KFAS 한국수산과학회지 Korean Journal of Fisheries and Aquatic Sciences
한수지 58(6), 816-823, 2025

Original Article
Korean J Fish Aquat Sci 58(6),816-823,2025

머신러닝 기반 남극은암치(*Pleuragramma antarctica*) NASC 예측 모델 개발

이사라 · 오우석 · 나형술[1,2] · 손우주[1] · 김정훈[3] · 이경훈[4]*

국립부경대학교 어업기술안전연구소, [1]한국해양과학기술원 극지연구소 해양대기연구본부, [2]한국과학기술연합대학교대학원 극지과학과, [3]한국해양과학기술원 극지연구소 생명과학본부, [4]국립부경대학교 해양생산관리학부

Antarctic Silverfish *Pleuragramma antarctica* Nautical Area Scattering Coefficient (NASC) Prediction Using a Machine Learning-Based Model

Sara Lee, Wooseok Oh, Huoungsul Na[1,2], Wuju Son[1], Jeong-Hoon Kim[3] and Kyounghoon Lee[4]*

Institute of Fisheries Technology and Safety, Pukyong National University, Busan 48513, Republic of Korea
[1]Division of Ocean&Atmosphere Science, Korea Polar Research Institute, Incheon 21190, Republic of Korea
[2]Department of Polar Science, University of Science and Technology, Daejeon 34113, Republic of Korea
[3]Division of Life Science, Korea Polar Research Institute, Incheon 21990, Republic of Korea
[4]Division of Marine Production System Management, Pukyong National University, Busan 48513, Republic of Korea

This study aimed to develop a machine learning-based prediction model for the nautical area scattering coefficient (NASC) of the Antarctic silverfish *Pleuragramma antarcticum*, a key species in the Southern Ocean. Acoustic survey data from the Ross Sea from 2018 to 2023 were integrated with environmental variables, including depth, temperature, salinity, survey period, survey area, and grid location, to construct Random Forest regression models. Separate models were trained on the adults and juveniles. For adults, continuous variables were standardized using z-scores. Meanwhile, juvenile models were standardized using raw values. Model training was performed using MATLAB TreeBagger with grid search optimization. The performance was evaluated by hold-out validation. The adult model achieved high accuracy ($R^2 \approx 0.76$, RMSE≈2.10), with depth, temperature, and salinity identified as the most influential predictors. The juvenile model showed lower explanatory power ($R^2 \approx 0.38$, RMSE≈2.54), often underestimating high NASC values. Adults are more strongly governed by physical conditions, whereas juveniles are influenced by additional biological or ecological factors. Random Forest models can effectively predict adult silverfish NASC using limited environmental inputs, supporting the improved interpretation of acoustic data and ecosystem-based management in polar environments.

Keywords: Machine learning, Randomforest, Antarctic silverfish, Nautical area acoustic coefficient

서 론

남극은암치(*Pleuragramma antarcticum*)는 남극 해양생태계에서 가장 풍부하고 생태학적으로 중요한 중층성 어류로, 남극 먹이사슬에서 다양한 상위 포식자들과의 먹이 관계를 통해 에너지 전달 경로의 핵심을 이루는 종이다(Vacchi et al., 2004; Ainley and Siniff, 2009; Ainley et al., 2024). 특히 로스해를 포함한 남극 대륙 주변 해역에서 대규모 군집을 이루며 서식하고 있으며, 이들의 공간 분포와 자원량 변화를 정량적으로 파악하는 것은 생태계 구조 및 기후변화에 따른 반응을 이해하는 데 필수적이다. 해양생물의 분포를 추정하기 위해 수중음향 기법이 널리 활용되고 있으며, NASC (nautical area scattering coefficient)는 상대적 개체군 밀도 및 분포 패턴을 정량화하는 핵심 지표로 사용된다. 하지만 NASC는 음향반사강도를 기반으로 한 지표로, 실제 어류의 분포와 밀도를 결정짓는 수온, 염분, 해류 등의 해양환경 요인을 직접적으로 반영하지 않기 때문

***Corresponding author:** Tel: +82. 51. 629. 5889 Fax: +82. 51. 629. 5886
E-mail address: klee71@pknu.ac.kr

https://doi.org/10.5657/KFAS.2025.0816
Korean J Fish Aquat Sci 58(6), 816-823, December 2025
Received 19 September 2025; **Revised** 22 October 2025; **Accepted** 18 November 2025
저자 직위: 이사라(박사후연구원), 오우석(대학원생), 나형술(선임연구원), 손우주(대학원생), 김정훈(선임연구원), 이경훈(교수)

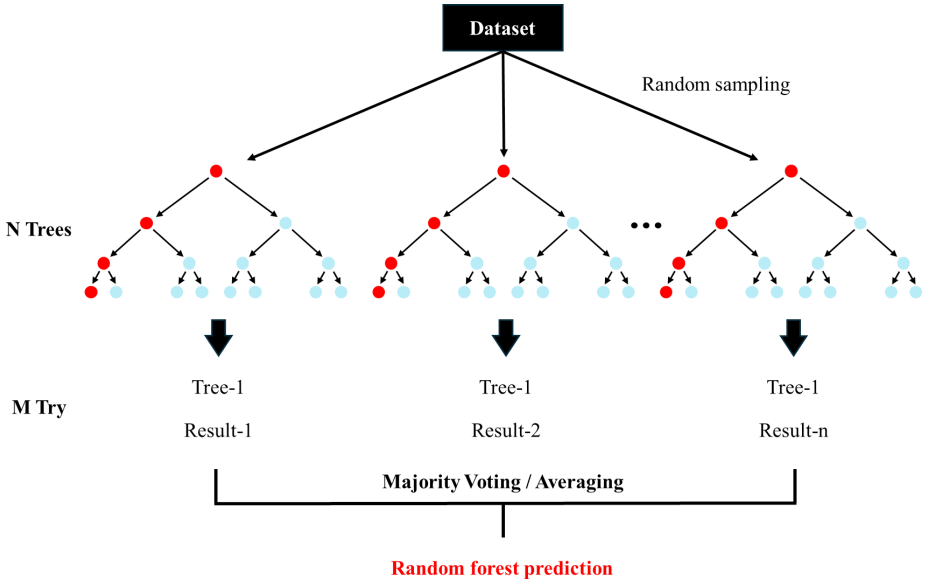

Fig. 1. Conceptual diagram of random forest.

에, 단독으로는 생태적 해석에 한계가 존재한다. 즉, NASC 자체는 어류의 분포를 정량화하는 데 유용하지만, 어류의 분포와 개체군 동태는 환경 변화에 민감하게 반응하기 때문에, 이러한 환경 변수를 통합적으로 고려할 때 음향자료의 해석 정확도뿐만 아니라 생태학적 해석력 또한 크게 향상될 수 있다. 최근에는 NASC와 환경 변수 간과 같은 복잡한 상호작용과 비선형적 관계를 정량적으로 모델링하기 위해 머신러닝(machine learning) 기법이 다양한 연구에 도입되고 있다(Woo et al., 2018; Islam et al., 2021). 그 가운데 랜덤포레스트(random forest)는 높은 예측 정확도와 변수 간 관계 파악 능력으로 주목받고 있다. 랜덤포레스트는 Breiman (2001)이 제안한 앙상블 학습 방법으로, 다수의 결정트리를 통합하여 예측 성능을 향상시킨 기법이다. 각 트리는 원본 데이터로부터 복원 추출(bootstrap sampling)된 데이터셋을 이용하여 개별적으로 학습되며, 이 과정을 배깅(bagging)이라 한다. 학습된 다수의 트리들은 CART (classification and regression) 방법을 토대로 구현되며 회귀문제에서는 평균값을, 분류 문제에서는 다수결 투표를 통해 최종 예측값을 산출한다(Fig. 1). 따라서 본 연구에서는 2018년부터 2023년까지 로스해에서 수행된 음향 조사 자료를 기반으로, 수심, 수온, 염분과 같은 환경 변수 및 조사 시기, 조사 해역 정보를 설명 변수로 활용하여 남극은암치의 NASC 값을 예측하는 랜덤포레스트 회귀 모델을 구축하였다. 이를 통해 제한된 환경 변수 조건에서의 예측 정확도를 평균제곱근오차(root mean square error, RMSE), 평균절대오차(mean absolute error, MAE), 결정계수(coefficient of determination, R^2), OOB (out-of-bag) error 등의 성능 지표를 통해 평가하였으며, 변수 중요도 분석을 통해 남극은암치의 분포를 지배하는 주요 환경 요인을 도출하고자 한다. 또한 성어와 치어를 대상으로 각각 구축된 모델의 예측 성능과 오차를 비교함으로써, 생애 단계에 따라 NASC 변동에 대한 환경 변수의 영향이 다르게 나타나는 경향을 확인하였다.

재료 및 방법

조사해역 및 음향 · 환경자료

조사해역은 남극 로스해 대륙붕(74–76°S, 165–175°E)과 로스 해구(76–78°S, 170–178°E)로 설정하였다. 2018년부터 2023년까지 과학어군탐지기 EK 60, EK 80 (Simrad; Kongsberg Maritime AS, Horten, Norway)를 이용하여 연속 음향 조사를 수행하였다. 수집된 음향자료는 Echoview 12 (version 12; Echoview Software Pty. Ltd., Hobart, Australia)를 이용해 time varied threshold (TVT) 방법으로 배경 잡음을 제거하고, Impulse 및 Electrical Noise 필터를 적용하였다. 이후 주파수별 평균체적후방산란강도(mean volume backscattering strength, MVBS)를 산출한 뒤, 주파수차($\triangle MVBS = MVBS_{200} - MVBS_{120}$)를 이용하여 남극은암치의 신호를 분리하였다. 남극은암치 성어는 $-1.42 \leq \triangle MVBS_{200-120} \leq 0.60$ dB, 치어는 $-1.22 \leq \triangle MVBS_{200-120} \leq 0.74$ dB 범위를 적용하였다. 식별된 남극은암치의 신호를 1 nmi 수평 간격, 10 m 수직 셀 단위로 적분하여 NASC (m^2/nmi^2)로 변환하였다. 유사한 시점에 투하한 SBE 911plus CTD (Sea-Bird Electronics, Bellevue, WA, USA) 프로파일로부터 수온(°C) 및 염분(psu)을 추출하였으며, 위도·경도는 GPS로 기록하였다. 수집된 자료에 대해 이상치와 결측치를 사전 처리하였다. 결측치는 CTD 프로파일 미기록, GPS 오류, 강한 잡음으로 인한 NASC 산출 불가 등에서 발생

하였다.

학습모델파라미터 설정

본 연구에서는 성어와 치어 자료를 구분하여 각각 랜덤포레스트 회귀 모델을 구축하였다. 전체 1,186개의 원시 샘플 중 이상치 및 결측치 판정 결과, 성어 모델과 치어 모델에는 1,064개의 유효 샘플이 최종 활용되었으며, 유효 샘플에 대한 변수별 기초통계량은 다음과 같다(Table 1). 종속 변수는 NASC로, 분포의 왜도를 완화하기 위해 로그 변환 [log(NASC+1)]을 적용하였다. 설명 변수로는 수심, 수온, 염분, 조사 시기(period), 조사 해역(area), 격자 위치(GridID)를 포함하였다(Table 2).

성어 모델의 경우, 수심·수온·염분은 평균과 표준편차를 이용한 z-스코어 정규화를 적용하여 단위 차이를 보정하고 변수 중요도의 일관성을 확보하였다. 반면 치어 모델은 수온 등 물리적 변수의 절대값이 생태적 의미를 가지므로 정규화를 하지 않았다. 위·경도 좌표는 공간 변이를 반영하기 위해 격자 단위로 이산화하여 GridID를 생성하였으며, 조사 시기(early Summer, Summer, late Summer)와 해역(Cape adare, CA; Oates Land, OL; Ross Sea Polynya, RSP; Terra nova Bay Polynya, TNBP)은 범주형 변수로 인코딩하였다.

Table 1. Descriptive statistics of environmental and acoustic variables used in the NASC prediction model for juvenile Antarctic silverfish *Pleuragramma antarcticum*

	Minimum	Maximum	Mean	Standard deviation
Depth (m)	20.00	250.00	133.95	69.10
Latitude (°)	-78.20	-66.51	-73.93	2.67
Longitude (°)	-175.00	179.98	124.24	116.18
Temperature (°)	-1.99	1.43	-1.16	0.78
Salinity (psu)	31.06	34.78	34.47	0.30

NASC, Nautical area scattering coefficient. Variables include sampling depth, temperature, salinity, and NASC integrated over 250 m depth (unit, m²/nmi²)

Table 2. Predictor variables used in random forest models

Variable	Unit	Associated driver (s)
Depth	m	Aquatic (Vertical habitat)
Latitude, longitude	°N, °E	Spatial
Water temperature	°C	Aquatic
Water salinity	psu	Aquatic
Period (Early/Mid/Late Summer)	— (factor)	Seasonal
Area (CA/OL/RSP/TNBP)	— (factor)	Regional
NASC	$m^2\ nmi^{-2}$	Response

CA, Cape adare; OL, Oates land; RSP, Ross Sea Polynya; TNBP, Terra nova Bay Polynya; NASC, Nautical area scattering coefficient.

모델 학습은 MATLAB의 TreeBagger 함수를 사용하였으며, 부트스트랩 샘플링을 기본값으로 활성화하여 각 트리별 데이터 셋을 복원추출 방식으로 구성하였다. 각 결정트리는 데이터 특성에 기반하여 최대 분할 수(maximum number of splits)가 자동으로 결정되도록 설정하였으며, 최소 말단 노드 크기(minimum leaf size)를 5로 제한함으로써 과도한 분할에 따른 과적합을 방지하였다. 클래스 균형 옵션은 회귀 모델 특성상 적용되지 않았으며, 하이퍼파라미터는 그리드 서치(grid search) 기법으로 검증하였다(Scowen et al., 2021). 최종 모델은 6개 설명 변수와 1개 종속 변수(NASC)로 구성되었으며, 성어와 치어 모델은 각각 0.10° 및 0.15° 공간 해상도로 집계되었다. 각 격자 셀의 남서 모서리 좌표를 문자열로 결합하여 생성한 GridID는 범주형 설명 변수로 활용하였다.

랜덤포레스트

본 연구에서는 남극은암치의 NASC값을 예측하기 위하여 수심, 수온, 염분 등의 환경 변수와 조사 시기 및 조사해역 정보를 설명 변수로 사용하는 랜덤 포레스트 회귀모델을 구축하였다. 모델 입력은 각 측정 시점의 공간 좌표에 해당하는 격자 단위로 구성되었으며, 이를 기반으로 조사 시점의 NASC 값을 추정하였다. 예측 모델은 다음 식 (1)과 같은 구조로 표현할 수 있다.

$$Predicted\ NASC(x,y)=Model(D,T,S,P,A,G) \quad \cdots\cdots \quad (1)$$

설명 변수는 수심(D), 수온(T), 염분(S), 조사시기(P), 조사해역(A), 격자위치(Grid ID) 다섯 개이며, 종속 변수는 NASC이다. 왜도와 과첨도를 완화하기 위해 다음 식 (2)와 같이 변환하여 학습하였다.

$$y=log(NASC+1) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2)$$

연속형 변수(수심, 수온, 염분)는 훈련 자료의 평균 μ와 표준편차 σ를 이용하여 식 (3)과 같이 z-score 표준화하였다. 여기서 X_i는 i번째 관측값을 의미하며, μ_X와 σ_X는 각각 변수 X의 훈련 자료 평균과 표준편차를 나타낸다. 따라서 $X_{z,i}$는 정규화 과정을 거친 i번째 표준화 값으로, 변수 간 단위 차이를 보정하고 모델 입력 시 일관성을 확보하기 위해 사용되었다.

$$X_{z,i}=\frac{X_i-\mu_X}{\sigma_X} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (3)$$

모델의 일반화 성능을 신뢰성 있게 평가하기 위해, 훈련 데이터에서 설정된 전처리 및 파라미터를 테스트 데이터에 동일하게 적용하였다(Regier et al., 2023). 사전 실험에서 치어 자료는 연속 변수의 Z-score 표준화 시 설명력이 크게 감소하여, 최종 모델에서는 치어는 원 단위, 성어는 Z-score 표준화를 적용하였다.

격자 해상도 설정은 사전 실험을 통해 모델 성능과 표본 수

간 균형을 고려하여 최적화하였다. 성어 모델은 0.10° 해상도에서 RMSE=2.09, R^2=0.71로 가장 높은 예측력을 보였으며, 치어 모델은 0.15° 격자에서 R^2=0.61로 가장 우수한 설명력을 나타냈다. 이에 따라 각 모델에 해당 해상도를 최종 적용하였다.

또한, 특정 변수가 집중 학습되는 것을 방지하기 위해, 의사결정트리의 노드를 분할할 때 총 q개의 독립 변수 중 매번 새로 선택된 m개의 독립 변수가 노드 분할 시 고려 대상이 된다. 이때 변수의 개수 m은 일반적으로 $m \approx \sqrt{q}$또는 $m \approx log_{2}{}_{q}$로 결정된다. 랜덤포레스트 회귀 학습 시, 총 B개의 부트스트랩 샘플 집합으로부터 생성된 개별 트리 T_b는 각각의 입력값 x에 대해 예측값 $T_b(x)$을 산출한다(Lee and Lee, 2020). 이들 트리의 예측값을 평균하여 최종 예측값을 다음과 같이 결정한다(식 4).

$$\hat{f}(x)=\frac{1}{B}\sum_{b=1}^{B} T_b(x) \cdots\cdots\cdots (4)$$

본 연구에서는 NASC 값의 분포 왜도를 완화하고 정규성을 확보하기 위해 로그 변환 log(NASC+1)을 적용하였으며, 예측 결과는 다음과 같은 역변환을 통해 실제 단위(m^2/nmi^2)의 NASC 값으로 복원하였다.

$$\widehat{NASC}=exp(\hat{f}(x))-1 \cdots\cdots\cdots\cdots\cdots (5)$$

모델검증

본 연구에서는 전체 데이터를 학습용과 평가용으로 8:2 비율로 고정 분할하여 모델 성능을 평가하는 hold-out 방식(hold-out validation)을 사용하였다. 이는 일반적인 교차검증 방식에 비해 계산 부담이 적고 구현이 간단하며, 전체 데이터의 양이 비교적 충분한 본 연구 조건에서 적절한 방법으로 판단되었다.

학습은 80%의 데이터를 기반으로 수행하였으며, 나머지 20%는 독립된 테스트 세트로 활용하여MAE, RMSE, 정규화 RMSE (normalized RMSE, NRMSE), R^2를 산출하였다. 또한, 랜덤포레스트(random forest) 회귀 모델의 내부 일반화 성능을 추정하기 위해 out-of-bag 기반 평균제곱오차(OOB-MSE) 및 평균제곱근오차(OOB-RMSE)를 병행하여 산출하였다.

결 과

관측자료

전체 NASC 값은 치어에서 0.00–41.33 (m^2/nmi^2), 성어에서 0.00–222.53 (m^2/nmi^2)로 분포하였으며, 중앙값은 각각 0.018, 0.038 (m^2/nmi^2)로 나타났다. 평균값은 치어에서 0.90, 성어에서 1.91 (m^2/nmi^2)로 성어에서 더 높은 경향을 보였다. 로그 변환 [log(NASC+1)] 결과, 중앙값은 치어 0.008, 성어 0.016으로 산출되었으며, 평균값은 각각 0.279, 0.464로 계산되어 데이터의 분산이 현저히 감소하였다.

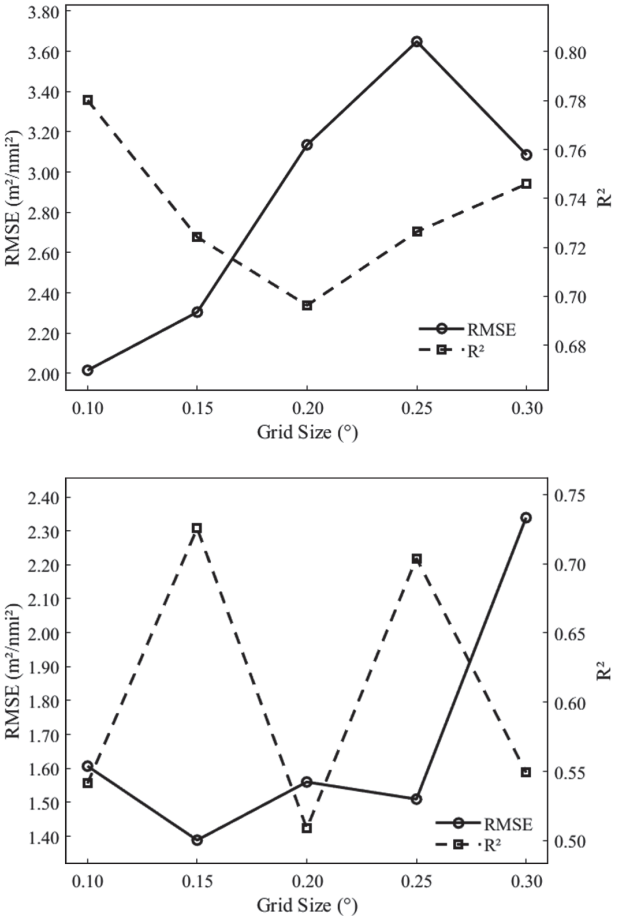

Fig. 2. Effect of spatial grid resolution on random forest model performance. a, Adult; b, Juvenile. Each line indicates RMSE (blue, left y-axis) and R^2 (red, right y-axis) across different grid sizes (0.1–0.3°).

매개변수 최적화

치어 모델의 경우, 수심·수온·염분 변수에 대해 정규화를 적용한 결과 RMSE는 2.50에서 2.64로 증가하고, 결정계수 R^2는 0.39에서 0.34로 감소하는 등 예측 성능이 저하되었다. 이에 따라 치어 모델에는 정규화를 적용하지 않고 단위를 그대로 유지하였다. 또한, 본 연구에서는 공간 해상도 설정에 따른 예측 성능의 영향을 평가하기 위해, 0.1°, 0.2°, 0.25°의 격자 단위를 대상으로 사전 실험을 수행하였다. 그 결과, 치어 모델에서는 0.15° 격자 해상도에서 가장 우수한 예측 성능(RMSE=1.38, R^2=0.72)을 보여, 해당 해상도가 최적으로 판단되었다. 반면, 성어 모델의 경우 0.10° 격자 단위에서 RMSE=2.01, R^2=0.78로 예측력이 가장 높아, 보다 세밀한 공간 구분이 성어 분포 예측에 유리한 것으로 나타났다(Fig. 2).

하이퍼파라미터 검증에는 그리드 서치(grid search) 기법을 적용하였다. 탐색 대상 파라미터는 트리 개수(nTrees)를 100,

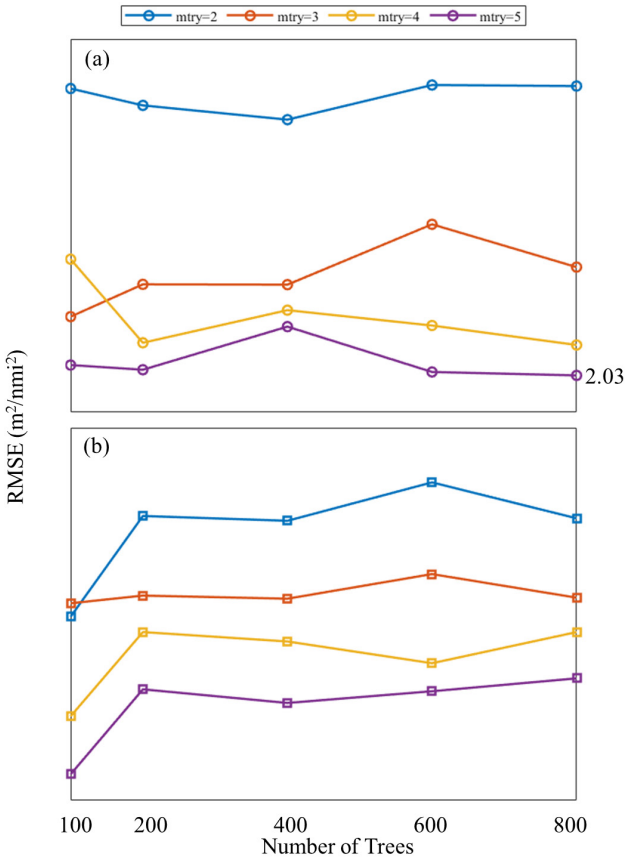

Fig. 3. RMSE variation according to the number of trees and mtry values in the random forest model for adult (a) and juvenile (b) Antarctic silverfish *Pleuragramma antarcticum*.

200, 400, 600, 800으로 분할 시 샘플링할 변수 수(mtry) 2, 3, 4, 5로 설정하였으며, 모델 학습의 재현성과 결과 일관성을 확보하기 위해 난수 생성기의 초기값을 고정하였다(Fig. 3). 데이터 분할은 한 번만 수행하여 전체 데이터의 80%를 학습용, 20%를 테스트용으로 고정하였고, 모든 하이퍼파라미터 조합 평가 및 최종 모델 검증에 동일한 학습·테스트 분할을 재사용함으로써, 모델 성능 비교 시 분할에 따른 변동성을 완전히 배제하였다.

각 조합별로 테스트 세트에 대한 RMSE를 계산한 결과, 성어 모델은 nTrees=800, mtry=5에서 2.03 m²/nmi²의 최저 RMSE를, 치어 모델은 nTrees=100, mtry=5에서 2.51 m²/nmi²의 최저 RMSE를 기록하였다. 이후, 선정된 최적 파라미터를 이용해 동일한 데이터 분할(학습 80%, 테스트 20%) 조건에서 모델을 재학습(retraining)하였으며, 재학습 결과의 RMSE 및 R^2 값은 그리드 서치 단계에서의 최저 RMSE 결과와 완전히 일치하였다.

모델적용

로그 변환된 NASC와 환경변수 간의 선형 관계를 파악하기 위하여 Pearson 상관계수 기반의 상관관계 분석을 수행하였다.

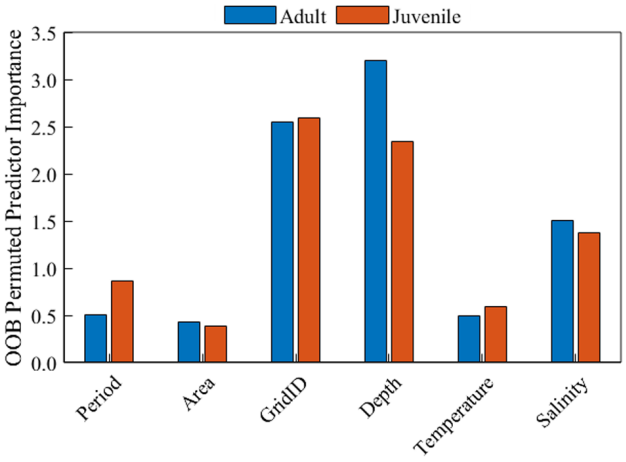

Fig. 4. Relative importance of predictor variables in the random forest models for adult and juvenile Antarctic silverfish *Pleuragramma antarcticum*.

성어와 치어 모델 모두에서 수심은 log(NASC+1)과 가장 뚜렷한 음의 상관관계를 나타냈으며, 이는 수심 증가에 따라 남극은암치의 음향산란신호가 감소하는 경향을 시사한다. 수온과 염분 역시 NASC와 음의 상관성을 보였으나, 그 절댓값은 각각 -0.19– -0.13 수준으로 상대적으로 약하였다. 이 결과는 개별 변수 단독으로는 NASC의 분포를 충분히 설명하기 어렵다는 점을 암시하며, 비선형적 상호작용을 포함하는 머신러닝 기반 분석의 필요성을 뒷받침한다. 한편, 환경 변수 간 상호 상관성은 전반적으로 낮은 수준을 유지하였으며, 설명 변수 간 다중공선성의 위험은 제한적인 것으로 판단된다.

OOB 기반 변수 중요도 분석 결과(Fig. 4), 성어 모델에서는 수심(z), 수온(z), 염분(z)이 전체 설명력의 75% 이상을 차지하는 핵심 변수로 도출되었다. 이들 세 변수는 수치적 환경(물리수괴 조건)과 직접적으로 연관되어 있어, 남극은암치 성어의 분포가 해양 물리 환경의 영향을 강하게 받는다는 점을 시사한다. 특히 수심(z)은 약 단일 변수 중 가장 높은 상대 중요도를 기록하였다. 치어 모델의 경우에도 동일한 세 연속형 변수가 주요한 예측 기여를 하였으나, 전반적으로 변수 간 기여도 분포는 보다 균일한 양상을 보였다.

반면, 조사 시기(period), 조사해역(area), 격자 위치(GridID)와 같은 범주형 변수는 성어 및 치어 모델 모두에서 상대적으로 낮은 중요도를 나타냈으며, 이는 시·공간적 차이보다는 수심·수괴 구조 등의 물리적 요인이 남극은암치 NASC의 주요 결정 요인임을 의미한다.

모델성능

최적화된 하이퍼파라미터(nTrees=800, mtry=5)를 적용한 성어(adult) 모델의 학습 및 예측 성능은 Fig. 5a에 제시된 OOB-MSE 곡선을 통해 확인할 수 있다. 트리 수 증가에 따라 OOB

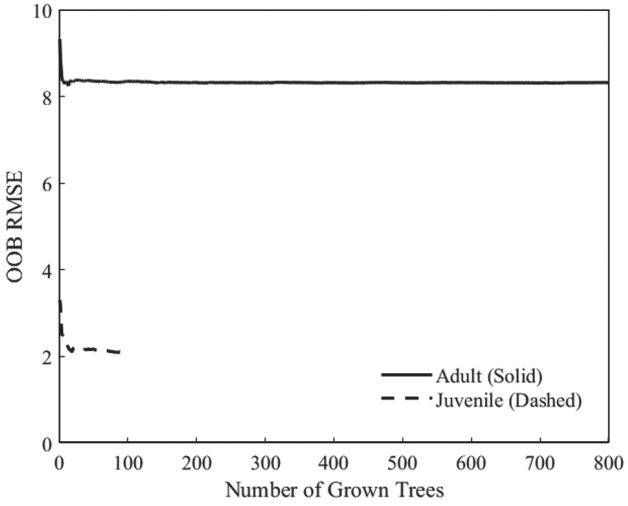

Fig. 5. Out-of-Bag mean squared error (OOB-MSE) curves by the number of trees in the random forest model for adult and juvenile Antarctic silverfish *Pleuragramma antarcticum*.

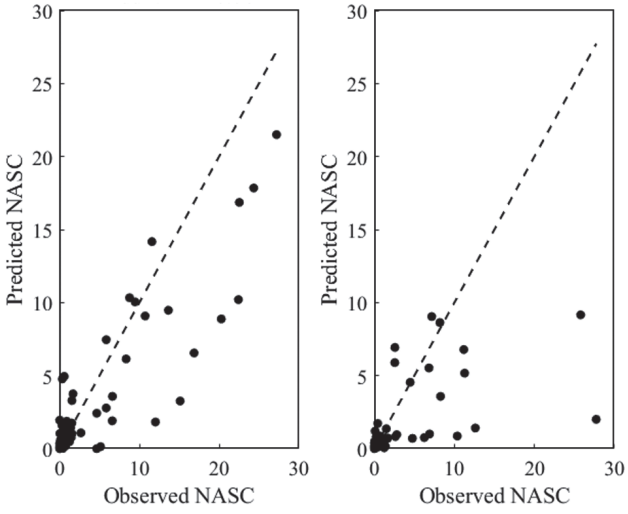

Fig. 6. Scatter plots comparing predicted and observed NASC values for adult (a) and juvenile (b) Antarctic silverfish *Pleuragramma antarcticum*. NASC, Nautical area scattering coefficient.

오차는 급격히 감소하였고, 약 100개 이상의 트리 이후에는 안정적으로 수렴하는 양상을 보여 학습 안정성과 과적합 방지 효과가 충분히 확보되었음을 시사한다. 테스트 세트에서의 예측 성능은 MAE=0.70, RMSE=2.09, NRMSE=0.07, R^2=0.76로 나타났으며, 이는 전체 NASC 분산의 약 77%를 설명할 수 있는 수준으로 매우 우수한 예측력을 갖춘 것으로 평가된다.

동일 모델의 OOB 기반 내부 검증 결과는 MSE=69.19, RMSE=8.31로 산출되었으며, 이는 로그 스케일에서 학습된 예측값을 지수 역변환하는 과정에서의 스케일 차이와, 부트스트

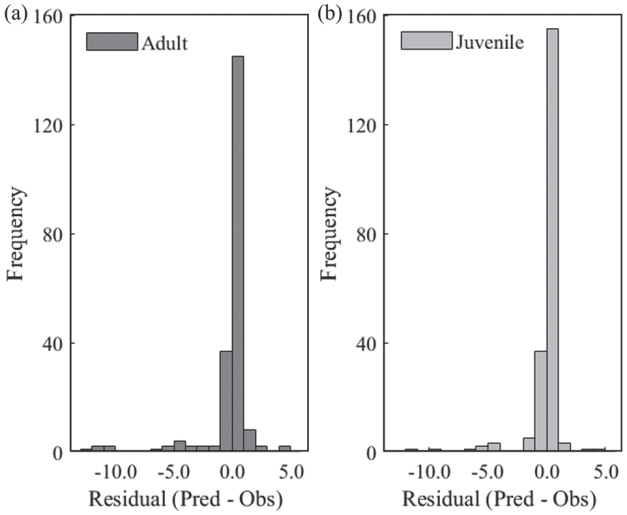

Fig. 7. Histogram of residuals (Predicted - Observed) for adult (a) and juvenile (b) Antarctic silverfish *Pleuragramma antarcticum* NASC prediction models. NASC, Nautical area scattering coefficient.

랩 샘플 분할 방식의 차이에서 기인한 것으로 해석된다. 그럼에도 불구하고 전반적인 예측 정확도와 일반화 성능은 안정적인 수준을 유지하였다. 반면, 치어(Juvenile) 모델은 nTrees=100, mtry=5로 최적화되었으며, OOB-MSE 곡선(Fig. 5b)에서도 100개 미만의 트리 수에서 빠르게 수렴하는 양상을 보여 비교적 단순한 구조로도 모델이 안정화되었음을 보여준다. 테스트 세트 기준 성능은 MAE=0.61, RMSE=2.54, NRMSE=0.09, R^2=0.37로 나타나, 전체 NASC 분산의 약 39%만을 설명하는 수준에 머물렀다. 그러나 OOB 기반 검증에서는 MSE=4.36, RMSE=2.09으로 외부 테스트 세트보다 낮은 오차를 보여 일정 수준의 일반화 성능은 유지된 것으로 분석된다.

예측값과 관측값 간의 관계를 시각화한 Fig. 6에 따르면, 성어 모델(Fig. 6a)은 대부분의 예측값이 1:1 대각선 부근에 분포하며 전반적으로 높은 적합도를 보였다. 이에 반해, 치어 모델(Fig. 6b)은 NASC가 증가함에 따라 예측값이 점차 과소 추정되는 경향이 나타났다. Fig. 7에 제시된 잔차 분포는 각 모델의 예측 안정성과 잠재적 한계를 잘 보여준다. 성어 모델(Fig. 7a)의 잔차는 대부분 -2에서 +2 사이에 집중되어 있으며, 분포 형태는 전반적으로 대칭적이고 중앙값이 0에 근접하여 예측 오차가 안정적인 특성을 보였다. 그러나 잔차의 평균값은 0이지만 분포가 왼쪽으로 다소 치우쳐 있는 비대칭성이 관찰되어, 잔차 정규성 가정에는 일정 수준의 한계가 존재함을 시사한다. 이에 따라 해석 시에는 오차의 정규분포 가정에 기반한 통계적 신뢰도 추정에 유의가 필요하다.

반면, 치어 모델(Fig. 7b)은 -5 이하의 음의 잔차가 빈번히 발생하고 있으며, 전체적으로 잔차의 분포가 좌측으로 성어 모델에 비하여 왜도되어 있다. 이는 높은 NASC 값을 가진 샘플에

서 예측이 과소 추정되는 경향을 반영하며, 모델이 이상값 또는 높은 NASC 값에 대해 과도한 민감도 또는 예측력 저하를 보임을 의미한다. 이로 인해 예측 오차의 정규성 및 등분산성 가정이 크게 위배되며, 전체 모델 안정성 측면에서도 불확실성이 더 클 수 있다.

결론적으로, 성어 모델은 전반적으로 안정적인 예측 성능과 일반화 오차 수준을 확보하여, 환경 변수 기반 NASC 예측에 있어 효과적인 접근임을 입증하였다. 반면, 치어 모델은 예측 성능이 상대적으로 제한적이었으며, 향후 모델 개선을 위해서는 생물학적 변수의 추가 확보, 이상치 제거 등의 보완이 필요할 것으로 판단된다.

고 찰

본 연구에서는 수심, 수온, 염분 등 주요 물리적 환경변수와 조사 시기 및 해역 정보를 기반으로 남극은암치(*P. antarcticum*)의 음향산란 지표인 NASC를 예측하기 위한 랜덤포레스트(random forest) 회귀 모델을 구축하였다. 예측 성능 평가 결과, 성어 모델은 R^2=0.70, RMSE=2.09로 높은 설명력을 보였으며, 변수 중요도 분석에서도 수심(z), 수온(z), 염분(z)이 주요한 기여도를 나타냈다. 이러한 결과는 남극은암치 성어의 분포가 해양의 물리적 요인, 특히 수심대 및 수괴 구조에 의해 강하게 지배된다는 점을 시사한다(O'Driscoll et al., 2011). Sun et al. (2020) 역시 해역은 다르지만 랜덤포레스트 기반 NASC 예측에서 수온과 수심이 가장 중요한 변수로 작용함을 보고하였으며, 이는 본 연구의 결과와 일치한다. 다만 본 연구의 로스해 자료에서는 상층 수온의 변화 폭이 상대적으로 제한적이며, 해빙 형성 및 수괴 구조에 의해 염분의 분포 패턴이 더 뚜렷하게 나타나 염분과 수심의 기여도가 높게 평가되었다. 또한 잔차 분포가 ±2 범위 내에 집중되어 있어 모델의 안정성과 신뢰성이 높았으며, 물리 환경만으로도 성어 분포의 주요 변동성을 충분히 설명할 수 있음을 보여준다. 이는 성어가 일정한 환경 범위 내에서 군집하며, 물리적 서식 조건의 변화에 민감하게 반응하는 종의 특성과 부합한다.

반면 치어 모델은 R^2=0.61, RMSE=2.54로 상대적으로 낮은 예측력을 보였다. 이는 치어의 분포가 물리적 변수보다 생물학적·생태적 요인의 영향을 더 크게 받기 때문으로 해석된다. 남극은암치 치어는 해빙 가장자리나 폴리냐(polynya), 해빙 하부와 같은 미세한 환경 요인에 따라 서식 밀도가 달라지며, 이 해역은 먹이생물의 분포와 해빙 구조 변화에 따라 시공간적으로 크게 변동한다(Corso et al., 2022; Lee et al., 2024). 또한 치어는 포식 회피 행동이나 해류에 의한 확산 등 행동적 요인으로 인해 국지적 환경 요인에 대한 의존도가 낮은 경향을 보였으며, 이러한 특성은 단순한 수심·수온·염분 변수로는 포착하기 어렵다. 실제로 본 연구의 치어 모델에서는 높은 NASC 구간에서 과소추정 경향이 나타났고, 잔차 분포에서도 -5 이하의 음의 잔차가 빈번히 관측되었다. 이는 치어 분포의 높은 변동성과 미세환경 의존성을 반영하며, 물리적 변수 중심의 모델링 접근만으로는 예측의 한계가 있음을 보여준다. 이러한 결과는 기존 선행연구들과도 일치한다. Sun et al. (2020)은 랜덤포레스트 기반 예측 모델을 통해 표층 수온이 NASC 변화의 주요 설명 변수로 작용하며, 수심대별로 예측력에 영향을 주는 요인이 상이하다고 보고하였다. 또한, Gadeken et al. (2021)은 캘리포니아 해류계에서 위성 기반 표층 환경 변수만으로는 NASC의 절대값을 정확히 예측하기 어렵다고 지적하였으며, 본 연구의 치어 모델에서도 유사한 과소추정 경향 및 모델 불안정성이 관찰되었다.

한편, 모델을 시기별 및 해역별로 분리 적용한 결과, 일부 구간에서는 전체 모델보다 더 높은 예측 성능을 나타냈다. 특정 조사 정선에서는 치어 모델의 R^2가 0.80 이상으로 상승하였고, 성어 모델 역시 10C 해역에서 R^2=0.79 이상을 기록하며 RMSE 수준도 낮게 나타났다. 이는 일부 시기나 공간에서는 특정 환경 요인이 국지적으로 강하게 작용하여 예측 정확도가 크게 향상될 수 있음을 보여준다. 해역별 분석 결과에서는 RSP 및 TNBP 해역에서 성어와 치어 모두 비교적 안정적인 예측 성능을 보였지만, OL 해역에서는 설명력이 음수(R^2<0)로 나타나는 등 지역 간 예측 정확도의 편차가 존재하였다. 이는 NASC 예측에서 지역별 생태 특성과 환경 이질성을 반영한 공간 맞춤형 모델링의 필요성을 시사한다.

해양생물 분포 예측에 머신러닝 기법을 적용하는 연구는 최근 들어 다양한 해양 생태계에서 확산되고 있다. 랜덤포레스트는 변수 간 비선형성과 다중공선성 문제를 효과적으로 처리할 수 있어 특히 주목받고 있으며(Islam et al., 2021; Han et al., 2022), 수온, 수심 등의 물리적 요인과 음향 자료의 복합적 관계를 반영하여 높은 예측력을 달성한 사례가 다수 보고되고 있다. 본 연구에서도 변수 상호작용을 반영한 랜덤포레스트 모델을 통해 남극은암치 성어의 분포를 높은 신뢰도로 재현할 수 있었다.

본 연구 역시 기존 회귀 기반 예측과 달리, 변수 간 상호작용을 반영한 모델링을 통해 성어의 분포를 높은 신뢰도로 재현해냈으며, 최소한의 환경 변수만으로도 일정 수준 이상의 예측이 가능함을 실증하였다. 그러나 치어와 같은 생물학적 변동성이 큰 개체군의 예측 정확도 향상을 위해서는 추가적인 설명 변수 확보가 필수적이다. 위성 기반 클로로필 농도, 해빙 농도, 플랑크톤 밀도 등 1차 생산 및 먹이 생물 관련 지표의 보완이나, 해양생물물리 모델과의 결합을 통해 미세 환경 특성을 반영하는 것이 필요하다. 또한 gradient boosting, XGBoost, 딥러닝 기반 예측 알고리즘 등과의 성능 비교를 통해 최적화된 예측 체계를 구축할 수 있을 것이다

결론적으로, 본 연구는 남극 로스해 해역에서 수심, 수온, 염분과 같은 주요 물리적 환경 변수가 남극은암치 성어의 NASC에 미치는 기여도를 분석하고, 변수 간 상호작용을 반영한 랜덤포레스트 기반 모델링이 강한 설명력을 가짐을 확인하였다. 이는 제한된 물리 환경 정보만으로도 남극은암치 분포의 주요 변동

성을 평가할 수 있음을 실증한 것으로, 극지 생물 모니터링 및 생태계 관리에 유용한 기초 자료를 제공한다. 하지만 본 연구는 주로 변수 중요도 분석과 모델 성능 평가에 초점을 두었으므로, 향후 연구에서는 공간 예측 지도를 구축하거나, 위성 데이터·해양 모형 등 외부 환경 자료를 추가로 통합하여 예측 범위를 확장할 필요가 있다. 더불어 치어와 같이 생물학적 변동성이 큰 개체군에 대해서는 클로로필 농도, 해빙 농도, 플랑크톤 분포 등 정량적 생태 지표를 추가하여 생물·환경 복합 모델링을 강화하거나 XGBoost, 딥러닝 등 다양한 기계학습 알고리즘과의 비교를 통해 최적화된 예측 체계를 마련할 수 있을 것이다.

사 사

본 과제(결과물)는 2025년도 교육부 및 부산시의 재원으로 부산RISE혁신원의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과이며(2025-RISE-02-001-006), 본 논문을 사려 깊게 검토하여 주신 심사위원님들과 편집위원님께 감사드립니다.

References

Ainley DG and Siniff DB. 2009. The importance of Antarctic toothfish as prey of Weddell seals in the Ross Sea. Antarct Sci 21, 317-327. https://doi.org/10.1017/S0954102009001953.

Ainley DG, Morandini V, Salas L, Nur N, Rotella J, Barton K and Anderson DP. 2024. Response of indicator species to changes in food web and ocean dynamics of the Ross Sea, Antarctica. Antarct Sci 36, 290-318. https://doi.org/10.1017/S0954102024000191.

Breiman L. 2001. Random forests. Mach Lear 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Corso AD, Steinberg DK, Stammerjohn SE and Hilton EJ. 2022. Climate drives long-term change in Antarctic silverfish along the western Antarctic Peninsula. Commun Biol 5, 104. https://doi.org/10.1038/s42003-022-03042-3.

Gadeken KR, Joseph MB, McGlinchy J, Karnauskas KB and Wall CC. 2021. Predicting subsurface sonar observations with satellite-derived ocean surface data in the California Current Ecosystem. PLoS One 16, e0248297. https://doi.org/10.1371/journal.pone.0248297.

Han H, Yang C, Zhang H, Fang Z, Jiang B, Su B, Sui J, Yan Y and Xiang D. 2022. Environment variables affect CPUE and spatial distribution of fishing grounds on the light falling gear fishery in the northwest Indian Ocean at different time scales. Front Mar Sci 9, 939334. https://doi.org/10.3389/fmars.2022.939334.

Islam MM, Kashem MA and Uddin J. 2021. Fish survival prediction in an aquatic environment using random forest model. Int J Artif Intell 10, 614-622. https://doi.org/10.11591/ijai.v10.i3.pp614-622.

Lee D and Lee S. 2020. Hourly prediction of particulate matter (PM2.5) concentration using time series data and random forest. KIPS Trans Softw Data Eng 9, 129-136. https://doi.org/10.3745/KTSDE.2020.9.4.129.

Lee S, Oh W, La HS, Son W, Kim JH and Lee K. 2024. Spatiotemporal distribution of Antarctic silverfish in the Ross Sea, Antarctica. Fishes 9, 47. https://doi.org/10.3390/fishes9020047.

O'Driscoll RL, Macaulay GJ, Gauthier S, Pinkerton M and Hanchet S. 2011. Distribution, abundance and acoustic properties of Antarctic silverfish (*Pleuragramma antarcticum*) in the Ross Sea. Deep Sea Res II Top Stud Oceanogr 58, 181-195. https://doi.org/10.1016/j.dsr2.2010.05.018.

Regier PJ, Ward ND, Myers-Pigg AN, Grate J, Freeman MJ and Ghosh RN. 2023. Seasonal drivers of dissolved oxygen across a tidal creek-marsh interface revealed by machine learning. Limnol Oceanogr 68, 2359-2374. https://doi.org/10.1002/lno.12426.

Scowen M, Athanasiadis IN, Bullock JM, Eigenbrod F and Willcock S. 2021. The current and future uses of machine learning in ecosystem service research. Sci Total Environ 799, 149263. https://doi.org/10.1016/j.scitotenv.2021.149263.

Sun M, Cai Y, Zhang K, Zhao X and Chen Z. 2020. A method to analyze the sensitivity ranking of various abiotic factors to acoustic densities of fishery resources in the surface mixed layer and bottom cold water layer of the coastal area of low latitude: A case study in the northern South China Sea. Sci Rep 10, 11128. https://doi.org/10.1038/s41598-020-67387-7.

Vacchi M, La Mesa M, Dalù M and Macdonald J. 2004. Early life stages in the life cycle of Antarctic silverfish, *Pleuragramma antarcticum* in Terra Nova Bay, Ross Sea. Antarct Sci 16, 299-305. https://doi.org/10.1017/S0954102004002135.

Woo SY, Jung CG, Kim JU and Kim SJ. 2018. Assessment of climate change impact on aquatic ecology health indices in Han river basin using SWAT and random forest. J Korea Water Resour Assoc 51, 863-874. https://doi.org/10.3741/JKWRA.2018.51.10.863.