文章编号：1000-0615(2022)12-2340-09　　水产学报, 2022, 46(12): 2340−2348

水产学报
JOURNAL OF FISHERIES OF CHINA
DOI: 10.11964/jfc.20210212651
Science Press

基于 PCA-GAM 的阿拉伯海公海鸢乌贼资源量空间分布预测模型研究

范秀梅[1,2]，崔雪森[1,2*]，唐峰华[1,2]，樊　伟[1,2]，伍玉梅[1,2]，张　衡[1,2]
(1. 中国水产科学研究院东海水产研究所，农业农村部远洋与极地渔业创新重点实验室，上海　200090；
2. 中国水产科学研究院渔业资源与遥感信息技术重点开放实验室，上海　200090)

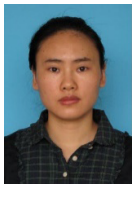
摘要：为了科学预测鸢乌贼资源量的分布，更加合理开发和利用其资源，实验利用2017—2019年阿拉伯海公海灯光围网鸢乌贼生产数据，结合同期的盐度、温度、混合层厚度、海面高度异常、叶绿素a浓度、海表流速、经度和纬度数据构建了阿拉伯海鸢乌贼渔场的PCA-GAM预报模型。环境因子间的相关性会形成多重共线性，易造成模型过拟合，降低模型的预报能力。基于主成分分析(principal component analysis，PCA)降维技术，将环境数据转变成少数几个不相关但保留重要信息的主成分(PCs)，将前8个PCs作为广义加性模型(GAM)的解释变量构建模型。利用交叉验证得到预报值和实际单位捕捞努力量渔获量(CPUE)[经过ln(CPUE+1)变换]相关系数均值为0.532 7，回归模型斜率的均值为0.708 7，截断的均值为1.471 1。模型预报的鸢乌贼资源量分布和实际的CPUE[经过ln(CPUE+1)变换]在空间上重叠度较高，表明PCA-GAM模型能够较好地预报阿拉伯海鸢乌贼资源量的空间分布。

关键词：鸢乌贼；广义加性模型(GAM)；主成分分析(PCA)；空间分布模型；阿拉伯海
中图分类号：S 932　　**文献标志码：**A

鸢乌贼(*Sthenoteuthis oualaniensis*)隶属柔鱼科(Ommastrephidae)鸢乌贼属(*Sthenoteuthis*)，俗名南鱿，繁殖力强，生长速率较快，生命周期短，属于大洋暖水性物种，广泛分布于热带和副热带的印度洋和太平洋区域[1-2]。2015年我国开始在阿拉伯海公海海域进行公海灯光围网捕捞作业，鸢乌贼是2017—2019年围网作业渔获物种类之一。阿拉伯海公海海域鸢乌贼渔场的形成及其分布主要与索马里海流流经区域广泛存在的上升流有关[3]。鸢乌贼是海洋中上层的渔业资源，其渔场对环境变化极为敏感。邵峰等[4]分析了2004年9月至2005年1月的阿拉伯海鸢乌贼资源的探捕调查数据，结果表明，阿拉伯海鸢乌贼的中心渔场分布在冷暖涡交汇处，并处于冷水涡边缘一侧，即海面高度异常值小于且接近0的海域。陈新军等[5]基于2003年9—11月的阿拉伯海鸢乌贼资源的探捕调查数据分析表明，10月下旬中心渔场分布在15°N~16°N，61°E附近海域，表温为27~29 °C，盐度为35.96~36.03。田思泉等[6]基于2004年9—12月的阿拉伯海鸢乌贼资源的探捕调查数据，利用广义加性模型(generalized additive model, GAM)分析表明，鸢乌贼产量与表温、50和200 m水温及各层盐度的关系密切。

GAM已经被广泛应用于研究渔场资源分布

收稿日期：2021-02-23　　修回日期：2021-03-25
资助项目：国家重点研发计划(2019YFD0901405)；中国水产科学研究院东海水产研究所中央级公益性科研院所基本业务费专项(2021M06)
第一作者：范秀梅(照片)，从事海洋渔业相关研究，E-mail：fxm1fxm@163.com
通信作者：崔雪森，从事海洋渔业相关研究，E-mail：cui1012@sh163.net



与环境因子之间的关系，其本质是利用高阶的多项式来模拟环境变量对渔场资源分布的线性和复杂的非线性作用[7-9]。国内外学者利用 GAM 对各大洋的渔场资源与环境因子的关系进行了一系列的研究和探讨，并利用 GAM 在渔场空间分布预测等方面做了相关研究[10-16]。主成分分析 (principal component analysis，PCA) 是通过降维技术把多个相关变量转换为不相关的几个少数主成分 (PCs) 的统计方法[17]，其主要目的就是将许多相关性很高的变量转化成少数几个相互独立的变量，用较少的变量去解释原资料中的大部分信息。孙宵等[18] 利用多年海州湾附近海域的底拖网调查数据及时空、环境等因子构建了基于主成分分析的短吻红舌鳎 (*Cynoglossus joyneri*) 资源 PCA-GAM 预测模型，经过交叉验证表明，PCA-GAM 模型的拟合度和预测效果均优于普通的 GAM。

国内外还未见阿拉伯海鸢乌贼资源量空间分布预测的相关报道，为了更好地了解和可持续开发利用阿拉伯海鸢乌贼资源，本研究拟基于 PCA-GAM 模型，利用 2017—2019 年我国在阿拉伯海公海捕捞作业的渔获数据，结合 0、50、100、150 和 200 m 水层的盐度、温度、混合层厚度、海面高度异常、叶绿素 a (chlorophyll a，Chl-a) 浓度、海表流速、经度及纬度，建立阿拉伯海鸢乌贼资源量空间分布预报模型，并进行交叉验证，以期为鸢乌贼资源的开发和利用提供科学依据。

1　材料与方法

1.1　数据来源

鸢乌贼的捕捞数据来自中国 2017—2019 年阿拉伯海公海灯光围网商业渔捞日志，数据记录包括作业日期、经度、纬度、鱼种、产量等。2017—2019 年鸢乌贼渔获量的空间分布如图 1 所示，鸢乌贼中心渔场主要分布在临近阿拉伯半岛的海域。图 2 显示了 2017—2019 年各月的鸢乌贼总产量和各月单位捕捞努力量渔获量 (catch per unit effort，CPUE) 的均值及标准误差 (standard error，SE)。各月的产量分布不均，因夏季热带气旋多发[19]，阿拉伯海公海围网作业主要在 9 月至翌年 5 月。

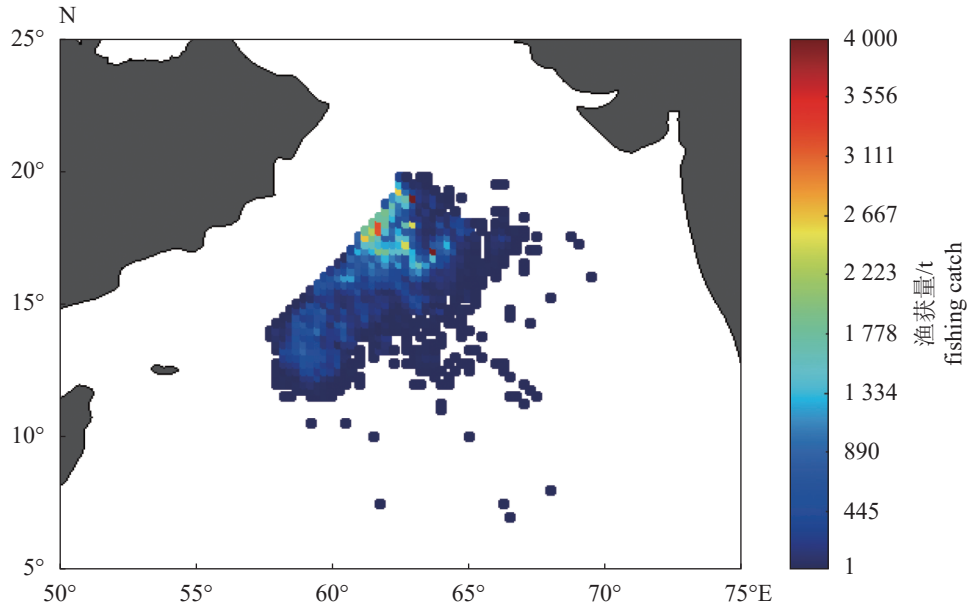


图 1　2017—2019 年阿拉伯海鸢乌贼渔获量的空间分布

Fig. 1　Distribution of fish catches of *S. oualaniensis* in the Arabian Sea during 2017-2019

叶绿素 a 浓度数据来自 CMEMS (the Copernicus Marine Environment Monitoring Service) 提供的全球再分析数据 (https://resources.marine.copernicus.eu/?option=com_csw&task=results)，时间分辨率为月，空间分辨率为 1/4°。海水盐度 (salinity，S)、温度 (temperature，T)、混合层厚度 (mixed layer thickness, MLT)、海面高度异常、海表流速 (sea surface vector，SSV) 数据是 CMEMS 提供的基于现场观测和卫星观测的全球海洋 3D 网格的 4 级再处理 (reprocessed，REP) 数据，时间分辨率

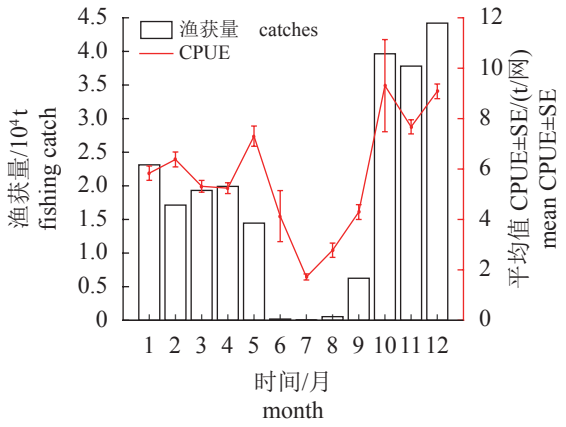


图 2 2017—2019 年阿拉伯海鸢乌贼渔获量和 CPUE 的月分布

Fig. 2 Monthly distribution of fish catches of *S. oualaniensis* in the Arabian Sea during 2017-2019

为月，空间精度为 1/4°。海平面异常 (sea level anomaly, SLA)为海面高度数据距平值。环境数据范围为阿拉伯海鸢乌贼渔场，经度、纬度范围：5°N ~ 20°N、50°E ~ 75°E。

1.2 处理方法

构建和验证 PCA-GAM 模型的流程 首先处理环境和捕捞数据；接着随机采样抽取 80% 的数据构建 PCA-GAM 模型，20% 的数据进行预报，并重复 100 次；最后对上一步交叉验证的结果进行统计分析、验证，并进行鸢乌贼资源量空间分布的预报。

数据处理 实验中建立阿拉伯海鸢乌贼 GAM 的因变量为 CPUE，自变量包括环境变量 (0、50、100、150 和 200 m 水层的 S 和 T、MLT、SLA、Chl-a、SSV)，空间变量 (经度、纬度)。CPUE 的计算公式：CPUE=Catch/N，其中 Catch 为 0.25°×0.25°网格内各月总的渔获量，N 为 0.25°×0.25°网格内各月总的作业网次，单位为 t/网。

利用主成分分析的方法对变量进行降维，得到相互正交的主成分 PCs，然后将 PCs 作为 GAM 的解释变量。为消除各环境变量标准差不同的影响，环境变量值先标准化后再进行主成分分析。CPUE 中存在 0 或较小的值，这里用变量 min_cpue 表示这个较小的值，当 CPUE<min_cpue 表示其对应的时空位置不足以称为渔场，实验中 min_cpue=1 t/网。用二值变量 y 表示阿拉伯海鸢乌贼渔场是否出现[20]，即 y=0 表示 CPUE≤min_cpue，y=1 表示 CPUE>min_cpue。用条件概率值$p=P(y=1|\mathrm{PCs}, \mathrm{lon}, \mathrm{lat})$表示在主成分 PCs、经度、纬度的条件下鸢乌贼渔场出现的概率，那么 1−p 表示在主成分 PCs、经度、纬度的条件下鸢乌贼渔场不发生的概率。CPUE 和二值数据 y 与上述主成分分量 PCs 按照空间位置匹配，总有效记录 3 579 条。

PCA−GAM 模型 GAM 可以模拟自变量对因变量的线性、非线性作用，这里将利用 PCs、经度、纬度构建的模型称为 PCA-GAM 模型。PCA-GAM 预测模型的建立分为两步，第一步建立渔场概率值 p 的模型：PCA-GAM1 模型，第二步建立渔场资源量的模型：PCA-GAM2 模型。

PCA-GAM1 模型属于广相加模型中的 logistic 回归模型，因变量为二项分布，公式：

$$\mathrm{logit}\,(p)=\sum_1^n s\,(\mathrm{PC}_i)+\mathrm{s}\,(\mathrm{lon},\mathrm{lat})+\varepsilon \quad (1)$$

式中，PC_i为第i个主成分，n 为模式中选用的主成分分量的总个数，s 表示样条平滑函数，lon 表示经度，lat 表示纬度，ε为模型的截距，常数值，连接函数设为‘logit’函数，误差函数设为二项分布。‘logit’表示 logit 变换，$\mathrm{logit}\,(p)=\ln\left(\frac{p}{1-p}\right)$。主成分分析对经度、纬度等分类变量作用不显著[21]，故空间变量单独建立平滑项。

PCA-GAM2 模型属于广义相加模型中的多重线性、非线性回归模型，因变量服从高斯分布。CPUE 具有较明显的偏态性，需利用对数函数纠偏后再进行拟合[22]，公式：

$$\ln\,(\mathrm{CPUE}+1)=\sum_1^n s\,(\mathrm{PC}_i)+\mathrm{s}\,(\mathrm{lon},\mathrm{lat})+\varepsilon \quad (2)$$

式中，CPUE 表示 0.25°×0.25°网格内的单位捕捞努力量渔获量，其他变量含义与公式 (1) 一致，连接函数设为 ‘identity’，误差设为高斯分布 $N(\mu, \sigma^2)$。利用 3.6.0 版 R 软件 (https://www.r-project.org/) 的 mgcv 软件包进行 GAM 模型的构建和计算。

模型验证 因数据量有限，利用交叉验证的方法来验证模型的预报结果。随机采样 80% 的数据来构建 PCA-GAM 模型，剩下的 20% 数据用来进行预报，如此反复采样和计算 100 次后，进行结果的统计和检验。基于 PCA-GAM1 模型预报的渔场概率用符号 P'表示，基于 PCA-GAM2模型预报的渔场 CPUE 产量用符号 C 表示，则最终预报值 C'可通过渔场发生的概率 P'和C的积获得[21]，表达公式：

$$C'=P'*C \quad (3)$$

将预报值C'和观测值 CPUE 利用线性回归拟合：

$$C'=a(\ln(\mathrm{CPUE}+1))+b \quad (4)$$

式中，a 为斜率，b 表示截距，a 越接近 1，b 越接近 0 表示预报结果越接近观测值[23]。计算 100 次的模型预报值和观测值的相关系数，并获得相关系数的概率密度图和均值，相关系数值越大，表明预报效果越好。将预报的资源量空间分布与观测值在空间上进行叠加，二者重叠度越高表明预报效果越好。

2　结果

2.1　环境因子的相关系数

0、50、100、150 和 200 m 水层的盐度和水温、MLT、SLA 、Chl-a、SSV 的相关系数[21]如表 1 所示。部分变量之间的线性相关性比较显著，表 1 中的粗体表示相关系数超过 0.5，出现较强的线性相关，直接输入 GAM，会引起模型的过拟合，降低模型的拟合度和预报能力。

2.2　主成分分析

将环境变量 (0、50、100、150 和 200 m水层的盐度和水温、MLT、SLA、Chl-a、SSV) 进行主成分分解后得到前 8 个主成分 PC1~PC8。PC1 平均方差解释率占比 24.35% (±2.13%)，PC2 占比 17.49% (±1.98%)，PC3 占比 13.73% (±2.11%)，前 3 个主成分的平均方差解释率占比 55.58% (±2.43%)，PC4~PC8 共占比 22.97% (±1.40%)，前8 个主成分的平均方差解释率占比 87.34% (±0.86%)。

将 PC1~PC8 与各环境因子载荷的绝对值以箱图的形式显示 (图 3)。PC1 是前 8 个主成分中包含信息量最多的，PC1~PC8 的载荷整体逐渐降低。0 ~ 200 m 水层的盐度和温度在 PC1 中载荷最大，PC2、PC3 次之，在 PC4~PC8 中，其载荷小于 0.2。MLT 的载荷在 PC1、PC2 中最高，PC3 次之，在 PC4~PC8 中载荷小于 0.2。SLA 的载荷在 PC1、PC2 中最高，PC3、PC4、PC5、PC6 次之。Chl-a 的载荷在 PC1、PC2、PC3、PC4 中最高，PC5、PC6 次之。SSV 的载荷在 PC2、PC3、PC4 中最高，PC1、PC5、PC6 次之。SLA、Chl-a、SSV在 PC7、PC8 中载荷均小于 0.2。

2.3　模型分析

将模型拟合结果中 P 值小于显著性水平 0.001 的变量认定为对模型的影响有显著性，并统计 100 次模型拟合中变量显著性次数的占比，结果显示在表 2 中。在 PCA-GAM1 模型中，PC1、PC4、PC5、PC6、经度、纬度显著性占比较高，100 次建模中超过 83 次都为显著，PC2 的显著性占比较低，占比 9%，PC3、PC7、PC8 的显著性占比接近 0，表明可只用 PC1、PC2、PC4、PC5、

表 1　各水层环境因子之间的相关系数

Tab. 1　Correlation coefficients between each environmental factor

	S0	S50	S100	S150	S200	T0	T50	T100	T150	T200	MLT	SLA	Chl-a	SSV
S0	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
S50	**0.78**	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00
S100	**0.72**	**0.76**	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S150	0.34	0.27	**0.54**	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S200	0.22	0.06	0.25	**0.71**	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
T0	0.23	0.15	0.40	0.17	0.08	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
T50	0.18	0.39	0.35	0.12	−0.15	0.28	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T100	0.27	0.47	**0.52**	0.06	−0.41	0.15	**0.64**	—	0.00	0.00	0.00	0.00	0.00	0.00
T150	0.15	0.29	0.43	0.38	−0.10	0.14	**0.60**	**0.85**	—	0.00	0.00	0.00	0.00	0.00
T200	0.14	0.17	0.37	**0.63**	0.45	0.10	0.35	0.45	**0.78**	—	0.03	0.00	0.00	0.00
MLT	−0.05	0.07	−0.18	−0.16	−0.21	**−0.70**	0.19	0.21	0.17	0.04	—	0.00	0.00	0.00
SLA	−0.03	0.00	0.29	0.29	−0.06	0.48	**0.51**	**0.57**	**0.74**	**0.63**	−0.18	—	0.00	0.00
Chl-a	−0.12	−0.12	−0.29	−0.24	−0.04	**−0.55**	−0.20	−0.24	−0.32	−0.30	0.44	−0.44	—	0.00
SSV	−0.41	−0.44	−0.38	−0.09	0.07	−0.04	−0.32	−0.43	−0.32	−0.18	−0.15	−0.09	0.09	—

注：表中下三角为相关系数，上三角为P值

Notes: Values in the lower triangle of the table are correlation coefficients with bold font being significant at correlation coefficients(> 0.5), and values in the upper triangle are P values

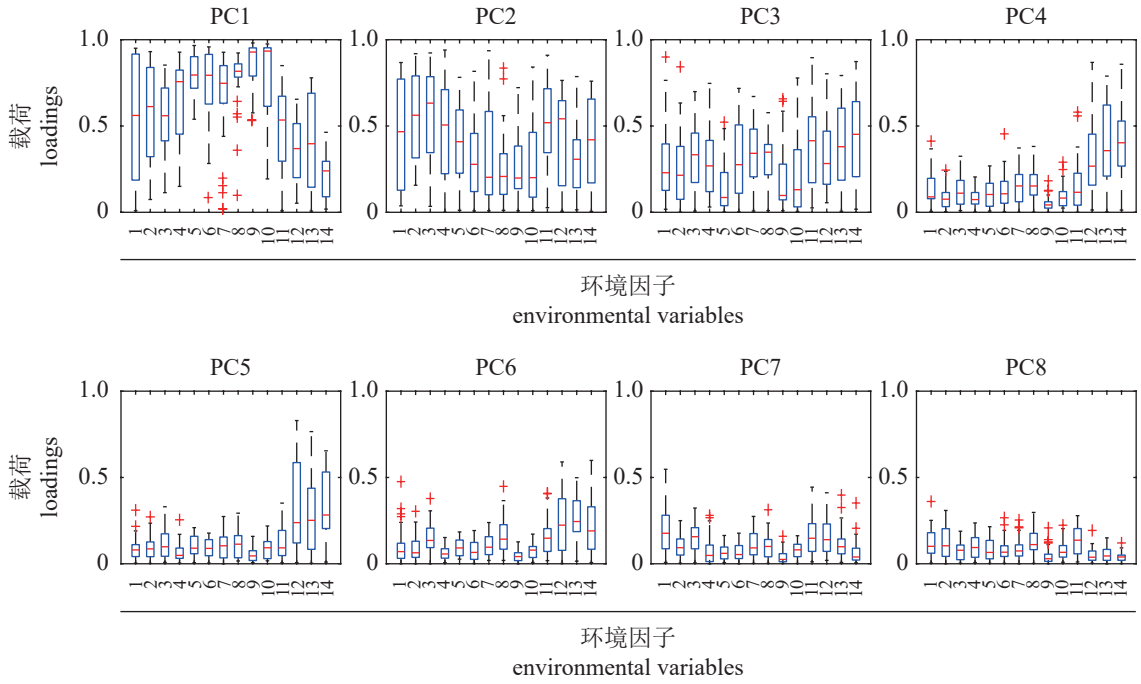


图 3 100 次模拟的各主成分中环境因子的载荷

Fig. 3 Box plots of correlation coefficients between environmental variables and each principle component over 100 simulation runs

1. S0, 2. S50, 3. S100, 4. S150, 5. S200, 6. T0, 7. T50, 8.T100, 9. T150, 10. T200, 11. MLT, 12. SLA, 13. Chl-a, 14. SSV

表 2 100 个 PCA-GAM 模型中各变量显著性的占比

Tab. 2 Proportion of 100 simulation runs in which a factor was indentified as significant

因子 factors	阶段1/% stage1	阶段2/% stage2
PC1	100	100
PC2	9	100
PC3	0	53
PC4	83	100
PC5	95	100
PC6	100	100
PC7	1	43
PC8	0	39
(lon,lat)	100	100

PC6、经度、纬度建立 PCA-GAM1 模型。在 PCA-GAM2 模型中，PC1、PC2、PC4、PC5、PC6、经度、纬度显著性占比较高，100 次建模中均显著，PC2、PC7、PC8 的显著性占比稍低，分别占比 53%、43%、39%，表明需利用 PC1 至 PC8、经度、纬度建立 PCA-GAM2 模型。

2.4 模型验证

交叉验证 100 次模拟中，预报值和观测值相关系数的变化范围为 0.459 2~0.614 3，均值为 0.532 7，方差为 0.034 2 (图 4)。预报值和观测值之间的一元一次回归模型的斜率 a 的均值为 0.708 7，方差为 0.054 1，截断 b 的均值为 1.471 1，方差为 0.484 8 (表 3)。

鸢乌贼分布预报 利用 PCA-GAM 进行鸢乌贼渔场资源量空间分布的预测，建立模型使用的数据为剔除预测月份后剩余的数据，包括环境

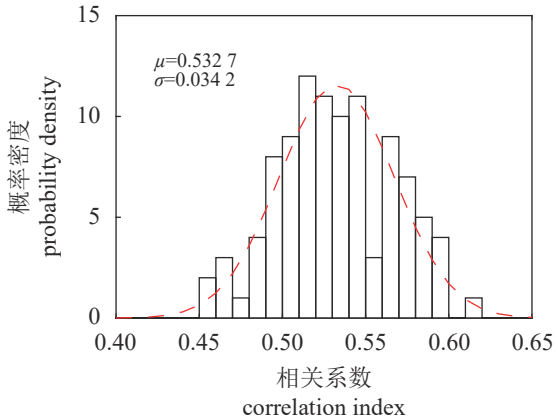


图 4 100 次模拟的测试数据中观测和预报 CPUE 的相关系数的概率分布

Fig. 4 Probability distribution of correlation coefficients between observed and predicated CPUE for the test data in the cross-validation over 100 simulation runs

表 3　观测和预报值回归模型的系数

Tab. 3　Coefficients for the regression models of observed and predicted values

系数 coefficients	PCA-GAM 模型 PCA-GAM model
相关系数　correlation coefficient	0.532 7±0.034 2
斜率a　slope(a)	0.708 7±0.054 1
截距b　intercept (b)	1.471 1±0.484 8

注：表中的系数标注了标准差
Note: Standard deviances were provided for these values

数据、经度、纬度、CPUE[经过 ln(CPUE+1) 变换]，预报使用的数据为当月的环境数据、经度、纬度。将 2019 年 1—4 月，9—12 月实际的 CPUE [经过 ln(CPUE+1) 变换] 与预报值进行空间叠加对比，从而检验渔场预测的效果 (图 5)。鸢乌贼 CPUE 高值区主要分布在临近阿拉伯半岛的海域，PCA-GAM 模型预测的鸢乌贼高产量区与实际 CPUE 分布大致相符，表明 PCA-GAM 能够预测阿拉伯海鸢乌贼资源的分布。

3　讨论

阿拉伯海域鸢乌贼中心渔场形成的内在动力为季风吹拂引起的上升流，上升流引起底层海水上涌，带来了丰富营养盐，使得藻类繁殖旺盛，叶绿素浓度增加。上升流区域由于底层冷水与表层水交汇，导致海水温度、盐度的变化大。海洋环境因素对鸢乌贼生物量的分布有着重要的影响[24-25]，例如海水温度不仅影响头足类自身卵的孵化率、成体的大小等[26]，还可能影响头足类饵料生物的丰度，故其对海水温度较为敏感[27-28]。海面高度异常与海水的辐聚、辐散或者下降、涌升有关，往往能够带来底层海域丰富的营养盐[4]，温跃层深度会限制鸢乌贼昼夜的垂直移动，适宜的海水盐度有利于鸢乌贼与海水环境的渗透作用[25]，叶绿素 a 浓度影响海域的饵料状况，海表地转流会影响近海表层中型鱼类到公海洄游和觅食[29]。鸢乌贼昼夜会垂直迁移，白天在水下较深处，夜晚到近表层水域觅食[2]，故海水温度和盐度的垂直结构会对鸢乌贼的产量和渔场形成有一定的影响[30]。田思泉等[6] 从鱼探仪映像和生产实际情况看，鸢乌贼主要栖息在 200 m 以上浅水层，故选择 0~200 m水层的海水温度、盐度数据来研究。

由于一些海洋环境变量之间存在较高的相关性 (表 1)，故实验中利用了主成分分析对环境变量进行了降维[31]，去除多重共线性，将得到的前 8 个主分量 PC1~PC8 作为 GAM 的解释变量，构建 PCA-GAM 模型。不同主成分中各变量的重要性通过变量在其上的载荷来体现，载荷的大小等于其各自的相关系数，也即是反映该变量与主成分之间关系的密切度[17]。从图 4 中可见，PC1 中包含了各水层盐度、温度的大部分信息，包含了 MLT、SLA、

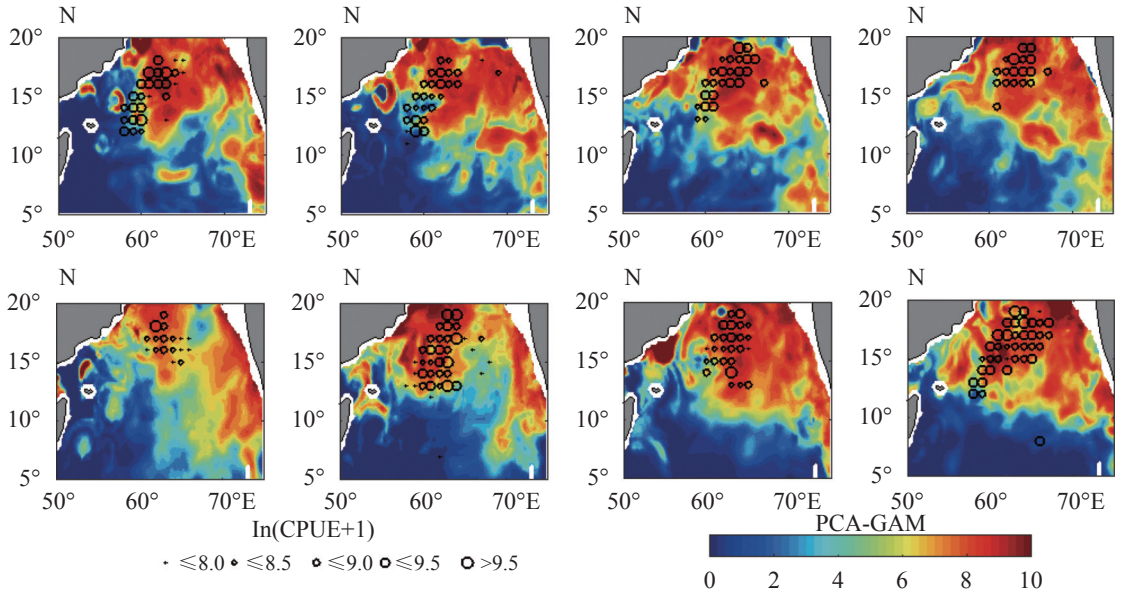


图 5　阿拉伯海鸢乌贼 2019 年各月 CPUE 与 PCA-GAM 模型预报结果对比

Fig. 5　The monthly observed CPUE and forecast results distribution of *S. oualaniensis* based on PCA-GAM in the Arabian Sea in 2019

Chl-a、SSV 的部分信息。PC2、PC3 中包含了各水层盐度、温度、MLT、SLA、Chl-a、SSV 的部分信息。PC4~PC5 包含了 SLA、Chl-a、SSV 的部分信息。PC1~PC8、经度、纬度在 PCA-GAM2 中的显著性占比都较高 (表 2)，表明 0、50、100、150 和 200 m 水层的盐度、温度、MLT、SLA、Chl-a、SSV 对鸢乌贼资源分布的影响较大。

PCA-GAM 模型中包括 2 个子模型：计算阿拉伯海鸢乌贼渔场发生概率的 logistics 回归模型 PCA-GAM1，计算 CPUE[经过 ln(CPUE+1) 变换] 值的多重线性、非线性模型 PCA-GA2，最终的预测值为 2 个子模型结果的乘积，2 个模型耦合可以一定程度缓解 CPUE 中 0 值膨胀的影响[32]。在 100 次的 8 折交叉验证中，预报值和观测值相关系数的均值超过 0.5(图 4)，表明该模型具有较好的预测性能。利用该模型预报的 2019 年 1—4 月、9—12 月的鸢乌贼资源量空间分布与观测值重叠度较高 (图 5)，表明该模型能够较好地预报阿拉伯海公海鸢乌贼资源量的空间分布。

（作者声明本文无实际或潜在的利益冲突）

参考文献 (References):

[1] Liu B L, Chen X J, Li J H, *et al*. Age, growth and maturation of *Sthenoteuthis oualaniensis* in the eastern tropical Pacific Ocean by statolith analysis[J]. Marine and Freshwater Research, 2016, 67(12): 1973-1981.

[2] Snÿder R. Aspects of the biology of the giant form of *Sthenoteuthis oualaniensis* (Cephalopoda: Ommastrephidae) from the arabian sea[J]. Journal of Molluscan Studies, 1998, 64(1): 21-34.

[3] 余为, 陈新军. 印度洋西北海域鸢乌贼9-10月栖息地适宜指数研究[J]. 广东海洋大学学报, 2012, 32(6): 74-80.
Yu W, Chen X J. Analysis on habitat suitability index of *Sthenoteuthis oualaniensis* in Northwestern Indian Ocean from september to october[J]. Journal of Guangdong Ocean University, 2012, 32(6): 74-80 (in Chinese).

[4] 邵锋, 陈新军. 印度洋西北海域鸢乌贼渔场分布与海面高度的关系[J]. 海洋科学, 2008, 32(11): 88-92.
Shao F, Chen X J. Relationship between fishing ground of *Symlectoteuthis oualaniensis* and sea surface height in the northwest Indian ocean[J]. Marine Sciences, 2008, 32(11): 88-92 (in Chinese).

[5] 陈新军, 钱卫国, 田思泉. 阿拉伯海北部公海海域鸢乌贼资源密度及其分布[J]. 海洋科学进展, 2006, 24(3): 360-364.
Chen X J, Qian W G, Tian S Q. Resource density and distributions of *Symlectoteuthis oualaniensis* in open seas of Northern Arabian Sea[J]. Advances in Marine Science, 2006, 24(3): 360-364 (in Chinese).

[6] 田思泉, 陈新军, 杨晓明. 阿拉伯北部公海海域鸢乌贼渔场分布及其与海洋环境因子关系[J]. 海洋湖沼通报, 2006(1): 51-57.
Tian S Q, Chen X J, Yang X M. Study on the fishing ground distribution of *Symlectoteuthis oualaniensis* and its relationship with the environmental factors in the high sea of the Northern Arabian Sea[J]. Transactions of Oceanology and Limnology, 2006(1): 51-57 (in Chinese).

[7] Guisan A, Edwards Jr T C, Hastie T. Generalized linear and generalized additive models in studies of species distributions: setting the scene[J]. Ecological Modelling, 2002, 157(2-3): 89-100.

[8] Liu X X, Wang J, Zhang Y L, *et al*. Comparison between two GAMs in quantifying the spatial distribution of *Hexagrammos otakii* in Haizhou Bay, China[J]. Fisheries Research, 2019, 218: 209-217.

[9] Lan K W, Lee M A, Wang S P, *et al*. Environmental variations on swordfish (*Xiphias gladius*) catch rates in the Indian Ocean[J]. Fisheries Research, 2015, 166: 67-79.

[10] Li M K, Zhang C L, Xu B D, *et al*. A comparison of GAM and GWR in modelling spatial distribution of Japanese mantis shrimp (*Oratosquilla oratoria*) in coastal waters[J]. Estuarine, Coastal and Shelf Science, 2020, 244: 106928.

[11] 谢恩阁, 周艳波, 冯菲, 等. 中国南海外海鸢乌贼灯光罩网渔业CPUE标准化研究[J]. 大连海洋大学学报, 2020, 35(3): 439-446.
Xie E G, Zhou Y B, Feng F, *et al*. Catch per unit effort (CPUE) standardization of purpleback flying squid *Sthenoteuthis oualaniensis* for Chinese large-scale lighting net fishery in the open sea of South China Sea[J]. Journal of Dalian Ocean University, 2020, 35(3): 439-446 (in Chinese).

[12] Carvalho F C, Murie D J, Hazin F H V, *et al*. Spatial predictions of blue shark (*Prionace glauca*) catch rate and catch probability of juveniles in the Southwest Atlantic[J]. ICES Journal of Marine Science, 2011, 68(5): 890-900.

[13] 李杰, 张鹏, 晏磊, 等. 南海中南部海域鸢乌贼CPUE影响因素的GAM分析[J]. 中国水产科学, 2020, 27(8): 906-915.
Li J, Zhang P, Yan L, *et al*. Factors that influence the catch per unit effort of *Sthenoteuthis oualaniensis* in the

central-southern South China Sea based on a generalized additive model[J]. Journal of Fishery Sciences of China, 2020, 27(8): 906-915 (in Chinese).

[14] Mohamed K S, Sajikumar K K, Ragesh N, *et al*. Relating abundance of purpleback flying squid *Sthenoteuthis oualaniensis* (Cephalopoda: Ommastrephidae) to environmental parameters using GIS and GAM in south-eastern Arabian Sea[J]. Journal of Natural History, 2018, 52(29-30): 1869-1882.

[15] 牛明香, 李显森, 徐玉成. 基于广义可加模型的时空和环境因子对东南太平洋智利竹筴鱼渔场的影响[J]. 应用生态学报, 2010, 21(4): 1049-1055.
Niu M X, Li X S, Xu Y C. Effects of spatiotemporal and environmental factors on the fishing ground of *Trachurus murphyi* in Southeast Pacific Ocean based on generalized additive model[J]. Chinese Journal of Applied Ecology, 2010, 21(4): 1049-1055 (in Chinese).

[16] Su N J, Chang C H, Hu Y T, *et al*. Modeling the spatial distribution of swordfish (*Xiphias gladius*) using fishery and remote sensing data: approach and resolution[J]. Remote Sensing, 2020, 12(6): 947.

[17] 黄嘉佑. 气象统计分析与预报方法 [M]. 北京: 气象出版社, 1990.
Huang J Y. Methods of meteorological statistical analysis and forecast[M]. Beijing: China Meteorological Press, 1990 (in Chinese).

[18] 孙霄, 张云雷, 刘笑笑, 等. 两种GAM模型对海州湾短吻红舌鳎(*Cynoglossus joyneri*)资源分布预测效果的比较研究[J]. 海洋学报, 2020, 42(6): 20-28.
Sun X, Zhang Y L, Liu X X, *et al*. Evaluation of the prediction effect of two GAMs on the distribution of *Cynoglossus joyneri* in the Haizhou Bay[J]. Haiyang Xuebao, 2020, 42(6): 20-28 (in Chinese).

[19] 雷茜, 罗红霞, 白林燕, 等. 阿拉伯海区域气溶胶时空动态变化及海域叶绿素a浓度特征[J]. 生态学报, 2019, 39(9): 3110-3120.
Lei Q, Luo H X, Bai L Y, *et al*. Spatio-temporal dynamic characteristics of aerosol and chlorophyll a concentration in the Arabian Sea[J]. Acta Ecologica Sinica, 2019, 39(9): 3110-3120 (in Chinese).

[20] 崔雪森, 伍玉梅, 周爱忠, 等. 基于Logistic回归模型的西非沿海欧洲沙丁鱼渔场与环境因素关系模型的构建[J]. 大连海洋大学学报, 2016, 31(2): 211-218.
Cui X S, Wu Y M, Zhou A Z, *et al*. A Logistic regression model on relationship between sardine *Sardina pilchardus* fishing ground and environmental variables in the coastal waters of West Africa[J]. Journal of Dalian Ocean University, 2016, 31(2): 211-218 (in Chinese).

[21] Zhao J, Cao J, Tian S Q, *et al*. A comparison between two GAM models in quantifying relationships of environmental variables with fish richness and diversity indices[J]. Aquatic Ecology, 2014, 48(3): 297-312.

[22] Yu J, Hu Q W, Tang D L, *et al*. Response of *Sthenoteuthis oualaniensis* to marine environmental changes in the north-central South China Sea based on satellite and *in situ* observations[J]. PLoS One, 2019, 14(1): e0211474.

[23] Chang J H, Chen Y, Holland D, *et al*. Estimating spatial distribution of american lobster *Homarus americanus* using habitat variables[J]. Marine Ecology Progress Series, 2010, 420: 145-156.

[24] 江淼. 南海鸢乌贼资源现状及发展对策研究 [D]. 天津: 天津农学院, 2018.
Jiang M. The current situation and countermoves on development and utilization of *Sthenoteuthis oualaniensis* in South China Sea[D]. Tianjin: Tianjin Agricultural University, 2018 (in Chinese).

[25] 招春旭. 南海鸢乌贼渔场时空分布及其预报模型构建 [D]. 湛江: 广东海洋大学, 2017.
Zhao C X. The spatial-temporal distribution and the construction of the prediction model of the purpleback flying squid (*Sthenoteuthis oualaniensis*) in South China Sea[D]. Zhanjiang: Guangdong Ocean University, 2017 (in Chinese).

[26] 陆化杰, 童玉和, 刘维, 等. 厄尔尼诺年春季中国南海中沙群岛海域鸢乌贼的渔业生物学特性[J]. 水产学报, 2018, 42(6): 912-921.
Lu H J, Tong Y H, Liu W, *et al*. Fisheries biological characteristics of *Sthenoteuthis oualaniensis* in the spring season in the El Niño year of 2016 in the Zhongsha Islands waters of South China Sea[J]. Journal of Fisheries of China, 2018, 42(6): 912-921 (in Chinese).

[27] Chen X J, Liu B L, Tian S Q, *et al*. Fishery biology of purpleback squid, *Sthenoteuthis oualaniensis*, in the northwest Indian Ocean[J]. Fisheries Research, 2007, 83(1): 98-104.

[28] Chen X J, Zhao X H, Chen Y. Influence of El Niño/La Niña on the western winter-spring cohort of neon flying squid (*Ommastrephes bartramii*) in the northwestern Pacific Ocean[J]. ICES Journal of Marine Science, 2007, 64(6): 1152-1160.

[29] 杨胜龙, 范秀梅, 张忭忭, 等. 阿拉伯海域围网渔场时空分布及其与海表环境因子的关系[J]. 中国农业科技导报, 2019, 21(9): 149-158.
Yang S L, Fan X M, Zhang B B, *et al*. Spatial-temporal distribution of purse seine in Arabian sea and its relationship with sea surface environmental factors[J]. Journal of Agricultural Science and Technology, 2019, 21(9): 149-

158 (in Chinese).

[30] 晏磊, 张鹏, 杨炳忠, 等. 南海鸢乌贼产量与表温及水温垂直结构的关系[J]. 中国水产科学, 2016, 23(2): 469-477.
Yan L, Zhang P, Yang B Z, *et al*. Relationship between the catch of *Symplectoteuthis oualaniensis* and surface temperature and the vertical temperature structure in the South China Sea[J]. Journal of Fishery Sciences of China, 2016, 23(2): 469-477 (in Chinese).

[31] Bierman P, Lewis M, Ostendorf B, *et al*. A review of methods for analysing spatial and temporal patterns in coastal water quality[J]. Ecological Indicators, 2011, 11(1): 103-114.

[32] Barry S C, Welsh A H. Generalized additive modelling and zero inflated count data[J]. Ecological Modelling, 2002, 157(2-3): 179-188.

Research on the prediction model of spatial distribution of *Sthenoteuthis oualaniensis* in the open sen Arabian Sea based on PCA-GAM

FAN Xiumei [1,2], CUI Xuesen [1,2*], TANG Fenghua [1,2], FAN Wei [1,2], WU Yumei [1,2], ZHANG Heng [1,2]

(1. *Key Laboratory of Oceanic and Polar Fisheries*, *Ministry of Agriculture and Rural Affairs*, *East China Sea Fisheries Research Institute*, *Chinese Academy of Fishery Sciences*, *Shanghai* 200090, *China*;
2. *Key and Open Laboratory of Remote Sensing Information Technology in Fishing Resource*, *East China Sea Fisheries Research Institute*, *Chinese Academy of Fishery Sciences*, *Shanghai* 200090, *China*)

Abstract: In order to scientifically predict the distribution of *Sthenoteuthis oualaniensis* and to utilize its resources,this study established the PCA-GAM prediction model of *S. oualaniensis* was established based on the production data of light seine in the open sea Arabian Sea from 2017 to 2019, combined with the data of salinity, temperature, 0, 50, 100, 150 and 200 m water layers, mixed layer thickness, sea level anomaly, chlorophyll *a* concentration, sea surface velocity, longitude and latitude. The correlation between environmental factors will cause multicollinearity, resulting in over-fitting of the model, and reducing the prediction ability of the model. The environmental data were transformed into a few unrelated principal components (PCs) which retained important information of these environmental factors based on the application of dimension reduction techniques such as principle component analysis (PCA). The average variance explanation rate of the top 8 PCs accounted for 87.34% (±0.86%). The top 8 PCs were taken as explanatory variables of the GAM model to construct the prediction model of the distribution of *S. oualaniensis*. The establishment of PCA-GAM prediction model was divided into two-stage GAM. The first stage GAM is to estimate the presence probability of *S. oualaniensis*. The second stage GAM is to estimate the log-transformed CPUE of *S. oualaniensis*. The overall log-transformed CPUE was the product of the results from the first and second stages of the GAM. The eight fold cross-validation results showed that the mean of the correlation coefficients between the predicted values and the practical CPUE (log-transformed) was 0.532 7, the mean of the slopes of the regression models was 0.708 7, and the mean of truncation values was 1.471 1. The degree of overlap between the predicted values and the practical CPUE (log-transformed) from January to April and September to December 2019 was very high in spatial distribution, which indicated that the PCA-GAM model was able to predict the spatial distribution of *S. oualaniensis* in the Arabian Sea.

Key words: *Sthenoteuthis oualaniensis*; generalized additive model, GAM; principle component analysis (PCA); spatial distribution model; Arabian Sea

Corresponding author: CUI Xuesen. E-mail: cui1012@sh163.net

Funding projects: National Key Research and Development Program of China (2019YFD0901405); Central Public-interest Scientific Institution Basal Research Fund, ECSFR, CAFS (2021M06)