Journal of the Korean Society of Marine Environment & Safety
Research Paper
Vol. 29, No. 1, pp. 024-035, February 28, 2023, ISSN 1229-3431(Print) / ISSN 2287-3341(Online) https://doi.org/10.7837/kosomes.2023.29.1.024

인공지능 기법을 활용한 한반도 해역의 수질평가지수 예측모델 개발

김성수* · 손규희** · 김도연*** · 허장무**** · 김성은*****†
* ㈜아라종합기술 대리, ** ㈜세광종합기술단 실장, *** ㈜아라종합기술 이사,
**** ㈜아라종합기술 사원, ***** ㈜아라종합기술 대표이사

Development of a Water Quality Indicator Prediction Model for the Korean Peninsula Seas using Artificial Intelligence

Seong-Su Kim* · Kyuhee Son** · Doyoun Kim*** · Jang-Mu Heo**** · Seongeun Kim*****†
* Assistant Manager, ARA consulting & Technology Ltd, Incheon 21990, Korea
** Executive Director, SEKWANG Engineering consultants co., Ltd, Seoul 04521, Korea
*** Director, ARA consulting & Technology Ltd, Incheon 21990, Korea
**** Staff, ARA consulting & Technology Ltd, Incheon 21990, Korea
***** CEO, ARA consulting & Technology Ltd, Incheon 21990, Korea

요 약 : 급격한 산업화와 도시화로 인해 해양 오염이 심각해지고 있으며, 이러한 해양 오염을 실효적으로 관리하기 위해 수질평가지수(Water Quality Index, WQI)를 마련하여 활용하고 있다. 하지만 수질평가지수는 다소 복잡한 계산과정으로 인한 정보의 손실, 기준값 변동, 실무자의 계산오류, 통계적 오류 등의 불확실성(uncertainty)을 내포하고 있다. 이에 따라 국내·외에서 인공지능 기법을 활용하여 수질평가지수를 예측하기 위한 연구가 활발히 이루어지고 있다. 본 연구에서는 해양환경측정망 자료(2000 ~ 2020년)를 활용하여 우리나라 전 해역 즉, 5개의 생태구에 대한 WQI를 추정할 수 있는 가장 적합한 인공지능기법을 도출하기 위해 총 6가지의 기법(RF, XGBoost, KNN, Ext, SVM, LR)을 실험하였다. 그 결과, Random Forest 기법이 다른 기법에 비해 가장 우수한 성능을 보였다. Random Forest 기법의 WQI 점수 예측값과 실제값의 잔차 분석 결과, 모든 생태구에서 시간적 및 공간적 예측 성능이 우수한 것으로 나타났다. 이를 통해 본 연구에서 개발한 Random Forest 기법은 높은 정확도를 바탕으로 우리나라 전해역에 대한 WQI를 예측 가능할 것으로 사료된다.

핵심용어 : 수질평가지수, 인공지능 기법, 기계학습, 해양환경, 생태구

Abstract : *Rapid industrialization and urbanization have led to severe marine pollution. A Water Quality Index (WQI) has been developed to allow the effective management of marine pollution. However, the WQI suffers from problems with loss of information due to the complex calculations involved, changes in standards, calculation errors by practitioners, and statistical errors. Consequently, research on the use of artificial intelligence techniques to predict the marine and coastal WQI is being conducted both locally and internationally. In this study, six techniques (RF, XGBoost, KNN, Ext, SVM,* and *LR) were studied using marine environmental measurement data (2000–2020) to determine the most appropriate artificial intelligence technique to estimate the WOI of five ecoregions in the Korean seas. Our results show that the random forest method offers the best performance as compared to the other methods studied. The residual analysis of the WQI predicted score and actual score using the random forest method shows that the temporal and spatial prediction performance was exceptional for all ecoregions. In conclusion, the RF model of WQI prediction developed in this study is considered to be applicable to Korean seas with high accuracy.*

Key Words : *Water Quality Index, Artificial Intelligence, Machine Learning, Marine Environments, Ecoregion*

* First Author : seongsukim@aracnt.com, 070-7585-5432
† Corresponding Author : godhkim@naver.com, 070-7585-5432

1. 서 론

우리나라는 급격한 산업화와 도시화가 이뤄지며, 그에 따른 연안 및 항만의 개발로 인해 도시주변에 위치한 연안 해역의 수질오염이 심화되며 해양환경관리에 대한 필요성이 대두되고 있다. 이에 따라 해양 오염을 효율적으로 관리하고 해양환경과 해양생태계를 보전·복원하기 위해서 국가 차원에서 해양 수질을 지속적으로 모니터링하고 있다(Kim et al., 2022). 전국 연안의 해양환경 상태를 정기적으로 조사하고 해양환경실태를 종합적으로 파악하여 해양환경정책수립의 기본 자료로 활용하기 위해 해양환경공단(Korea marine environment management, KOEM)에서는 1997년부터 측정을 시작하였으며, 2010년 종합모니터링 체제가 구축되어 현재 항만환경측정망(50개 정점), 하천영향 및 반폐쇄성환경측정망(230개 정점), 연안해역환경측정망(145개 정점)에서 계절별(2월, 5월, 8월, 11월)로 조사를 수행하고 있다.

연안 및 해양 환경을 포괄적으로 이해하기 위해 조사 항목 중 일부를 활용하여 수질평가지수(Water Quality Index, WQI)를 통해 수질 등급을 결정한다. 현재 국내 해양은 연안해역과 근해역을 분류인자(수심, 해류, 탁도, 조위차, 기후)에 따라 서해 중부, 서남해역, 대한해협, 동해, 제주로 크게 5가지 생태구역으로 분류하여 관리하고 있다. 해양수산부고시 제2018-10호「해양환경관리법」에 따라 여러 수질 항목(저층 용존산소포화도, 표층 용존무기질소, 표층 용존무기인, 표층 식물플랑크톤, 투명도)을 생태구별로 다른 기준 값을 적용하여 WQI 값을 산출하고 있다(Table 1).

Table 1. Reference value of each parameters to calculate Water Quality Index (WQI)

Eco-Region	Chl-a (μg/L)	DO_{sat} (%)	DIN (μg/L)	DIP (μg/L)	Transparency (m)
Middle West Sea	2.2	90	425	30	1.0
Southwest Sea	3.7		230	25	0.5
Straits of Korea	6.3		220	35	2.5
East Sea	2.1		140	20	8.5
Jeju	1.6		165	15	8.0

일반적으로 WQI를 산출하는 방법은 4단계를 통해 이뤄진다. 먼저, 분석에 사용할 수질항목을 선정한다. 두 번째는 각 항목의 자료를 단위가 없는 보조 지수(sub-index)로 변환한다. 세 번째는 각 인자에 대하여 가중치를 산출한다. 마지막으로 각 수질 인자들을 집계 함수(Aggregation Function)를 활용하여 합산 후 WQI 점수에 따라 등급화 한다(Uddin et al., 2021). 현재 해양환경관리법에서 고시하는 WQI 계산법도 이러한 프로세스를 따른다(Rho et al., 2012). 먼저 전문가의 의견에 따라 부영양화의 원인 항목인 용존무기질소(DIN)와 용존무기인(DIP), 일차반응항목인 클로로필(Chl-a)과 투명도, 이차반응항목인 저층용존산소 포화도가 수질 분석 항목으로 선정되어있다. 각 항목들은 생태구 해역별 통계분석을 통해 정해진 기준값에 따라 1~5 사이의 보조 지수(Sub-Index, SI)로 환산된다(Table 2).

Table 2. Sub-index scoring equations for each parameters

Sub-index score	Chl-a(μg/L), DIN(μg/L), DIP (μg/L)	DO_{sat}(%), Transparency(m)
1	≤ reference value(RV)	≥ RV
2	< RV + (0.10×RV)	> RV − (0.10×RV)
3	< RV + (0.25×RV)	> RV − (0.25×RV)
4	< RV + (0.50×RV)	> RV − (0.50×RV)
5	≥ RV + (0.50×RV)	≤ RV − (0.50×RV)

WQI 점수는 항목별 보조 지수에 가중선형합산법(weighted linear combination method)를 통해 설정된 가중치를 곱하여 산출되며(Eq.1), 도수분포 특징에 따라 정해진 1~5단계의 등급으로 평가하게 된다(Table 3).

$$WQI = 10 \times DO_{SI} + 6 \times [(Chl\text{-}a_{SI}) + (Trans_{SI})]/2 + 4 \times [(DIN_{SI}) + (DIP_{SI})]/2 \quad (1)$$

Table 3. WQI score criteria to evaluate water quality status

WQI grade	I (Very Good)	II (Good)	III (Moderate)	IV (Bad)	V (Very Bad)
WQI score	≤ 23	23-34	34-46	47-59	≥ 60

이처럼 여러 단계를 거치는 복잡한 계산과정은 정보의 손실, 해양환경이 변화함에 따른 기준값 변동, 실무자의 계산오류, 등급 간의 근소한 점수차로 인한 통계적 오류 등의 불

확실성(uncertainty)을 야기한다(Uddin et al., 2021).

이에 따라 최근 국내외에서는 인공지능(artificial intelligence, AI)을 활용한 WQI 예측 모델 개발과 같이 기존 WQI 계산에서 나타나는 불확실성을 보완하기 위해 다양한 연구가 시도되었다. 먼저 국내 연구를 살펴보면, Jang et al.(2016)은 국내 현장 관측 자료와 위성자료(GOCI)에서 측정되는 반사도와 수질항목 산출물을 이용하여 기계학습 기반의 수질평가지수 추정 기법을 개발하였고 준수한 성능이 나타났지만, 저층 산소포화도는 추정할 수 없다는 한계점이 있다. Jeon et al.(2020)은 Random Forest(RF) 및 Support vector machine(SVM)을 통해 광양만의 WQI를 추정하였고, RF와 SVM은 F1 score 95 % 이상의 높은 예측 성능을 보였다. Kim et al.(2022)은 분류모델인 AdaBoost와 TPOT 알고리즘을 활용하여 시화호에서 WQI 등급을 예측하여 F1-score로 평가한 결과, 1~2등급에서는 높은 분류 성능(90 % 이상)을 보였지만 3~4등급에서는 낮은 분류 성능(70 % 이하)을 보였다. 국외에서는 좀 더 다양한 머신러닝 기법이 연구되었다. Abba et al.(2020)의 연구에서는 인도에 위치한 Yamuna강의 3개 정점에서 Back Propagation Neural Network(BPNN), Adaptive Neuro-Fuzzy Inferecne System (ANFIS), Support Vector Regression(SVR), Multi-Linear Regression (MLR) 등 여러 가지 인공지능 기법을 사용하여 WQI를 예측하였으며, 예측 성능을 향상시키기 위해 앙상블 기법(Neural Network Ensemble, NNE)을 활용하였다. Uddin et al.(2022b)은 Random Forest(RF), Decision Tree(DT), K-Neareset Neighbors (KNN), Extreme Gradient Boosting(XGB), Extra Tree(Ext), SVM, Linear Regression(LR), Gauusian Naive Bayes(GBN) 등 여러 가지 머신러닝 기법을 비교한 결과, 트리 기반의 DT, Ext와 앙상블 기반인 RF, XGB가 WQI의 예측 성능이 가장 뛰어난 것으로 나타났다. Gaya et al.(2020)은 인공신경망(ANN)과 ANFIS을 활용하여 WQI를 예측한 결과 MLR 보다 10 % 이상의 성능이 개선된 연구결과를 도출하였다.

이처럼 국내외로 다양한 인공지능 기반의 WQI 예측 방법이 연구되고 있지만, 국내 연구에서는 광양만, 시화호 등 특정 지역에 국한되어 연구된 바(Jeon et al., 2016; Kim et al., 2022), 우리나라 전 해역 대상으로 적용하기에는 어려운 실정이다. 따라서 본 연구에서는 우리나라 전 해역인 5개의 생태구(서해 중부, 서남해역, 대한해협, 동해, 제주)의 WQI를 추정할 수 있는 인공지능 기법을 개발하기 위해 6가지 머신러닝 기법(RF, XGB, Ext, SVM, LR)을 평가하고, 그 중 성능이 가장 우수한 기법의 예측값과 실측값의 편차를 시공간적으로 분석하여 WQI를 예측하기 가장 적합한 모델을 제시하고자 한다.

2. 재료 및 방법

2.1 분석항목

본 연구에서는 해양수산부 해양환경정보포털(https://www.meis.go.kr)에서 제공하는 해양환경측정망 자료를 사용하였으며, 생태구의 일반적인 특성을 반영하기 위해서 2000년 1월부터 2020년 12월까지 수집 가능한 모든 자료를 분기별(2, 5, 8, 11월)로 수집하였다. WQI 예측 모델 학습에 사용한 분석 항목으로 기존 해양환경관리법에서 WQI 산출에 사용하는 투명도(transparency), 용존산소농도(dissolved oxygen), 용존무기질소(dissolved inorganic nitrogen, DIN), 용존무기인(dissolved inorganic phosphate, DIP), 클로로필(chlorophyll a)을 채택하였다. 추가로 전 세계의 WQI 산출 모델들이 사용하는 분석항목 중 빈도가 가장 높은 항목으로 수온, 수소이온농도(pH), 용존산소농도(DO), 생물학적산소요구량(biological oxygen demand, BOD), 분원성 대장균군(faecal coliform, FC), 총용존고형물(total dissolved solid, TDS), 부유물질(Suspended Solid, SS), 탁도(Turbidity), 암모니아성질소(NH_3-N)를 선정하였는데(Uddin et al., 2021), 그 중 해양환경측정망에서 활용 가능한 수온, 수소이온농도, 부유물질을 본 연구의 분석항목으로 추가하였다. NH_3-N의 경우 해양환경측정망 자료에서 활용 가능하지만 기존 분석 항목인 DIN에 이미 포함되어 분석에서 제외하였다. 수집된 총 26,974개 자료를 모델 학습(70 %)과 검증(30 %)으로 구분하여 적용하였다.

2.2 인공지능 알고리즘

최근 기술이 발전함에 따라 여러 분야의 연구에서 인공지능을 통한 데이터 분석이 이뤄지고 있다. 그 중 수질환경 예측 분야에서도 통계 등을 활용한 전통적인 방법보다 기계학습이 더 효과적인 것으로 알려져 있다(Uddin et al., 2022a). 따라서 본 연구에서는 기존 연구들에서 효과적인 WQI 예측 성능을 보인 6가지 인공지능 모델(Random Forest, XGBoost, KNN, Ext, SVM, LR)을 구축하고, 비교 분석을 통해 한반도 내 WQI 산출에 가장 적합한 인공지능 모델을 개발하고자 한다.

먼저 Random forest(RF)는 의사결정나무(decision tree) 기반의 앙상블 모델로, 부트스트랩(bootstrap) 샘플링을 통해 다양한 입력 자료를 생성하고, 학습 과정에서 다수의 의사결정나무를 만들어 그 평균 예측치를 산출하는 방식이다. 분류(classification)와 회귀(regression) 분석에서 모두 사용할 수 있으며, 자세한 알고리즘은 Liaw and Wiener(2002)에 되어 있다.

Extra Tree(Ext) 또한 의사결정나무 기반의 앙상블 모델이지만, RF와 달리 각 결정나무를 만들 때 모든 학습 자료를

Table 4. Optimized hyper-parameters for each models

Model parameters	RF	Ext	XGB	KNN	SVM	LR
n_estimators	500	500	-	-	-	-
n_neighbors	-	-	-	5	-	-
max_features	8	8	-	-	-	-
max_depth	-	-	11	-	-	-
min_child_weight	-	-	6	-	-	-
subsample	-	-	1	-	-	-
colsample_bytree	-	-	1	-	-	-
eta	-	-	0.05	-	-	-
weights	-	-	-	‘distance’	-	-
metric	-	-	-	‘manhattan’	-	-
C	-	-	-	-	5	-
kernel	-	-	-	-	‘rbf’	-
epsilon	-	-	-	-	0.2	-

사용하는데, 이는 부트스트랩된 학습자료를 통해 분할의 가장 적합한 특성을 찾는 RF와 가장 구분되는 특징이다. 결정나무에서 특성을 무작위로 분할하게 되면 성능은 저하된다는 단점이 있지만, 더 많은 결정나무를 앙상블하기 때문에 과적합(over-fitting)을 방지하는 효과가 있다. 자세한 알고리즘은 Hannan and Anmala(2021)에서 설명하고 있다.

의사결정나무 기반의 앙상블 기법 중 Gradient Boosting은 잔차(Residual)를 이용하여 이전 모형의 약점을 보완한 새로운 모형으로 순차적으로 업데이트 시키는 기법이다. 이는 구현이 쉽고 정확도가 높지만, 과적합에 취약하다는 단점이 있다. 이를 병렬 처리, 과적합 규제, CART 앙상블 사용 등의 기능을 통해 개선한 모델이 Extreme Gradient Boosting(XGBoost)이다. XGBoost는 특히 회귀 분석에서 성능이 우수한 것으로 알려져 있다(Huan et al., 2020; Khan et al., 2021; Tanha et al., 2020).

K-Nearest Neighbors(KNN)는 가장 일반적으로 많이 쓰이는 지도학습 알고리즘으로 데이터들 간의 거리를 측정하여 서로 가까운 K개의 데이터를 하나의 그룹으로 분류하여 예측값을 추정한다. 본 연구에서는 가장 보편적으로 사용되는 멘해튼 거리(Manhattan Distance)를 사용하여 WQI 예측값을 산출하였다. 자세한 알고리즘은 Modaresi and Araghinejad (2014)에 설명되어 있다.

Support Vector Machine(SVM)은 많이 쓰이는 기계학습 방법 중 하나로 입력 자료들을 가장 이상적으로 구별하는 초평면(hyper plane)을 구하는 방법이다(Mountrakis et al., 2011; Jang et al., 2016). 이때 보다 효율적인 초평면 탐색을 위해 학습자료를 고차원으로 변화시켜주는 커널함수를 활용할 수 있다. 커널 함수는 linear, polynomial, Gaussian, sigmoid, spectral angle 등이 있다(Kim et al., 2014).

선형 회귀(Linear Regression, LR)은 가장 일반적으로 알려진 기계학습 기법으로 여러 연구에서 WQI를 예측하는데 사용되어 왔다(Kadam et al., 2019; Grbčić et al., 2022). 이와 같이 다양한 인공지능 기법을 활용하여 국내 수질환경에 가장 적합한 알고리즘을 찾아 WQI 예측에 활용하고자 한다.

2.3 모델 학습 및 성능 평가 기법

하이퍼파라미터는 최적의 훈련모델을 구현하기 위해 모델에 설정하는 변수로 각 인공지능 기법과 활용되는 자료마다 세팅해야 하는 값이 모두 다르다. 본 연구에서는 각 모델의 최적 하이퍼파라미터를 찾기 위해 Grid Search 교차 검증 방법을 채택하였다, Grid Search 교차 검증은 각 모델에서 설정해야하는 하이퍼파라미터를 순차적으로 변경해가며 모델 평가를 수행하고, 그 중 성능이 가장 좋은 하이퍼파라미터 조건을 도출하는 경험적 방법이다. 이를 통해 도출된 각 모델의 최적 하이퍼파라미터는 Table 4와 같다.

최적 하이퍼파라미터를 활용하여 생태구역별로 각각 6가지 머신러닝 기법(RF, XGB, KNN, Ext, SVM, LR)에 대하여 모델 학습을 진행하였다. 모델의 예측 성능 평가에는 학습된 각 모델에 검증 자료를 대입하여 산출된 예측값(predicted value)과 실제값(actual value)을 비교하는 여러 가지 평가지표

를 활용하였다. 각 모델의 예측 성능을 검증하기 위한 평가지표로는 평균 절대 편차(mean absolute error, MAE), 제곱근 편차 (mean squared error, MSE), 평균 제곱근 편차(root mean square error, RMSE)를 사용하였다. 각 평가지표는 식(2)~(4)을 통해 도출하였다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \quad (3)$$

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i - x_i)^2} \quad (4)$$

여기서, x_i는 실제값, y_i는 예측값, n은 데이터 개수이다.

3. 결과 및 고찰

3.1 모델별 예측 성능 평가

3.1.1 모델별 WQI 예측 성능 비교

본 연구에서는 20년(2000~2020)간 측정된 해양환경측정망 자료를 활용하여 6개의 머신러닝 기법(RF, XGB, KNN, Ext, SVM, LR)의 WQI 예측 성능을 생태구별로 비교 및 평가하였다. 모델의 성능을 평가하기 위해서 평가지표(MAE, MSE, RMSE)를 활용하여 교차 검증을 수행하였다. Fig. 1과 Table 5는 생태구별 각 모델의 교차 검증(MAE, MSE, RMSE) 및 결정계수 결과를 나타낸다. 교차검증 결과 3가지 평가지표에서 모두 RF의 예측 성능이 가장 높았고, 다음으로는 XGB, Ext가 높은 예측 성능을 보였으며, KNN, SVM, LR에서 상대적으로 낮은 예측 성능을 보였다. Fig. 1에서 결정계수(R^2)는 회귀분석에서 하나의 변수로 설명되는 종속변수의 분산(variance)의 정도 혹은 추정회귀식의 적합도를 의미하며, 1에 가까울수록 회귀모형이 잘 추정되었음을 의미한다. 결정계수는 RF에서 0.99로 가장 높게 나타났으며, Ext와 XGB에서 각각 0.98과 0.96로 높은 상관성을 보였다. 반면 KNN과 SVM에서는 결정계수가 각각 0.74와 0.64로 상대적으로 낮은 상관성을 보였으며, LR에서 가장 낮은 결정계수(0.59)가 나타났다. 모델 성능과 결정계수가 높았던 RF, XGB, Ext는 모두 의사결정나무 기반의 앙상블 모델로 의사결정나무에서 나온 결과 값을 평균하여 사용하기 때문에 예측 성능이 뛰어난 것으로 알려져 있고, 이는 과거 WQI 예측 연구에서도 결정나무기반의 모델의 WQI 예측 성능이 좋았다는 다수의 연구 결과와 일치한다(Bui et al., 2020; Grbčić et al., 2022; Haghiabi et al., 2018; Khan et al., 2021; Khullar and Singh, 2021; Uddin et al., 2022b).

3.1.2 생태구별 WQI 예측 성능 비교

본 연구에서 개발한 각 모델의 생태구별로 예측 성능을 비교하였다(Table 5). RF는 대한해협에서 평가 점수가 가장 높았으며, 제주와 동해에서 예측 성능이 가장 낮았다. XGB는 제주에서 예측 성능이 가장 높았고 대한해협과 동해에서 예측 성능이 낮은 것으로 나타났다. KNN에서는 제주에서 예측 성능이 가장 높았고 서해 중부에서 예측 성능이 가장 낮았다. Ext는 대한해협에서 평가 점수가 가장 높았고, 서해 중부에서 예측 성능이 가장 낮았다. SVM은 제주와 서남해역에서 예측 성능이 가장 높았고, 서해 중부에서 예측 성능이 가장 낮았다. LR은 제주에서 예측 성능이 가장 높았고, 서해 중부에서 예측 성능이 가장 낮았다.

위의 결과를 바탕으로 서해 중부가 대부분 모델(KNN, Ext, SVM, LR)에서 예측 성능이 가장 낮은 해역으로 나타났고, 제주해역이 대부분의 모델(XGB. KNN, SVM, LR)에서 예측 성능이 좋은 해역으로 나타났다.

생태구별로 예측 성능이 다르게 나타나는 이유를 확인하기 위해 생태구별 WQI 자료의 분포를 확인하였고(Fig. 2), WQI 등급은 Table 3의 기준을 적용하였다. 먼저 우리나라 전 연안의 자료 분포를 살펴보면 WQI 등급이 높아질수록 자료의 수가 적어지는 경향을 보이고, WQI 점수의 최빈값은 20.0으로 1등급 자료의 비율이 높은 정규분포(Normal Distribution)를 보인다. 생태구별로는 예측 성능이 가장 높았던 제주해역을 비롯한 서남해역, 대한해협, 동해에서는 전 연안의 자료와 비슷한 정규분포를 보였다. 하지만, 서해 중부는 1등급(30.8%)과 2등급(38.2%)이 최빈값을 갖는 쌍봉분포(bimodal distribution)로서 자료의 불균형(imbalanced data)을 보였다. 일반적으로 기계학습 방법에서는 대부분 균형 잡힌 데이터 세트를 기반으로 했을 때 전체적인 정확도가 높게 나타나는 것으로 알려져 있다(Mi, 2013). 데이터 불균형의 문제점은 예측 알고리즘이 종종 다수의 클래스에 편향되어 있어 소수의 클래스에 대한 오분류율을 더 높이기 때문이다(López et al., 2013). 따라서 대부분 모델(KNN, Ext, SVM, LR)에서 비균질한 자료 분포를 보인 서해 중부에서 예측 성능이 낮게 나타난 것으로 사료된다. 하지만 모델 평가에서 예측 성능이 가장 좋은 Random Forest는 서해 중부의 평가 지표가 다른 생태구와 비슷하게 나타나, 자료 분포에 영향을 크게 받지 않은 것으로 판단된다(Table 5).

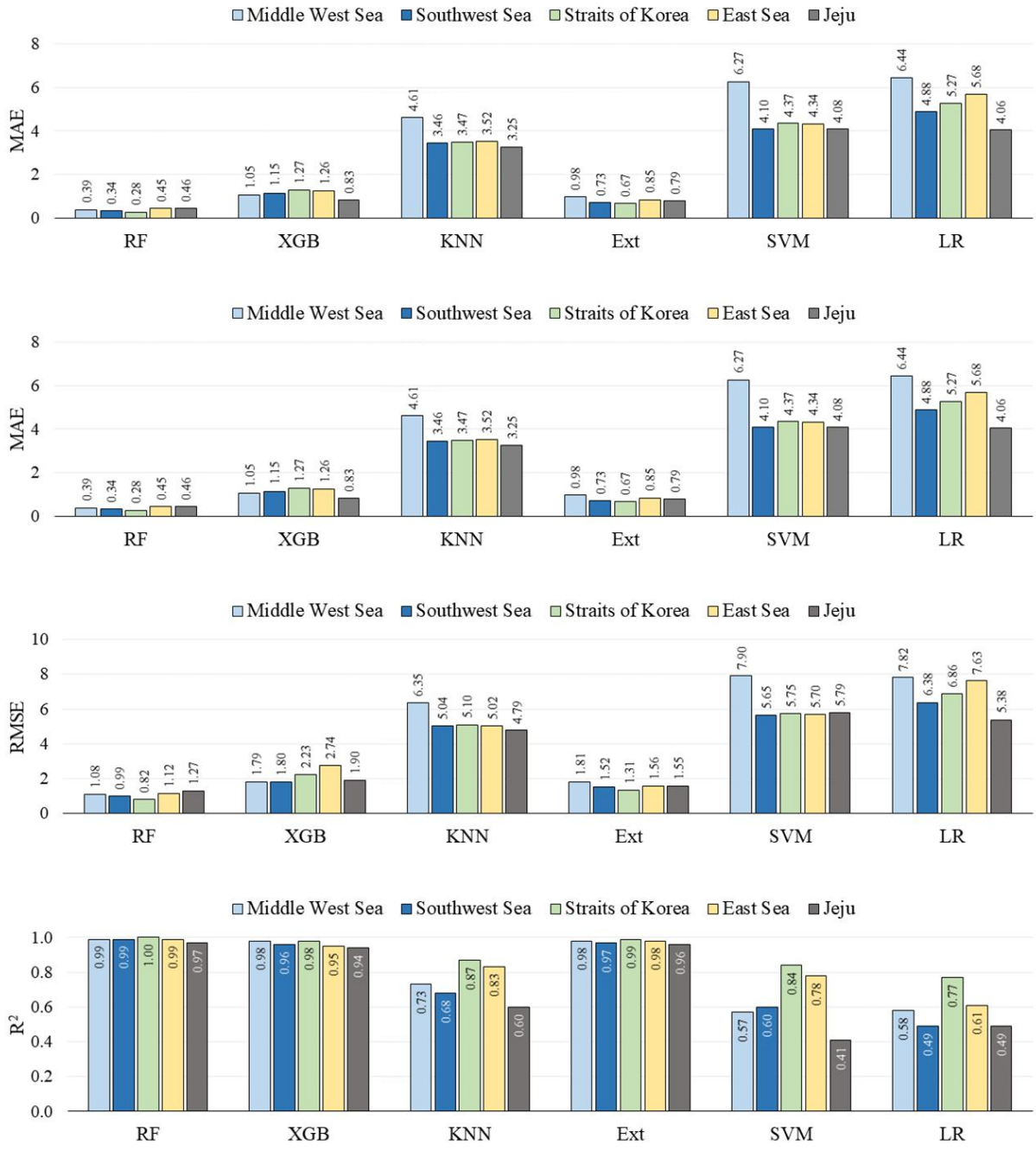

Fig. 1. Comparison of evaluation results from AI algorithms at each ecoregion.

Random Forest는 여러 번의 복원 추출한 자료를 사용하여 구축된 예측 모델들을 평균하는 배깅(bagging) 알고리즘으로 예측 모형간의 분산이 작아지기 때문에 자료 불균형의 영향을 적게 받게 된다. 이러한 알고리즘 때문에 Random Forest는 서해 중부에서도 예측 성능이 우수한 것으로 사료된다. 이에 따라 자료의 분포가 다양한 우리나라 전 해역을 예측하기에 가장 적합한 모델을 Random Forest로 선정하였고, Random Forest의 WQI의 예측값과 실제값을 비교 분석하여 WQI 예측 특성을 자세하게 파악하였다.

Table 5. 5-fold cross-validation results of various AI algorithms

Model	Region	MAE	MSE	RMSE	R^2	Rank
RF	Middle West Sea	0.40	1.16	1.08	0.99	3
	Southwest Sea	0.34	0.99	0.99	0.99	2
	Straits of Korea	0.28	0.67	0.82	1.00	1
	East Sea	0.45	1.26	1.12	0.99	4
	Jeju	0.46	1.6	1.27	0.97	5
	Mean	0.39	1.14	1.06	0.99	-
XGB	Middle West Sea	1.05	3.19	1.79	0.98	2
	Southwest Sea	1.15	3.23	1.80	0.96	3
	Straits of Korea	1.27	4.98	2.23	0.98	4
	East Sea	1.26	7.48	2.74	0.95	5
	Jeju	0.83	3.59	1.90	0.94	1
	Mean	1.11	4.49	2.09	0.96	-
KNN	Middle West Sea	4.61	40.27	6.35	0.73	5
	Southwest Sea	3.46	25.40	5.04	0.68	2
	Straits of Korea	3.47	26.01	5.10	0.87	3
	East Sea	3.52	25.18	5.02	0.83	4
	Jeju	3.25	22.92	4.79	0.60	1
	Mean	3.66	27.96	5.26	0.74	-
Ext	Middle West Sea	0.98	3.26	1.81	0.98	5
	Southwest Sea	0.73	2.32	1.52	0.97	2
	Straits of Korea	0.67	1.72	1.31	0.99	1
	East Sea	0.85	2.43	1.56	0.98	4
	Jeju	0.79	2.42	1.55	0.96	3
	Mean	0.80	2.43	1.55	0.98	-
SVM	Middle West Sea	6.27	62.39	7.90	0.57	5
	Southwest Sea	4.10	31.89	5.65	0.60	2
	Straits of Korea	4.37	33.10	5.75	0.84	4
	East Sea	4.34	32.44	5.70	0.78	3
	Jeju	4.08	33.52	5.79	0.41	1
	Mean	4.63	38.67	6.16	0.64	-
LR	Middle West Sea	6.44	61.19	7.82	0.58	5
	Southwest Sea	4.88	40.72	6.38	0.49	2
	Straits of Korea	5.27	47.05	6.86	0.77	3
	East Sea	5.68	58.27	7.63	0.61	4
	Jeju	4.06	28.98	5.38	0.49	1
	Mean	5.27	47.24	6.81	0.59	-

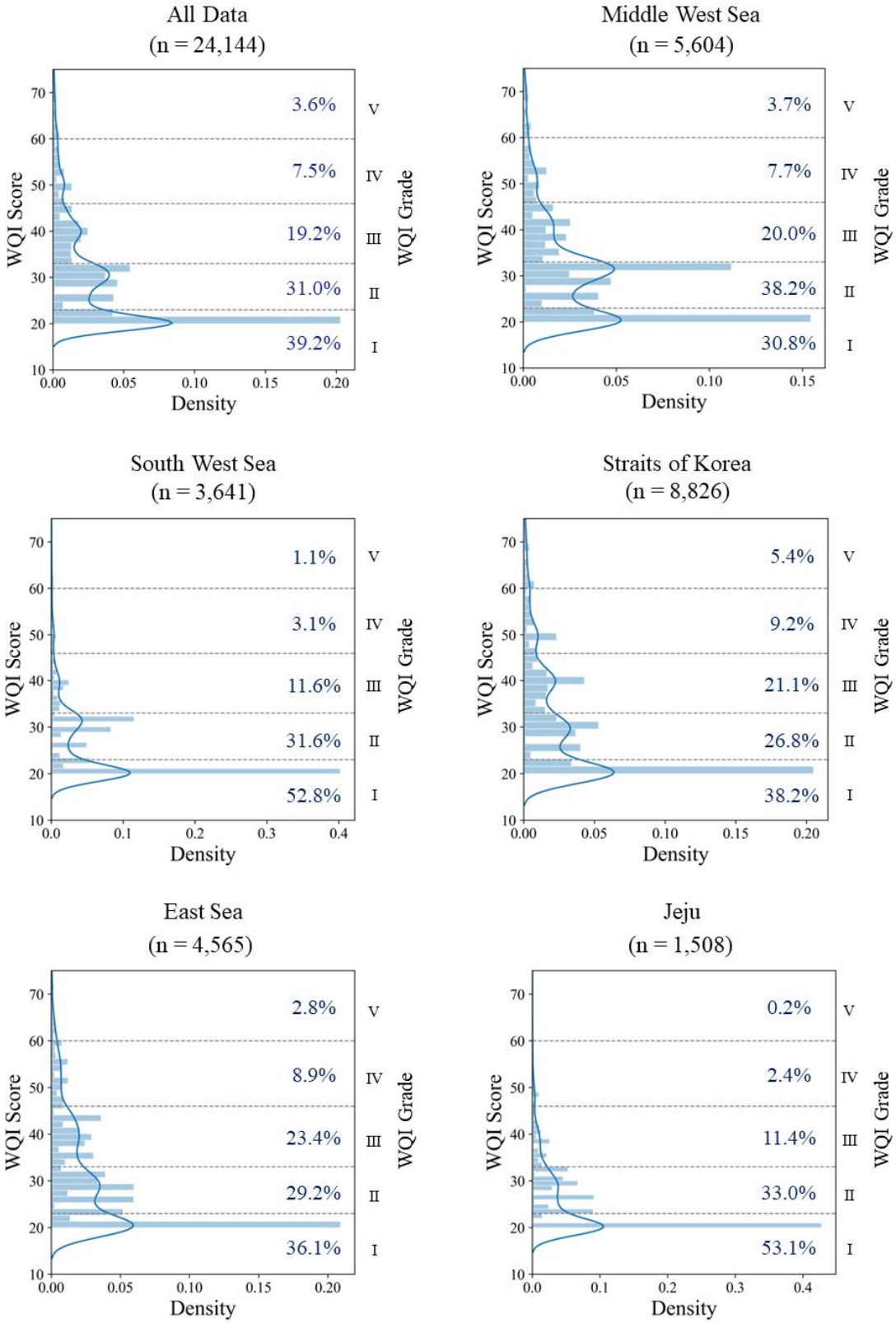

Fig. 2. Distribution of WQI score and grade by ecoregion.

3.2 Random Forest 예측 성능 평가

3.2.1 시간에 따른 예측 성능 분석

Random forest를 통해 산출된 WQI 예측 성능을 세부적으로 평가하기 위해 1999~2022년 데이터 범위 내에서 각 생태구의 잔차(residual)의 분포 특성을 분석하였다(Fig. 3). 잔차 평균은 0.004~0.029, 표준 편차는 0.66~0.98로 모든 생태구에서 잔차가 0에 가까운 값을 보였고, 87.60~91.42 %가 표준 편차 범위 내에 있는 것으로 나타났다(Table 6). 시계열 분포에서도 시간에 따른 유의미한 경향성은 관찰되지 않았다.

Table 6. Comparison of WQI residual statistical values

Region	mean	std	n	normal (%)
Middle West Sea	0.016	0.686	6378	87.60
South West Sea	0.029	0.660	4153	90.61
Straits of Korea	0.016	0.560	10015	88.30
East Sea	0.026	0.978	5138	91.20
Jeju	0.004	0.782	1724	91.42

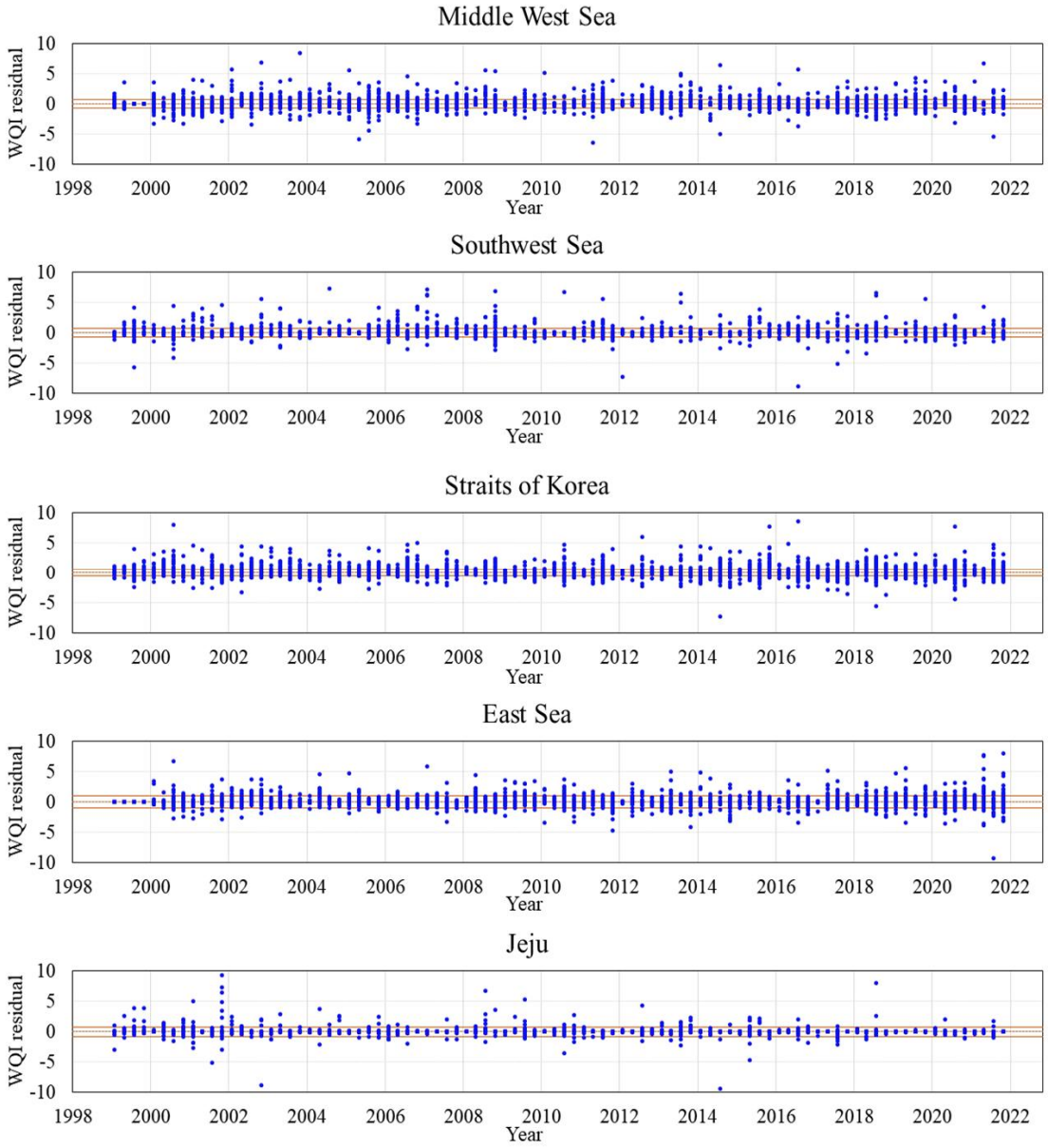

Fig. 3. Time series of WQI residual for each ecoregion.

3.2.2 공간에 따른 예측 성능

각 생태구내에서 지역별 WQI 예측 성능의 차이를 확인하기 위해 생태구별 WQI 실측값, 예측값, 그리고 잔차의 공간 분포를 분석하였다(Fig. 4). 앞서 연간 자료에서 시간에 따른 유의미한 성능 차이가 없었기 때문에 각 정점의 20년치 자료의 평균값을 분석에 사용하였다.

서해 중부의 WQI 실제값과 예측값은 인천, 시화호, 군산 일부 정점에서 47 이상(4등급)으로 높았고, 잔차는 최대 0.98로 나타났다. 서남해역은 목포, 고창, 여자만 일부 정점에서 34 이상(3등급)으로, 잔차는 최대 0.64로 나타났다. 대한해협은 낙동강 하구, 마산만 일부 정점에서 47 이상(4등급)으로 높았고 잔차는 낙동강하구에서 최대 1.03으로 나타났다. 동해에서는 영일만, 강릉, 축산, 죽변, 삼척, 구룡포 등에서 34 이상(3등급)으로 잔차는 영일만에서 최대 0.21로 나타났다. 제주는 전체적으로 2등급으로 나타났으며, 잔차는 제주 1,2 지점에서 최대 0.23으로 나타났다. 이와 같이 20년치 평균자료로 살펴봤을 때 모든 생태구역에서 실측값과 예측값이 거의 동일했고 잔차도 모든 지역에서 최대 1.03로 큰 차이를 보이지 않았으며, 잔차가 상대적으로 높은 지역은 WQI 등급이 높은 지역임을 알 수 있다.

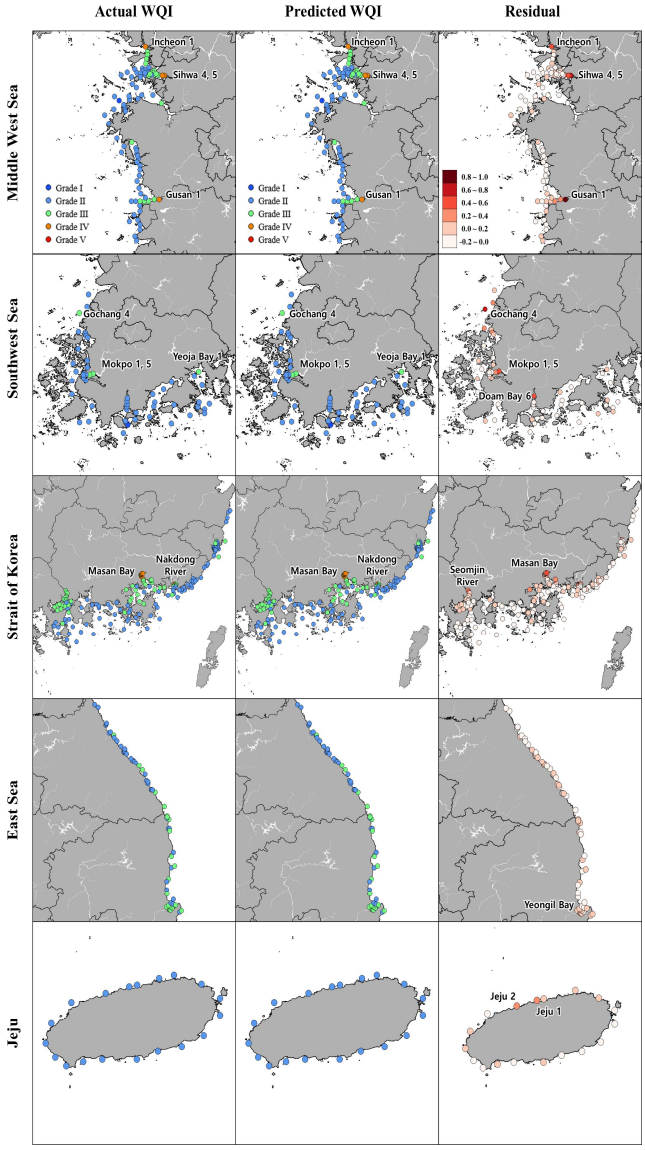

Fig. 4. Spatial distribution of residuals between predicted and actual WQI scores by Random Forest

4. 결 론

본 연구에서는 해양환경측정망 자료를 활용하여 우리나라 전 해역에서 수질평가지수(WQI)를 예측하는 인공지능 모델을 구축하였다. 최적 모델을 도출하기 위해 6가지 인공지능 알고리즘(RF, XGBoost, KNN, Ext, SVM, LR)의 모델별, 생태구별 예측 성능을 비교 하였고, 가장 성능이 좋은 모델의 예측값과 실제값의 잔차를 분석하여 WQI 예측 특성을 파악하였다.

모델 별 예측 성능은 Random Forest가 우수하였으며, 생태구별 예측 성능은 제주해역에서 가장 높고 서해 중부에서 가장 낮은 것으로 나타났다. 예측 성능이 가장 뛰어난 Random Forest 모델에 대한 심층 분석 결과, 모든 생태구에서 약 90 % 이상 자료의 잔차가 0에 가까웠고, 결정계수(R^2) 또한 0.97 이상으로 나타나 모델 성능의 우수성이 확인되었다.

국내 기존 연구에서 수행된 인공지능 기반의 적용 모델의 예측 성능은 높지만 국지적 해역을 대상으로 하여 다양한 수질 환경을 가진 우리나라 전 해역에 확대 적용하기에는 한계가 있다고 판단된다. 반면 본 연구에서 개발한 Random Forest 기반의 WQI 예측 모델은 우리나라 전 해역에서 장기간 관측된 해양수질자료(약 20년)를 기반으로 구축되었기 때문에 5개 생태구에 대한 각각의 해양 특성을 충분히 반영하고 있다고 예상된다. 향후 기존 5가지의 수질지표뿐만 아니라 다른 관측 수질지표까지 포함한 중요도를 평가하여 인공지능 기법을 활용한다면 WQI 평가 및 예측 정확도를 높일 수 있을 것으로 판단된다. 특히 생태계 정보를 연계한 인공지능 기반의 통합평가 및 예측이 구축된다면 해양환경 정책 추진 및 평가의 효율성을 높이는데 충분히 기여할 것으로 예상된다.

사 사

이 논문은 2023년도 정부(해양수산부)의 재원으로 해양수산과학기술진흥원-블루카본 기반 기후변화 적응형 해안 조성 기술개발 사업(KIMST-20220526)과 해양수산과학기술진흥원-과학기술기반 해양환경영향평가 기술개발 사업(KIMST-20210427)의 지원을 받아 수행된 연구임.

References

[1] Abba, S. I., Q. B. Pham, G. Saini, N. T. T. Linh, A. N. Ahmed, M. Mohajane, M. Khaledian, R. A. Abdulkadir, and Q. V. Bach(2020), Implementation of Data Intelligence Models Coupled with Ensemble Machine Learning for Prediction of Water Quality Index, Environmental Science and Pollution Research, 27(33), pp. 41524-41539.

[2] Bui, D. T., K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis(2020), Improving Prediction of Water Quality Indices Using Novel Hybrid Machine-Learning Algorithms, Science of the Total Environment, 721, p. 137612.

[3] Gaya, M. S., S. I. Abba, M. A. Abdu, A. I. Tukur, M. A. Saleh, P. Esmaili, and N. A. Wahab(2020), Estimation of Water Quality Index Using Artificial Intelligence Approaches and Multi-Linear Regression, IAES International Journal of Artificial Intelligence, 9(1), p. 126.

[4] Grbčić, L., S. Družeta, G. Mauša, T. Lipić, D. V. Lušić, M. Alvir, I. Lučin, A. Sikirica, D. Davidović, V. Travaš, D. Kalafatovicm, K. Pikelj, H. Fajkovic, T. Holjevic, and L. Kranjcevic(2022), Coastal Water Quality Prediction Based on Machine Learning with Feature Interpretation and Spatio-Temporal Analysis, Environmental Modelling & Software, 155, p. 105458.

[5] Haghiabi, A. H., A. H. Nasrolahi, and A. Parsaie(2018), Water Quality Prediction Using Machine Learning Methods, Water Quality Research Journal, 53(1), pp. 3-13.

[6] Hannan, A. and J. Anmala(2021), Classification and Prediction of Fecal Coliform in Stream Waters Using Decision Trees (DTs) for Upper Green River Watershed, Kentucky, USA, Water, 13(19), p. 2790.

[7] Huan, J., H. Li, M. Li, and B. Chen(2020), Prediction of Dissolved Oxygen in Aquaculture Based on Gradient Boosting Decision Tree and Long Short-Term Memory Network: A Study of Chang Zhou Fishery Demonstration Base, China, Computers and Electronics in Agriculture, 175, p. 105530.

[8] Jang, E., J. Im, S. Ha, S. Lee, and Y. G. Park(2016), 'Estimation of Water Quality Index for Coastal Areas in Korea Using GOCI Satellite Data Based on Machine Learning Approaches, Korean Journal of Remote Sensing, 32(3), pp. 221-234

[9] Jeon, S. B., H. Y. Oh, and M. H. Jeong(2020), Estimation of Sea Water Quality Level Using Machine Learning. Korea Spatial Information Society, Vol. 28, No. 4, pp. 145-152.

[10] Kadam, A. K., V. M. Wagh, A. A. Muley, B. N. Umrikar, and R. N. Sankhua(2019), Prediction of Water Quality Index Using Artificial Neural Network and Multiple Linear Regression Modelling Approach in Shivganga River Basin, India, Modeling Earth Systems and Environment, 5(3), pp. 951-962.

[11] Khan, I. U., N. Aslam, R. Alshehri, S. Alzahrani, M. Alghamdi, A. Almalki, and M. Balabeed(2021), Cervical Cancer Diagnosis Model Using Extreme Gradient Boosting and Bioinspired Firefly Optimization, Cervical Cancer Diagnosis Model Using Extreme Gradient Boosting and Bioinspired Firefly Optimization. Scientific Programming.

[12] Khullar, S. and N. Singh(2021), Machine learning techniques in river water quality modelling: a research travelogue, Water Supply, 21(1), pp. 1-13.

[13] Kim, S. B., J. S. Lee, and K. T. Kim(2022), WQI Class Prediction of Sihwa Lake Using Machine Learning-Based Models, J. The Sea: JOURNAL OF THE KOREAN SOCIETY OF OCEANOGRAPHY, 27(2), pp. 71-86.

[14] Kim, Y. H., J. Im, H. K. Ha, J. K. Choi, and S. Ha(2014), Machine learning approaches to coastal water quality monitoring using GOCI satellite data, GIScience & Remote Sensing, 51:2, pp. 158-174.

[15] Liaw, A. and M. Wiener(2002), Classification and regression by randomForest, R news, 2(3), pp. 18-22.

[16] López, V., A. Fernández, S. García, V. Palade, and F. Herrera (2013), An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Information Sciences, 250, pp. 113-141.

[17] Mi, Y.(2013), Imbalanced classification based on active learning SMOTE, Research Journal of Applied Sciences, Engineering and Technology, 5(3), pp. 944-949.

[18] Modaresi, F. and S. Araghinejad(2014), A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification, Water resources management, 28(12), pp. 4095-4111.

[19] Mountrakis, G., J. Im, and C. Ogole(2011), Support vector machines in remote sensing: A review, ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), pp. 247-259.

[20] Rho, T. K., T. S. Lee, S. R. Lee, M. S. Choi, C. Park, J. H. Lee, J. Y. Lee, and S. S. Kim(2012), Reference Values and Water Quality Assessment Based on the Regional Environmental Characteristics, The Sea : JOURNAL OF THE KOREAN SOCIETY OF OCEANOGRAPHY, Vol. 17, No. 2, pp. 45-58.

[21] Tanha, J., Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour(2020), Boosting methods for multi-class imbalanced data classification: an experimental review, Journal of Big Data, 7(1), pp. 1-47.

[22] Uddin, M. G., S. Nash, and A. I. Olbert(2021), A review of water quality index models and their use for assessing surface water quality, Ecological Indicators, 122, p. 107218.

[23] Uddin, M. G., S. Nash, A. Rahman, and A. I. Olbert(2022a), A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment, Water Research, 219, p. 118532.

[24] Uddin, M. G., S. Nash, M. T. M. Diganta, A. Rahman, and A. I. Olbert(2022b), Robust machine learning algorithms for predicting coastal water quality index, Journal of Environmental Management, 321, p. 115923.

Received : 2023. 01. 25.
Revised : 2023. 02. 13.
Accepted : 2023. 02. 24.