高 2 m，网纲长 6 m，囊网网目 20 mm。现场调查时，根据 GPS 定位，当调查船到达站位后，放下拖网，以 2 km/h 左右的航速拖曳 30 min，随即收网并整理两个囊袋中的渔获物。当渔获物数量较少时，记录全部渔获物的种类、数量、重量、体长等信息，渔获物数量较多时随机抽取一定比例的渔获物进行统计，最后换算成全部渔获物的数量。此外，使用 Hydrolab 水质分析仪等测量仪器同步测定各站点环境数据并予以记录。受调查时现场天气、海况等因素影响，多个航次未能完成全部站点的调查，各航次实际调查站点数量如表 2 所示。

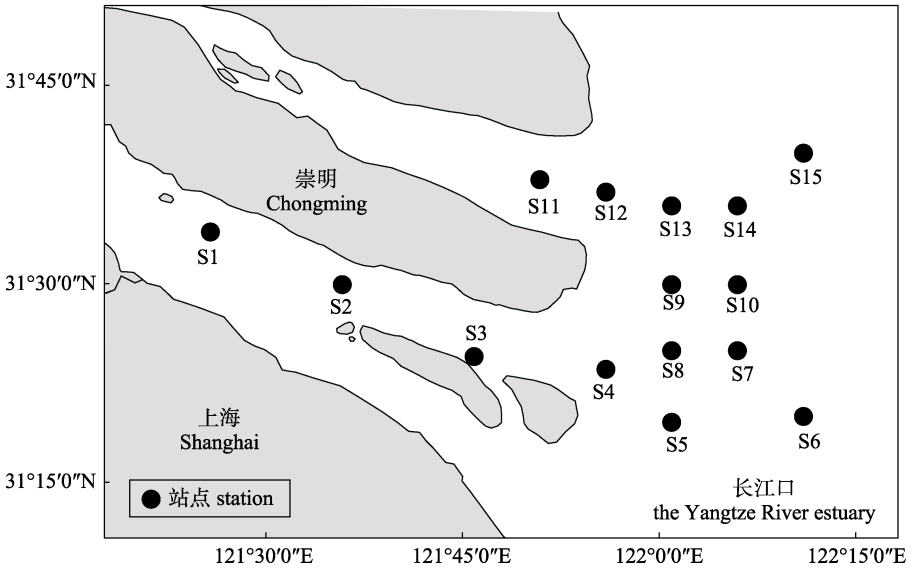

图 1　长江口渔业资源调查站点分布

Fig. 1　Spatial distribution of survey stations in the Yangtze River estuary

表 1　长江口资源调查站点位置

Tab. 1　Locations of survey stations in the Yangtze River estuary

站点 station	经度/(°) longitude	纬度/(°) latitude
S1	121.42	31.57
S2	121.58	31.50
S3	121.75	31.41
S4	121.92	31.39
S5	122.00	31.33
S6	122.17	31.33
S7	122.08	31.42
S8	122.00	31.42
S9	122.00	31.50
S10	122.08	31.50
S11	121.83	31.63
S12	121.92	31.62
S13	122.00	31.60
S14	122.08	31.60
S15	122.17	31.67

表 2　2012—2014 年长江口渔业资源调查实际调查站点数量

Tab. 2　Actual numbers of survey stations in the Yangtze River estuary in 2012–2014

年份 year	月份 month	站点数量 number of stations	未调查站点 station absent from survey
2012	2	13	S6, S7
	5	11	S2, S3, S6, S7
	8	12	S2, S6, S7
	11	11	S4, S6, S7, S8
2013	2	10	S1, S2, S3, S5, S7
	5	12	S2, S4, S7
	8	15	
	11	14	S8
2014	2	9	S1, S2, S3, S5, S9, S14
	5	11	S2, S5, S9, S14
	8	11	S2, S5, S9, S14
	11	10	S2, S5, S8, S9, S14

1.2 CART 算法的回归决策树

CART 算法是由 Breiman 等[9]提出的一种基于决策树的机器学习方法，根据响应变量类别和分裂准则(splitting criteria)的不同可分为分类决策树与回归决策树两种，考虑到本研究中的响应变量为鱼种丰富度，故使用回归树算法来构建模型。回归树的构建共包括两个过程[10-12]：一是树的种植，即采用二元递归分解法不断选择最适自变量将数据集分解到不断增加的同质子集当中，直至无法再继续分解；二是树的修剪，即确定树的大小(复杂性)，以减少树的分裂次数或叶节点数量，防止树过拟合。

1.2.1 回归决策树的种植 模型的响应变量为鱼种丰富度，即某一月份(季度)各站点所观测到的鱼种数量，解释变量为环境和时空因子，包括月份、经度、纬度、水深、水温、盐度、溶解氧、pH、叶绿素 a 和化学需氧量(chemical oxygen demand, COD)，共 10 个解释变量。在种植决策树之前，使用斯皮尔曼相关性分析对解释变量进行相关性检验，相关性大于 0.8 的变量将不用于模型的构建，以避免自变量间出现多重共线性[13]。相关性检验后，树在种植过程中通过不断选择最适因子来对树中节点进行二分，以探讨多因子与鱼类生物多样性之间的关系。

1.2.2 回归决策树的修剪 采用 1-SE 规则[9, 14]对树进行修剪，即与具有最小交叉验证相对误差(cross-validated relative error)的树的差异小于 1 个标准误且分裂次数更少的树被视为最优树。

1.3 模型的预测与验证

2012—2013 年的调查数据被用来进行回归树的构建，将经过修剪后得到的最优树作为预测模型。之后，使用预测模型对 2014 年长江口各站点的鱼种丰富度进行预测，并使用 2014 年的观测数据与之比较验证，验证结果基于平均相对误差(average relative error, ARE)、平均绝对误差(average absolute error, AAE)和均方误差根(root mean square error, RMSE)的大小，上述指标越接近于 0，表明预测值与观测值之间的差异越小[15]，3 个统计指数的计算公式如下：

$$\mathrm{ARE}=\frac{\sum_{i=1}^{n}(P_i-O_i)}{n} \tag{1}$$

$$\mathrm{AAE}=\frac{\sum_{i=1}^{n}\left|P_i-O_i\right|}{n} \tag{2}$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(P_i-O_i)^2}{n}} \tag{3}$$

式中，n 为样本数量；P_i 为第 i 个模型预测值；O_i 为第 i 个调查观测值。

2 结果与分析

2.1 回归决策树的训练结果

2012—2014 年各航次环境和时空变量统计摘要信息如表 3 所示。使用 2012—2013 年调查数据作为训练集进行回归树模型的建立(图 2)，树中各节点上方的变量即为决策树二分时选择的最优变量，判断语句为数据集的分类依据，判断结果为“是”的子集划分到左边的节点，判断结果为“否”的子集划分到右边的节点，以此类推，直至决策树不再分解。在所有被用作解释变量的因子中，只有水深没有被植入决策树中，这表明水深无法对长江口鱼类丰富度进行有效的量化划分。此外，决策树中多个变量被重复利用作为节点的分裂变量，这表明树的尺寸已偏大，分解次数过多，需要进行适当的修剪。

2.2 回归决策树的修剪结果

在回归决策树的种植过程中，决策树将原始数据集分为训练集和测试集两部分，进行交叉验证，以得到不同尺寸(节点数量)的树的交叉验证相对误差。根据 1-SE 准则，节点数量为 4 个的决策树可被视为最优树(图 3)，故对初始回归决策树进行相应的修剪，即将复杂性参数调整为 0.067。最优树结构图如图 4 所示。修剪后，树的结构得到显著优化，原始数据依据盐度、溶解氧和月份 3 个因子共分成 4 个子集，样本量从左到右依次减少。箱线图显示，鱼种丰富度从左到右逐渐递增，最高鱼种丰富度发生在盐度大于 5.25 且溶解氧大于 7.38 mg/L 的站点，其平均值为 10.75，最低丰富度出现在盐度小于5.25的43个样本中，其平均值为2.40。

表 3　各航次变量均值

Tab. 3　Summary of statistical information for variables in each trip

变量 variable	2012				2013				2014			
	2 月 Feb	5 月 May	8 月 Aug	11 月 Nov	2 月 Feb	5 月 May	8 月 Aug	11 月 Nov	2 月 Feb	5 月 May	8 月 Aug	11 月 Nov
丰富度 richness	2.62 (1.94)	4.45 (3.14)	6.42 (5.33)	4.73 (3.35)	4.20 (1.87)	3.17 (2.04)	4.93 (3.24)	6.36 (3.08)	3.22 (1.72)	2.73 (1.62)	4.09 (1.64)	3.00 (2.45)
经度/(°) longitude	121.90 (0.21)	121.95 (0.20)	121.93 (0.20)	121.89 (0.23)	122.02 (0.11)	121.95 (0.21)	121.93 (0.21)	121.93 (0.22)	122.02 (0.12)	121.94 (0.22)	121.94 (0.22)	121.93 (0.23)
纬度/(°) latitude	31.52 (0.11)	31.53 (0.11)	31.52 (0.11)	31.54 (0.10)	31.53 (0.11)	31.51 (0.12)	31.50 (0.11)	31.50 (0.11)	31.51 (0.12)	31.50 (0.12)	31.50 (0.12)	31.51 (0.12)
水深/m depth	5.98 (1.75)	6.09 (3.86)	5.68 (4.41)	8.86 (6.31)	5.98 (2.03)	6.57 (2.28)	6.15 (3.35)	5.97 (2.54)	5.00 (2.09)	6.27 (2.72)	5.60 (2.25)	6.60 (2.25)
水温/℃ temperature	6.27 (0.43)	22.30 (0.89)	27.28 (6.71)	13.75 (0.93)	11.03 (0.77)	21.69 (1.37)	27.65 (1.55)	10.61 (1.05)	9.23 (0.60)	19.67 (1.00)	29.52 (0.39)	16.28 (1.47)
盐度 salinity	14.24 (13.77)	6.04 (6.71)	8.53 (9.31)	14.19 (11.33)	17.35 (11.91)	8.83 (8.03)	11.25 (8.07)	12.43 (12.78)	8.12 (9.11)	6.43 (8.38)	4.92 (5.61)	8.85 (9.40)
溶解氧/(mg/L) dissolved oxygen	12.32 (0.39)	8.45 (0.26)	16.23 (11.52)	10.24 (0.44)	12.15 (0.20)	9.10 (0.13)	7.53 (0.26)	12.25 (0.30)	11.40 (0.11)	8.81 (0.07)	7.55 (0.56)	9.98 (0.45)
pH	7.72 (0.43)	7.81 (0.74)	8.04 (0.08)	8.14 (0.05)	7.87 (0.23)	8.16 (0.05)	8.03 (0.32)	7.84 (0.20)	8.26 (0.17)	8.21 (0.12)	8.29 (0.11)	8.10 (0.15)
叶绿素 a/(mg/L) chlorophyll a	2.23 (0.90)	2.49 (1.51)	7.87 (2.80)	1.73 (0.41)	0.93 (0.25)	1.46 (0.27)	1.13 (0.21)	1.04 (0.22)	0.51 (0.27)	0.28 (0.14)	0.48 (0.27)	0.53 (0.16)
化学需氧量/(mg/L) chemical oxygen demand	1.07 (0.28)	1.19 (0.56)	1.32 (0.30)	0.91 (0.38)	0.47 (0.27)	0.73 (0.30)	0.84 (0.31)	0.61 (0.33)	1.23 (0.32)	1.32 (0.45)	0.91 (0.32)	0.83 (0.39)

注: 括号中数字表示标准差.

Note: Figures in the brackets are standard deviations.

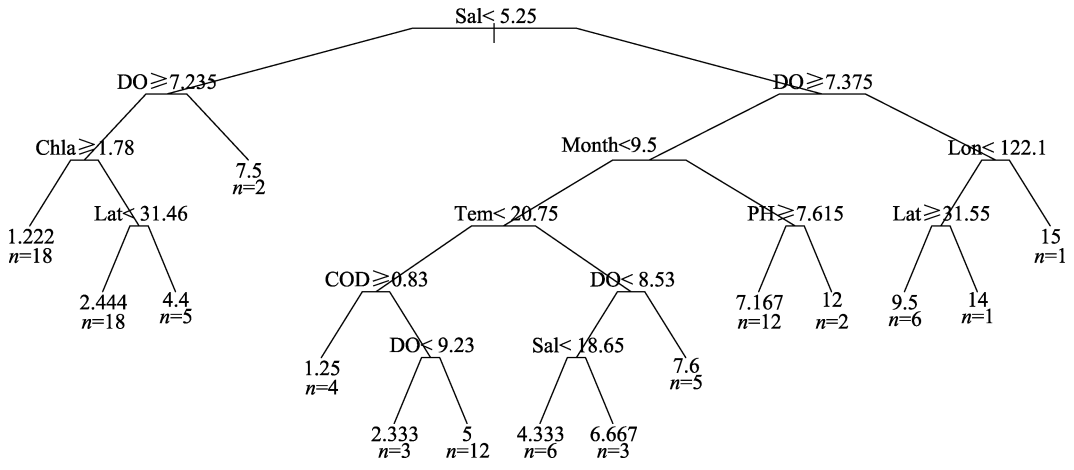

图 2　长江口鱼种丰富度回归决策树

Sal: 盐度; DO: 溶解氧; Chla: 叶绿素 a; Month: 月份; Lon: 经度; Lat: 纬度; Tem: 水温; COD: 化学需氧量; *n*: 样本量.

Fig. 2　Regression decision tree of fish species richness in the Yangtze River estuary

Sal: salinity; DO: dissolved oxygen; Chla: chlorophyll a; Month: month; Lon: longitude; Lat: latitude; Tem: water temperature; COD: chemical oxygen demand; *n*: sample size.

2.3　模型的预测与验证结果

2014 年长江口鱼种观测丰富度分布如图 5 所示。除冬季外, 其他三季都显示出北支河口鱼种丰富度略高于南支河口的趋势, 但季节上的丰富度变化尺度较小, 无明显的时间尺度变化规律。利用最优回归树模型, 结合 2014 年调查数据对 2014 年各站点的鱼种丰富度进行了预测, 预测结果具体数值如表 4 所示。时间上, 夏、秋季长江口鱼种丰富度显著高于春、冬季; 空间上, 秋季北支水域的鱼类多样性明显高于南支北港水域,

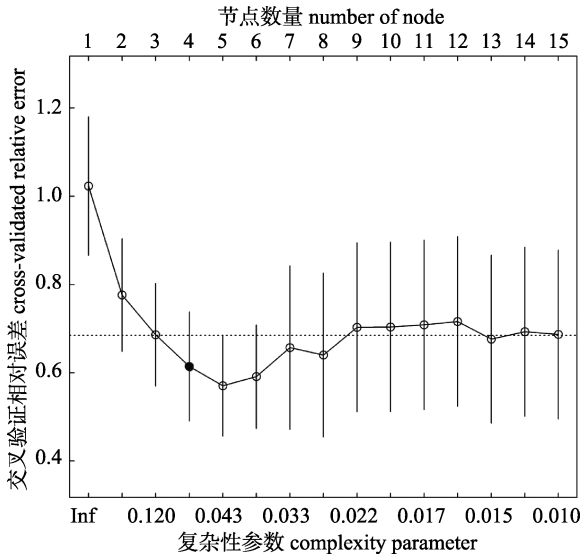

图 3 交叉验证相对误差随树中节点数量增加的变化情况

图中水平虚线为交叉验证相对误差最小值加 1 倍标准误，黑点表示满足 1-SE 准则的决策树大小的位置.

Fig. 3 Changes of cross-validated relative error along with the increase of nodes in decision tree

The horizontal dashed line represents the minimum cross-validated relative error adding a standard error, and the black dot represents the tree size which fitted 1-SE rule.

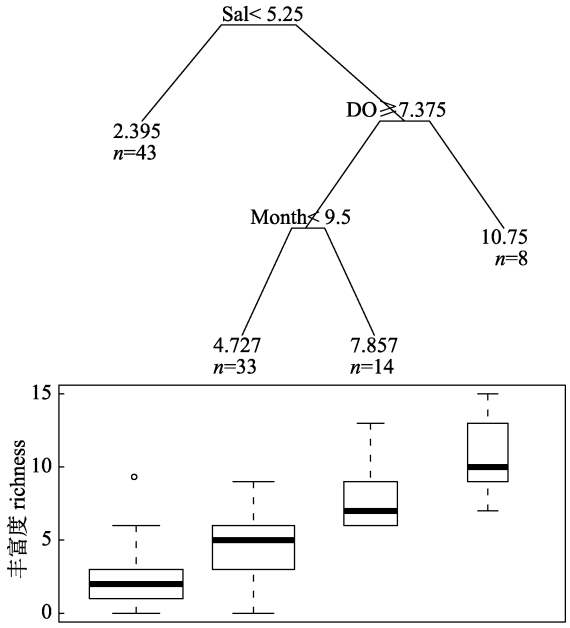

图 4 修剪后的回归决策树及各终端节点对应的鱼种丰富度大小

Sal: 盐度; DO: 溶解氧; Month: 月份; *n*: 样本量.

Fig. 4 Regression trees after pruning and the fish species richness corresponding to all the nodes

Sal: salinity; DO: dissolved oxygen; *n*: sample size.

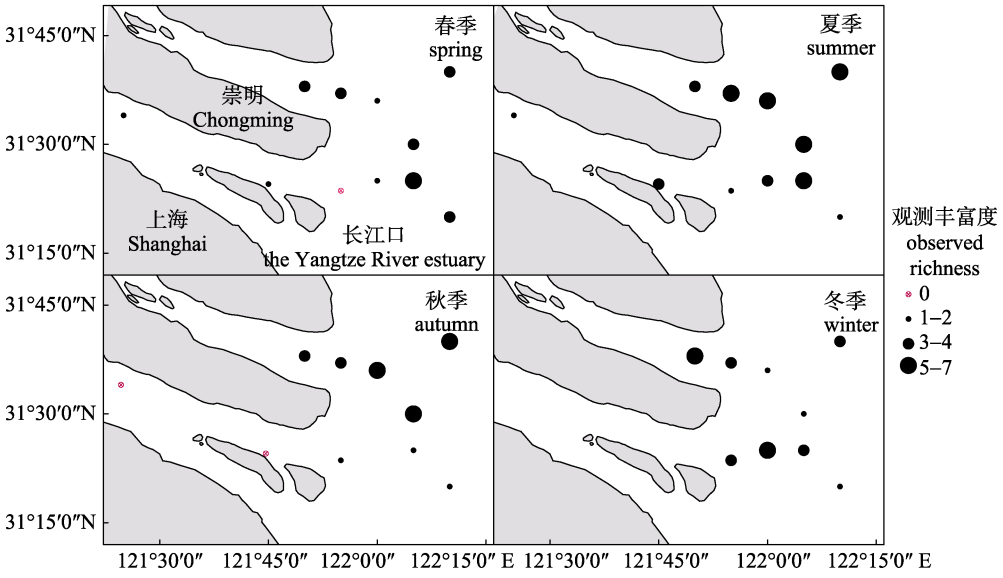

图 5 2014 年长江口鱼种丰富度观测值分布

Fig. 5 Distribution of fish species richness observed in the Yangtze River estuary in 2014

春、冬季之间则不存在明显的丰富度差异(图 6)。

RMSE, ARE 和 AAE 被用于验证回归树模型预测结果与实际观测数据之间的差异大小(图 7)。结果表明，回归树模型在夏季具有最佳的预测效果，其 AAE 值接近于 0，秋季预测结果最差，3 个指数统计量均为 4 个季度中的最大值(表 5)。总体上，春、夏季预测效果要优于秋、冬季，这也与图 5、图 6 之间的比较结果基本保持一致。

表 4　2014 年长江口鱼种丰富度观测值和预测值

Tab. 4　Observed and predicted fish species richness in the Yangtze River estuary in 2014

站点 station	观测丰富度 observed richness				预测丰富度 predicted richness			
	2 月 February	5 月 May	8 月 August	11 月 November	2 月 February	5 月 May	8 月 August	11 月 November
S1	–	2.00	2.00	0.00	–	2.15	7.50	2.15
S3	–	1.00	3.00	0.00	–	2.15	7.50	2.15
S4	4.00	0.00	2.00	1.00	2.15	2.15	2.15	2.15
S6	1.00	3.00	2.00	2.00	2.15	2.15	7.50	2.15
S7	3.00	5.00	6.00	2.00	2.15	2.15	2.15	2.15
S8	5.00	1.00	4.00	–	2.15	2.15	7.50	–
S10	2.00	4.00	6.00	7.00	3.79	3.79	6.00	7.17
S11	6.00	4.00	4.00	3.00	3.79	3.79	6.00	7.17
S12	4.00	4.00	5.00	4.00	3.79	3.79	6.00	7.17
S13	1.00	2.00	5.00	6.00	3.79	2.15	6.00	7.17
S15	3.00	4.00	6.00	5.00	2.15	3.79	6.00	7.17

注: “–”表示站点在该航次内未进行调查.

Note: “–” means the station was absent from survey in the corresponding trip.

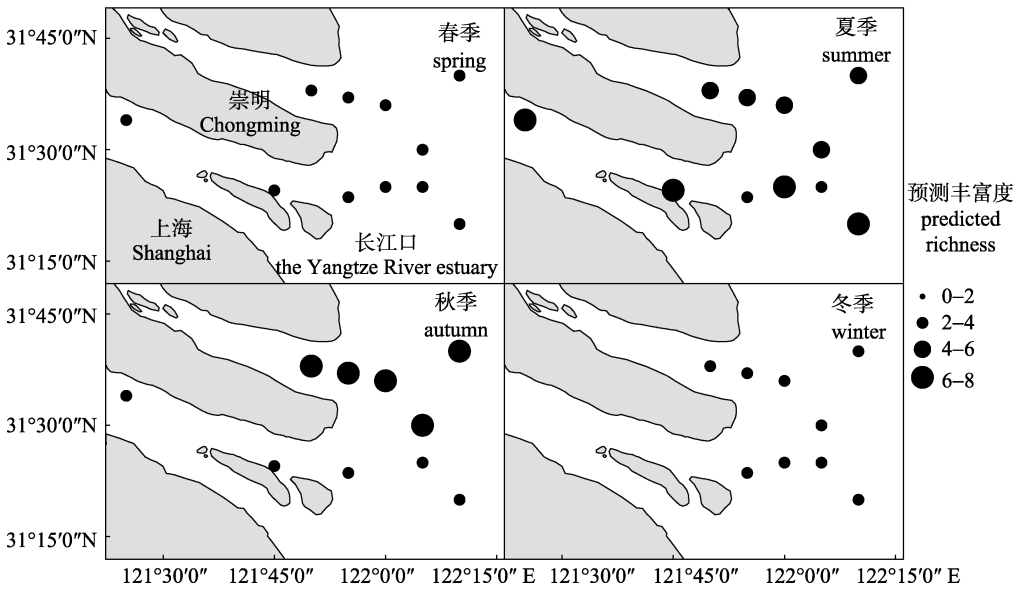

图 6　基于回归树算法的 2014 年长江口鱼种丰富度预测值分布

Fig. 6　Distribution of predicted fish species richness based on regression tree model in the Yangtze River estuary in 2014

3　讨论

根据最优回归树的节点结构, 盐度、溶解氧和月份(季节)是影响长江口鱼种丰富度的主要因子。长江口作为众多洄游性鱼类的洄游通道以及重要的鱼类育肥场、产卵场和饵料场, 其饵料丰富, 初级生产力发达, 环境梯度变化极快。而一些栖息于河口或进行过渡洄游鱼类的种群动态和丰富度分布通常都具有季节性特征[16], 故季度是影响长江口鱼种丰富度变化的重要时间因子。盐度梯度的变化也具有支配鱼类行为的作用, 对鱼类的渗透压和浮性鱼卵的漂流等都会产生影响[17], 长江口盐度梯度的变化会对生活在其中的鱼类的产卵、索饵、洄游等行为造成影响, 同时也会导致鱼类的群落结构和多样性发生改变, 如史赟荣等[8]发现表层和底层盐度对 2010 年春季长江口鱼类丰度和栖息密度的影响最大。一些研究[4, 18-19]分析了全球尺度下对河口鱼类生物多样性可能产生影响的因子, 并认为盐度的梯度变化就是使得河口鱼种丰富度改变最主要的环境驱动因子。此

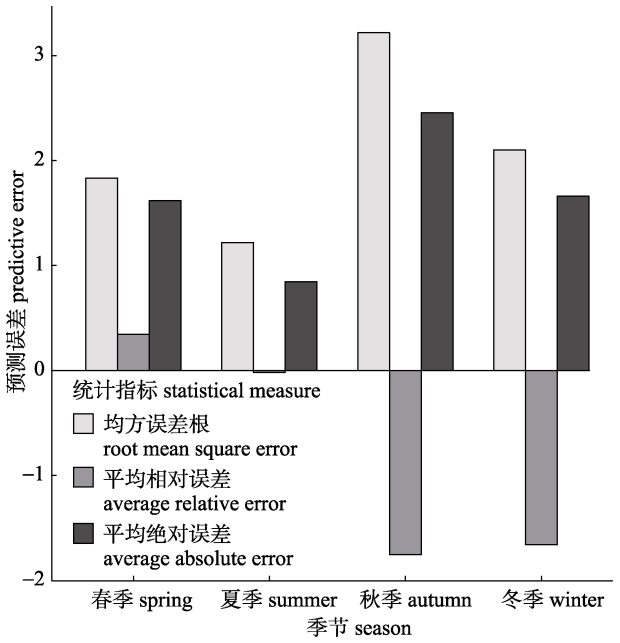

图 7 不同季节的预测误差比较

Fig. 7 Comparison of predictive error in different seasons

表 5 基于回归树算法的 2014 年长江口鱼种丰富度预测误差

Tab. 5 Predictive error for the fish species richness of the Yangtze River estuary using regression algorithm in 2014

月份 month	均方误差根 root mean square error	平均相对误差 average relative error	平均绝对误差 average absolute error
2 月 February	1.83	0.34	1.62
5 月 May	1.22	–0.02	0.85
8 月 August	3.22	–1.76	2.46
11 月 November	2.10	–1.66	1.66

外，Whitfield 等[18]运用 Remane 图解法对河口生物组成和盐度梯度之间的关系进行了评估，并认为盐度为 5.00~7.00 时河口的物种数量较小，这也与本研究的结果相符，盐度小于 5.25 的样本只具有均值为 2.40 的鱼种丰富度(图 4)。溶解氧浓度的大小作为水质量评价指标之一，对于河口的生态恢复有重要意义。有学者认为，随着河口溶解氧浓度水平的增加，鱼类的资源结构也会相应改善[20-23]，这或许能解释在本研究中有着最高鱼种丰富度的样本同时也具有较高的溶解氧浓度(图 4)。

回归树模型是一种非参数化的机器学习方法，其特点为不需要预先假设因变量与自变量之间的关系，而是通过二分递归的方式自动选择自变量，将原始数据集划分为具有同质特点的多个子集，并通过决策树的可视化来表达变量间的复杂关系[12]。同时，当变量间存在非线性关系、具有高阶的交互项或数据存在缺失等问题时，回归树模型具有优势。不仅如此，在物种分布预测[24]、河口鱼类丰富度预测[25]等研究方面，回归树相对于其他模型展现出了更好的预测能力。Pittman 等[26]认为在尝试对河口栖息地的鱼种丰富度与环境因子关系进行建模时，模型会被加入更多的非线性关系，故需要更加灵活的模型结构。尽管河口生态系统被认为是一种复杂的栖息地类型，但 CART 算法显然能够呈现出较好的预测结果[27]。

本研究使用基于 CART 算法的回归决策树对 2014 年长江口鱼类丰富度进行了预测，并与实际观测值进行了比较。结果表明，回归树模型能够合理解释多因子与鱼种丰富度之间的关系，但用以验证预测结果的统计指标之间存在季节性差异，春、夏季模型的预测值更接近实际的调查数据。笔者认为这可能与样本量大小有关，2014 年春、夏季的调查站点数量要多于秋、冬季，故用于拟合模型的训练数据更多，模型的预测效能也因此更为稳健。总体上，回归树模型仍呈现出较好的预测能力，这表明回归树模型可作为一种有效预测长江口鱼类生物多样性的方法，对长江口的生态系统的评估有着重要的科学指示意义。

参考文献:

[1] Costanza R, D'Arge R, Groot R D, et al. The value of the world's ecosystem services and natural capital[J]. Nature, 1998, 25(1): 3-15.

[2] Nicolas D, Lobry J, Lepage M, et al. Fish under influence: A macroecological analysis of relations between fish species richness and environmental gradients among European tidal estuaries[J]. Estuarine Coastal and Shelf Science, 2010, 86(1): 137-147.

[3] Shi Y R, Chao M, Quan W M, et al. Fish community diversity analyses in the Yangtze River estuary, China[J]. Journal of Fishery Sciences of China, 2012, 19(6): 1051-1059. [史赟荣, 晁敏, 全为民, 等. 长江口鱼类群落的多样性分析[J]. 中国水产科学, 2012, 19(6): 1051-1059.]

[4] Vasconcelos R P, Henriques S, França S, et al. Global patterns and predictors of fish species richness in estuaries[J]. Journal of Animal Ecology, 2015, 84(5): 1331-1341.

[5] Shen X Q, Shi Y R, Chao M, et al. Fish community structure

of the Yangtze River estuary in summer and autumn[J]. Journal of Fisheries of China, 2011, 35(5): 700-710. [沈新强, 史赟荣, 晁敏, 等. 夏、秋季长江口鱼类群落结构[J]. 水产学报, 2011, 35(5): 700-710.]

[6] Crooks S, Turner R K. Integrated coastal management: sustaining estuarine natural resources[J]. Advances in Ecological, 1999, 29(8): 241-289.

[7] Shen X Q, Shi Y R, Chao M, et al. Analysis of taxonomic diversity of fish community in Yangtze River estuary[J]. Progress in Fishery Sciences, 2013, 34(4): 1-7. [沈新强, 史赟荣, 晁敏, 等. 长江口鱼类群落分类学多样性变动的分析[J]. 渔业科学进展, 2013, 34(4): 1-7.]

[8] Shi Y R, Chao M, Quan W M, et al. Spatial variation in fish community of Yangtze River estuary in spring[J]. Journal of Fishery Sciences of China, 2011, 18(5): 1141-1151. [史赟荣, 晁敏, 全为民, 等. 2010 年春季长江口鱼类群落空间分布特征[J]. 中国水产科学, 2011, 18(5): 1141-1151.]

[9] Breiman L I, Friedman J H, Olshen R A, et al. Classification and regression trees (CART)[J]. Biometrics, 1984, 40(3): 358.

[10] Watters R, Deriso R B. Catch per unit of effort of bigeye tuna: A new analysis with regression trees and simulated annealing[J]. Inter-American Tropical Tuna Commission Bulletin, 2000, 21(8): 531-571.

[11] De'Ath G. Multivariate regression trees: A new technique for modeling species-environment relationships[J]. Ecology, 2002, 83(4): 1105-1117.

[12] Guan W J, Chen X J, Gao F, et al. Comparisons of regression tree and GLM performance in CPUE standardization[J]. Journal of Shanghai Ocean University, 2014, 23(1): 123-130. [官文江, 陈新军, 高峰, 等. GLM 模型和回归树模型在 CPUE 标准化中的比较分析[J]. 上海海洋大学学报, 2014, 23(1): 123-130.]

[13] Fielding A H, Haworth P F. Testing the generality of bird-habitat models[J]. Conservation Biology, 1995, 9(6): 1466-1481.

[14] Qian S. Environmental and Ecological Statistics with R[M]. Beijing: Higher Education Press, 2011: 204-207. [钱松. 环境与生态统计: R 语言的应用[M]. 北京: 高等教育出版社, 2011: 204-207.]

[15] Lynch D R, McGillicuddy Jr D J, Werner F E. Skill assessment for coupled biological/physical models of marine systems[J]. Journal of Marine Systems, 2009, 76(1-2): 1-3.

[16] Soyinka O O, Kuton M P, Ayoolalusi C I. Seasonal distribution and richness of fish species in the Badagry Lagoon, south-west Nigeria[J]. Estonian Journal of Ecology, 2010, 59(2): 147-157.

[17] Chen X J. Fish Resources and Fishing Grounds[M]. Beijing: China Ocean Press, 2014: 125-127. [陈新军. 渔业资源与渔场学[M]. 北京: 海洋出版社, 2014: 125-127.]

[18] Whitfield A K, Elliott M, Basset A, et al. Paradigms in estuarine ecology – A review of the Remane diagram with a suggested revised model for estuaries[J]. Estuarine, Coastal and Shelf Science, 2012, 97(1): 78-90.

[19] Sosa-López A, Mouillot D, Ramos-Miranda J, et al. Fish species richness decreases with salinity in tropical coastal lagoons[J]. Journal of Biogeography, 2007, 34(1): 52-61.

[20] Maes J, Stevens M, Ollevier F. The composition and community structure of the ichthyofauna of the upper Scheldt estuary: synthesis of a 10-year data collection (1991–2001) [J]. Journal of Applied Ichthyology, 2005, 21(2): 86-93.

[21] Maes J, Damme S V, Meire P, et al. Statistical modeling of seasonal and environmental influences on the population dynamics of an estuarine fish community[J]. Marine Biology, 2004, 145(5): 1033-1042.

[22] Damme S V, Struyf E, Maris T, et al. Spatial and temporal patterns of water quality along the estuarine salinity gradient of the Scheldt estuary (Belgium and the Netherlands): results of an integrated monitoring approach[J]. Hydrobiologia, 2005, 540(1-3): 29-45.

[23] Maes J, Stevens M, Breine J. Modelling the migration opportunities of diadromous fish species along a gradient of dissolved oxygen concentration in a European tidal watershed[J]. Estuarine Coastal and Shelf Science, 2007, 75(1): 151-162.

[24] Vayssiéres M P, Plant R E, Allen-Diaz B H. Classification trees: An alternative non-parametric approach for predicting species distributions[J]. Journal of Vegetation Science, 2000, 11(5): 679-694.

[25] França S, Cabral H N. Predicting fish species richness in estuaries: Which modelling technique to use?[J]. Environmental Modelling & Software, 2015, 66(66): 17-26.

[26] Pittman S J, Costa B M, Battista T A. Using lidar bathymetry and boosted regression trees to predict the diversity and abundance of fish and corals[J]. Journal of Coastal Research, 2009, 25(6): 27-38.

[27] Pihl L, Cattrijsse A, Codling I, et al. Habitat Use by Fishes in Estuaries and Other Brackish Areas[M]// Fishes in Estuaries. Blackwell Publishing, 2002: 10-53.

Prediction of fish species richness in the Yangtze River estuary using CART algorithm

DAI Libin[1, 2, 3], CHEN Jinhui[4], TIAN Siquan[1, 2, 3, 5], GAO Chunxia[1, 2, 3, 5], WANG Jiaqi[1, 2, 3], DU Xiaoxue[1, 2, 3], WANG Xuefang[1, 2, 3, 5]

1. College of Marine Science, Shanghai Ocean University, Shanghai 201306, China;
2. National Data Centre for Distant-Water Fisheries of China, Shanghai 201306, China;
3. Key Laboratory of Sustainable Exploitation of Oceanic Fisheries Resources, Ministry of Education, Shanghai 201306, China;
4. Superintendence Department of Shanghai Yangtze Estuarine Nature Reserve for Chinese Sturgeon, Shanghai 200092, China;
5. National Distant-water Fisheries Engineering Research Center, Shanghai 201306, China

Abstract: The main ecological patterns and functioning of estuarine ecosystems are difficult to evaluate owing to natural and human induced complexity and variability on biodiversity. Therefore, there is an increased demand to analyze and predict the relationships between the environment and the distribution of biota in estuarine ecosystems. Biodiversity is viewed as the variety of life, encompassing variations from the gene to ecosystem levels, and is commonly expressed as species richness. The patterns of biodiversity in the Yangtze River Estuary have remained largely unexplored, despite the increasing understanding of the importance of estuarine ecosystems and the existing knowledge on the variability of fish communities within estuaries and their environmental drivers. As a transitional system, the Yangtze River Estuary, a typical ecotone, is the largest estuarine ecosystem in the western Pacific Ocean. It establishes links between the marine and freshwater ecosystems in the East China Sea; persistent environmental fluctuations in this estuarine ecosystem creates considerable physiological demands on the species that inhabit this ecosystem. Predictive modelling techniques are being increasingly used to determine major habitat requirements that affect species distribution. Important technological advancements have benefited predictive distribution modelling, and new and sophisticated methods have been developed for use in statistical models that are applied to ecology. The prediction of fish biodiversity has important scientific implications for evaluating the Yangtze River Estuary ecosystem. Based on fishery and environmental data collected in 2012-2013, a regression tree model was built to predict fish species richness in the Yangtze River Estuary. The node structure of the optimal decision tree model indicated that salinity, dissolved oxygen, and month (i.e. season) were three factors affecting fish biodiversity in the Yangtze River Estuary. In addition, the data observed in 2014 was used to validate the predictive performance of the tree-based model by calculating root mean square error (RMSE), average absolute error (AAE), and average relative error (ARE), which were often used as statistical indicators to compare fitted value and observed value in modelling studies. The results showed that the prediction performance was better in spring and summer than in autumn and winter, and generally, the model presents a fair predictable ability indicating the feasibility to predict fish species richness by utilizing a classification and regression trees (CART) algorithm. Estuarine ecosystems are often considered a complex mosaic of habitat types, and their fish biodiversity are best predicted through a CART algorithm. In the present study, in terms of predictive performance, CART could be viewed as an appropriate technique to predict fish species richness in the Yangtze River Estuary.

Key words: classification and regression tree; the Yangtze River estuary; fish species richness; prediction

Corresponding author: CHEN Jinhui. E-mail: 1114260882@qq.com