Journal of the Korea Organic Resources Recycling Association, 32(3), pp.19-29 (2024)
ISSN 1225-6498 eISSN 2508-3015 https://doi.org/10.17137/korrae.2024.32.3.19

Original Paper

입력변수 구성에 따른 총유기탄소(TOC) 예측 머신러닝 모형의 성능 비교

이소현[a], 박정수[b†]

Comparison of the Performance of Machine Learning Models for TOC Prediction Based on Input Variable Composition

Sohyun Lee[a], Jungsu Park[b†]

(Received: Jul. 31, 2024 / Revised: Aug. 14, 2024 / Accepted: Aug. 22, 2024)

ABSTRACT: Total organic carbon (TOC) represents the total amount of organic carbon contained in water and is a key water quality parameter used, along with biochemical oxygen demand (BOD) and chemical oxygen demand (COD), to quantify the amount of organic matter in water. In this study, a model to predict TOC was developed using XGBoost (XGB), a representative ensemble machine learning algorithm. Independent variables for model construction included water temperature, pH, electrical conductivity, dissolved oxygen concentration, BOD, COD, suspended solids, total nitrogen, total phosphorus, and discharge. To quantitatively analyze the impact of various water quality parameters used in model construction, the feature importance of input variables was calculated. Based on the results of feature importance analysis, items with low importance were sequentially excluded to observe changes in model performance. When built by sequentially excluding items with low importance, the performance of the model showed a root mean squared error-observation standard deviation ratio (RSR) range of 0.53 to 0.55. The model that applied all input variables showed the best performance with an RSR value of 0.53. To enhance the model's field applicability, models using relatively easily measurable parameters were also built, and the performance changes were analyzed. The results showed that a model constructed using only the relatively easily measurable parameters of water temperature, electrical conductivity, pH, dissolved oxygen concentration, and suspended solids had an RSR of 0.72. This indicates that stable performance can be achieved using relatively easily measurable field water quality parameters.

Keywords: feature importance, machine learning, total organic carbon, water quality management, XGBoost

초 록: 총 유기 탄소 (total organic carbon, TOC)는 물에 포함된 유기 탄소의 총량을 나타내며 BOD, COD와 함께 수중의 유기물질량에 대한 정량적인 지표로 활용되는 대표적인 수질 항목이다. 본 연구에서는 대표적인 앙상블 (ensemble) 머신러닝 알고리즘의 하나인 XGBoost (XGB)를 이용하여 TOC를 예측하는 모형을 구축하였다. 모형의 구축을 위한 독립변수로는 수온, pH, 전기전도도, 용존 산소 농도, 생물화학적 산소요구량, 화학적 산소요구량, 부유물질, 총질소, 총인 및 유량을 활용하였다. 또한 모형의 구축에 활용된 다양한 수질 항목의 영향에 대한 정량적인 분석을 위해 입력변수의 feature importance를 산정하였으며, 이를 기반으로 변수중요도에 따라 중요도가 낮은 항목을

[a] 국립한밭대학교 환경공학과 석사과정 (Graduate Course, Department of Civil and Environmental Engineering, Hanbat National University)
[b] 국립한밭대학교 환경공학과 부교수 (Associate Professor, Department of Civil and Environmental Engineering, Hanbat National University)
† Corresponding author(e-mail: parkjs@hanbat.ac.kr)

순차적으로 제외하여 모형의 성능 변화를 분석하였다. 변수중요도가 낮은 항목을 순차적으로 제외하여 구축한 모형의 성능은 RSR (root mean squared error-observation standard deviation ratio) 0.53~0.55의 범위를 보였으며, 전체 입력변수를 적용한 모형의 RSR 값은 0.53로 가장 우수한 성능을 보이는 것으로 분석되었다. 또한 모형의 현장 적용성을 높이기 위해 현장 측정이 상대적으로 용이한 측정항목을 중심으로 모형을 구축하고 성능을 분석하였다. 분석결과 상대적으로 측정이 용이한 항목인 수온, pH, 전기전도도, 용존산소농도, 부유물질농도만으로 구축된 모형의 경우에도 RSR 값이 0.72로 분석되어 상대적으로 측정이 용이한 현장 수질측정항목만을 이용하는 경우에도 안정적인 성능의 확보가 가능할 수 있음을 확인하였다.

주제어: 머신러닝, 변수중요도, 수질관리, 총유기탄소, XGBoost

1. 서 론

하천 및 저수지 등의 환경, 먹는 물 안전성 및 수생태계 건강성 확보를 위해 지속적인 수질관리가 필요하다. 수질관리를 위해 측정되는 여러 항목 중 하나인 총유기탄소 (Total Organic Carbon, TOC)는 물에 포함된 유기 탄소의 총량을 나타내며, 수중의 유기물의 양을 나타내는 정량 지표로 수질 오염의 척도로 활용되는 항목이다[1]. 수중의 유기물량을 정량적으로 나타내는 지표는 TOC 외에 생물화학적 산소요구량(Biochemical Oxygen Demand, BOD), 화학적 산소요구량(Chemical Oxygen Demand, COD) 등을 들 수 있으며, BOD와 COD는 수질관리 지표로 지속적으로 사용되어온 대표적인 항목이다. 하지만 난분해성 유기물의 증가와 COD_{MN} 의 유기물 산화율이 제한적인 한계 등을 극복하기 위해 TOC를 수질오염 총량관리를 위한 유기물의 정량지표로 활용하는 방안이 추진되었으며 2021년부터 공공하수처리시설의 방류수 수질기준중 COD가 TOC로 변경되어 적용되는 등 유기물질의 정량 지표로 TOC의 중요성이 커지고 있다[2]. 또한 TOC의 적정한 관리를 위한 TOC의 예측을 위한 모형의 구축에 대한 관심도 지속되고 있다.

최근 머신러닝 등 다양한 데이터 기반 모형을 수질 예측 및 관리에 활용하기 위한 연구가 계속되고 있다. Lee 등(2020)은 낙동강 중류의 2개 보 지점에서의 수질과 수량 항목에 대한 Chlorophyll-*a* (Chl-*a*)을 목표변수로 하여 두 지점에서 중요인자를 통해 의사결정 나무 (Decision Tree, DT), 랜덤 포레스트 (Random Forest, RF), elastic net, 그래디언트 부스팅 (Gradient Boosting, GB)을 비교 분석하였고[3], Jun 등(2020)은 낙동강 하류 유역의 3개의 수질 및 유량 모니터링 지점을 대상으로 총 11개의 측정 항목에 베이지안 정규화 신경망 (Bayesian regularized neural networks, BRNN), 서포트 벡터 머신 (Support Vector Machine, SVM), 신경망 (Neural network, NN) 3개의 머신러닝을 적용하여 예측 성능을 비교하였다[4]. Park 등(2023)은 낙동강 본류의 8개의 다기능 보 지점의 5년간의 Chl-*a* 농도를 예측하기 위해 RF, XGBoost (extreme gradient boosting, XGB)를 사용하였으며, 설명 가능한 인공지능 (explainable artificial intelligence, XAI) 기법인 SHAP을 활용하여 모델 예측 결과를 해석하였다[2]. 하천 등에서 수질항목의 예측을 위해 고도화된 머신러닝 모형의 적용을 위한 다양한 연구가 수행되어왔으며, Nafsin 등(2023)은 Milwaukee Metropolitan Sewerage District (MMSD)에서 제공하는 밀워키 강 유역의 수질자료를 활용하고 인공신경망 (artificial neural network ,ANN), SVM, gradient boosting model (GBM), RF 등을 활용하여 TOC와 대장균을 예측하는 연구를 수행한 바 있다[5]. 최근의 연구들은 다양한 머신러닝 알고리즘을 이용하여 대상 항목을 예측하는 예측 모형의 구축을 수행하는 연구와 함께 모형의 결과에 대한 정량적인 해석을 통해 유역 수질관리를 위한 의사결정에 활용할 수 있는 방안을 제시하는 등 머신러닝을 하천 관리 효율성을 높이는데 활용하기 위한 연구들을 포함하고 있다. 하지만 머신러닝을 이용한 TOC의 예측에 대한 연구는 아직 초기 단계로 향후 지속적인 연구가 필요하다.

머신러닝 모형의 구축을 위해서는 모형 구축에 필

요한 양질의 현장 측정자료의 확보가 중요하며 이러한 현장자료의 취득을 위해서는 많은 인력, 시간 및 비용이 소요된다. 본 연구에서는 다양한 분야에서 우수한 성능을 보여 널리 사용되는 대표적인 앙상블 머신러닝 모형인 XGB 모형을 이용하여 하천에서 TOC 농도를 예측하는 모형을 구축하였으며, 모형 구축에 적용되는 입력자료의 구성에 따른 모형의 성능 비교를 수행하였다[6]. 특히 상대적으로 자료 측정이 용이한 기초 항목을 중심으로 모형을 구축하고 그 성능에 대한 비교를 수행하여, 모형의 구축시 입력자료의 선택적 적용을 통해 현장 자료 취득에 필요한 비용과 노력을 최소화하여 머신러닝 모형의 현장 적용 효율을 높일 수 있는 방안을 제시하였다.

2. 재료 및 방법

2.1. 연구대상 지역

본 연구에서는 환경부 국립환경과학원에서 총괄 운영하는 물환경정보시스템의 수질 총량측정망 중 금호C 지점 (Site No.: 2012A70)에서 제공하는 2007년 1월 1일부터 2023년 12월 31일까지 측정된 806회의 주간 측정자료를 활용하여 모형의 구축 및 분석을 수행하였다[7].

연구 대상 지역인 금호강은 경상북도 포항시에서 발원하는 낙동강의 지류로 영천시, 대구광역시 등을 거쳐 낙동강에 합류된다[8]. 금호강은 낙동강의 제1지류로 인근 지역에 생활 및 공업용수를 공급하는 수원이며, 지역 주민들의 친수공간으로 활용되어 수질 및 수생태계의 지속적인 관리가 필요한 지역이다[9].

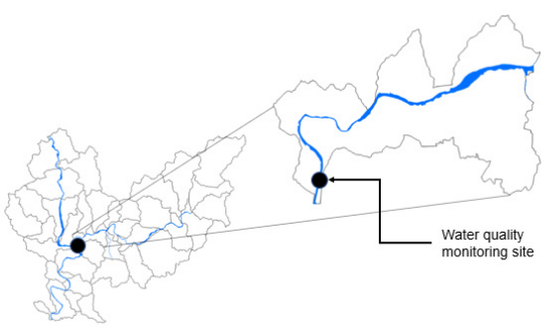


Fig. 1. Research site.

2.2. 입력자료 구축

현장 측정자료 중 수온 (temperature, TEMP), 수소이온농도 (potential of hydrogen, PH), 전기전도도 (electrical conductivity, EC), 용존산소 (dissolved oxygen, DO), BOD, COD 부유물질 (suspended solids, SS), 총질소 (total nitrogen, TN), 총인 (total phosphorous, TP), 유량 (discharge, Q)을 독립변수로 활용하였으며, TOC를 종속변수로 하여 모형을 구축하였다.

모형 구축에 사용한 입력자료 중 TOC는 1.12%의 결측값을 포함하였으며, TOC를 제외한 나머지 항목은 1% 미만의 결측치를 포함하여 결측값이 많지 않은 편이었으며 모형의 구축을 위해 python open source library인 scikit-learn의 K-Nearest Neighbors (KNN) 알고리즘을 이용하여 결측치에 대한 보간을 수행하였다[10].

총 803회의 측정 자료 중 2007년 1월 1일부터 2018년 12월 31일까지의 563회의 측정자료를 모형의 학습을 위한 training 자료, 2019년 1월 1일부터 2023년 12월 31일까지의 240회의 측정자료를 모형성능의 평가를 위한 testing 자료로 활용하여 training과 testing에 사용된 자료의 비율은 약 70%와 30%가 되도록 구성하였다.

2.3. 모형 구축

XGB는 extreme gradient boosting의 약자로 GB 모델의 과적합 문제를 개선하고 병렬학습이 지원되도록 구조를 변형한 알고리즘이며 오픈 소스로 제공되고 있다[6]. XGB 모형은 weak learner라고 불리는 개별 모형들로 구성되며, 이전 단계의 weak learner의 학습 결과를 다음 단계의 weak learner의 학습에 반영하여 단계적으로 모형의 성능을 향상시키는 boosting 방법을 사용하는 모형으로, 다양한 분야에서 활발하게 사용되는 앙상블 머신러닝 모형이다[6,11]. XGB 모형의 구축 및 최적 hyperparameters (i.g., learning rate, max depth, number of estimators) 결정을 위해 grid-search 및 cross validation을 수행하였으며, 모형의 구축은 python open source library인 XGB 및 scikit-learn을 이용하였다[10,12].

2.4. 모델 평가 기준

모형의 성능을 비교하기 위해 nash-sutcliffe coefficient of efficiency (NSE), root mean squared error (RMSE) 및 root mean squared error-observation standard deviation ratio (RSR) 3개의 성능지표를 이용하여 모형의 성능을 평가하였다(Eq. 1~3). 계산식에서 y_t 및 $\hat{y_t}$ 은 각각 시간 t에서의 실측값과 모형의 예측값을 나타내고, $\overline{y_t}$ 는 실측값의 평균을 나타내며, n은 자료 수를 나타낸다. RMSE는 모형의 예측값과 실측값의 차이를 정량화하는 지수로 그 값이 0에 가까울수록 모형의 예측 성능이 우수함을 나타낸다[13,14]. RSR과 NSE는 각각 0~∞과 -∞~1.0의 범위의 값을 가지며, RSR은 그 값이 약 0.7 미만인 경우, NSE는 1에 가까운 값을 가질수록 모형이 실측값을 잘 예측하는 것으로 판단한다[14]. RMSE는 실측값과 모형 예측값의 오차를 정량적으로 제시할 수 있어 모형 성능 평가에 널리 활용되는 지수이며, RSR은 RMSE를 표준화한 성능지표로, NSE와 함께 모형의 성능에 대한 보다 절대적인 비교 평가가 가능하다.

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2}{\sum_{t=1}^{n}\left(y_t - \overline{y}_t\right)^2} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \widehat{y_t})^2}{n}} \quad (2)$$

$$\mathrm{RSR} = \sqrt{\frac{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{n}\left(y_t - \overline{y}_t\right)^2}} \quad (3)$$

3. 결과 및 고찰

3.1. 입력자료의 특성 분석

모형의 구축에 사용된 입력변수의 평균, 최소, 최대 및 표준편차 값을 아래 표에 제시하였다.

TOC의 평균은 6.08이며, 3.6~14.2 범위의 값을 가지는 것으로 확인되었다. train에 사용된 TOC 값의 경우, 평균이 6.24이고, 3.6~14.2 범위의 값을 가지고 test에 사용된 TOC 값의 경우, 평균은 5.67이고, 3.7~.9 범위의 값을 가지는 것으로 확인되었다.

3.2. 모델 결과 및 변수중요도

전체 입력자료를 모두 적용한 모델의 독립변수는 TEMP, PH, EC, DO, BOD, COD, SS, TN, TP, Q로, 종속변수는 TOC로 이루어져 있다. 구축된 XGB 모형

Table 1. Characteristics of Input Variables

		Average	Min	Max	Standard deviation
Independent variables	TEMP (℃)	17.25	-1.00	33.80	8.18
	PH	7.99	6.40	10.20	0.53
	EC (μS/cm)	708.24	164.00	1340.00	209.84
	DO (mg/L)	10.89	5.10	18.10	2.52
	BOD (mg/L)	3.27	0.70	15.50	1.77
	COD (mg/L)	8.36	4.80	17.80	2.00
	SS (mg/L)	11.87	0.40	221.00	16.79
	TN (mg/L)	6.13	2.60	11.29	1.66
	TP (mg/L)	0.23	0.03	0.97	0.25
	Q (m^3/s)	36.03	0.00	1569.06	85.83
Dependent variable	TOC (mg/L)	6.08	3.60	14.20	1.44

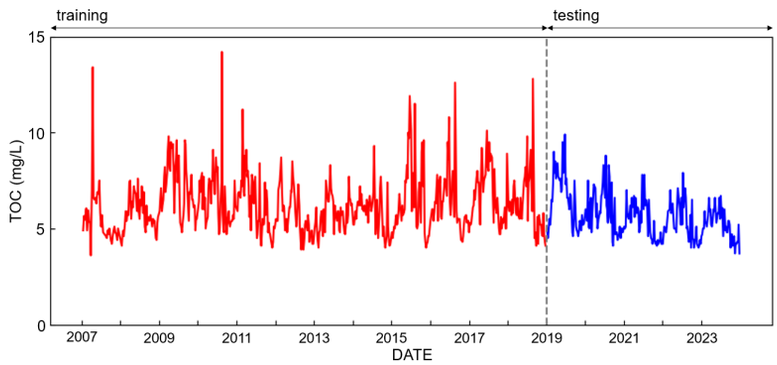


Fig. 2. TOC observations used for model development.

의 (M1_1) testing 자료에 대한 성능 평가 결과 RMSE는 0.61, RSR은 0.53, NSE는 0.72로 분석되었다. 구축된 모형의 실측값과 모형의 예측값에 대한 1:1 그래프를 그려본 결과를 Fig. 3에 제시하였다.

입력변수의 선정이 모형 성능에 미치는 영향에 대한 분석을 위해 python open source library인 feature importance 분석 알고리즘인 plot_importance (XGBoost)를 이용하여 각 독립변수가 모형에 미치는 상대적 중요도를 분석하여 Fig. 4에 제시하였다.

TOC 예측 모형의 독립변수로 사용된 항목들 중 또 다른 유기물 지표인 COD가 모형에 미치는 영향이 가장 큰 변수로 확인되었으며, 변수중요도는 COD > SS > TN > TP > PH > BOD > Q > EC 순으로 분석되었다. TEMP와 DO는 모형에 미치는 영향력이 없는 것으로 분석되었다. 또한, 비점오염원 등 다양한 오염원의 유입에 따라 발생되는 SS 및 TN, TP 등 영양염류의 농도도 모형의 예측 결과에 미치는 영향이 상대적으로 높은 편이었다. 이는 TN, TP 등 영양염류와 함께 유기물을 함유한 다양한 오염원의 유입에 따른 영향으로 판단된다. 또한, 수질의 이화학적 특성을 반영하는 기초항목인 PH, EC 및 유량 등의 영향은 상대적으로 높지 않은 것으로 분석되었으며, 생물학적으로 분해가 가능한 유기물 양을 나타내는 BOD의 영향도 상대적으로 크지 않은 것으로 확인되었다.

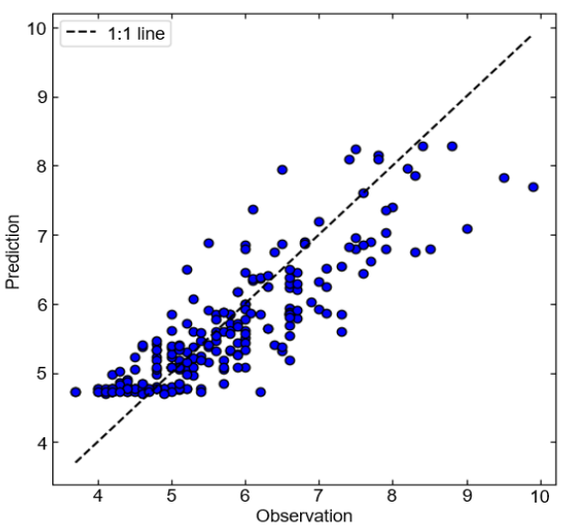


Fig. 3. Comparison of model (M1_1) predictions of TOC with observations.

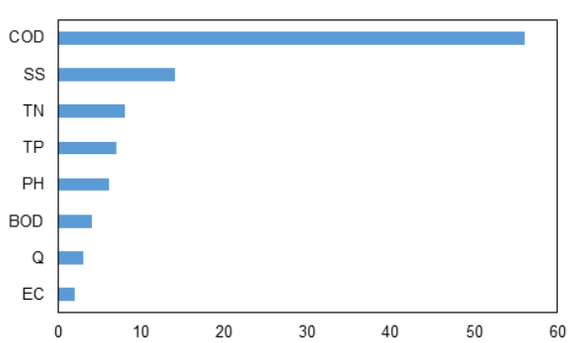


Fig. 4. Feature importance of independent variables in the XGBoost model.

3.3. 입력자료 구성이 모형 성능에 미치는 영향 분석

3.3.1. 변수중요도를 고려한 변수 선정에 따른 모형 성능 비교

적용되는 항목에 따른 성능 분석을 위해 다양한 변수들로 XGB 모델을 구축하였고 각 XGB 모델별 성능도 알아보았다. 우선 TOC를 예측하는데 연관성이 작았던 변수들부터 순차적으로 제외하여 M1_1~M1_10까지 10개의 모형을 구축하여 입력변수의 구성이 TOC 예측을 위한 모형의 성능에 미치는 영향을 분석하였다(Table 2).

입력변수의 상대적 중요도에 따라 다양하게 구축된 모형의 성능을 비교한 결과 모든 항목을 사용한 M1_1의 NSE, RMSE, RSR 값이 각각 0.72, 0.61, 0.53으로 가장 좋은 성능을 보이는 것으로 분석되었으며, M1_2~M1_4의 경우 M1_1과 동일한 수준의 성능을 보이는 것으로 확인되었다. 전체적으로 변수중요도에 따라 순차적으로 입력변수를 제외한 모형의 (M1_2~M1_10) 성능은 NSE는 0.70~0.72, RMSE는 0.61~0.64, RSR은 0.53~0.55의 범위를 가지는 것으로 분석되었으며 3가지 성능지표가 모두 유사한 경향을 보여 모형의 성능이 안정적으로 유지되는 것을 확인하였다 (Table S1). 이는 모형 구축 시 유기물의 정량 지표로 입력변수 중 모형 구축에 미치는 영향력이 가장 높은 COD를 입력변수로 적용한 결과로 판단된다.

Table 2. Independent Variables Used in Model 1

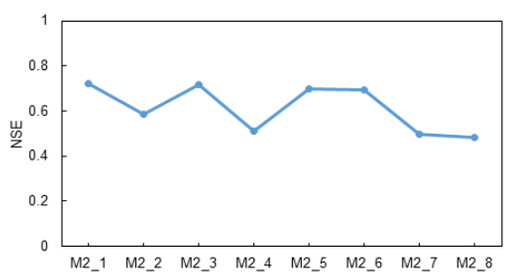
MODEL 1	Water quality items
M1_1	TEMP, PH, EC, DO, BOD, COD, SS, TN, TP, Q
M1_2	PH, EC, DO, BOD, COD, SS, TN, TP, Q
M1_3	PH, EC, BOD, COD, SS, TN, TP, Q
M1_4	PH, BOD, COD, SS, TN, TP, Q
M1_5	PH, BOD, COD, SS, TN, TP
M1_6	PH, COD, SS, TN, TP
M1_7	COD, SS, TN, TP
M1_8	COD, SS, TN
M1_9	COD, SS
M1_10	COD

3.3.2. 모형의 현장 적용성을 고려한 변수 선정에 따른 성능 비교

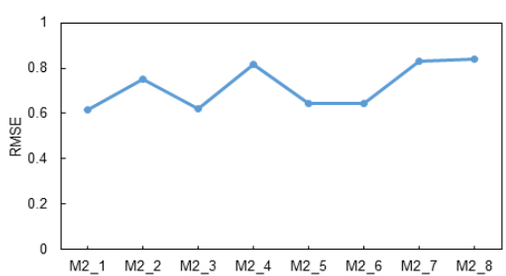
머신러닝 모형은 입력자료 확보가 중요하여, 현장

Table 3. Independent Variables Used in Model 2

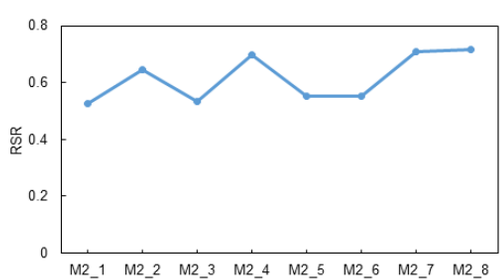
MODEL 2	Water quality items
M2_2	TEMP, PH, EC, DO, BOD, SS, TN, TP, Q
M2_3	TEMP, PH, EC, DO, COD, SS, TN, TP, Q
M2_4	TEMP, PH, EC, DO, SS, TN, TP, Q
M2_5	TEMP, PH, EC, DO, COD, SS, Q
M2_6	TEMP, PH, EC, DO, COD, SS
M2_7	TEMP, PH, EC, DO, SS, Q
M2_8	TEMP, PH, EC, DO, SS



(a) NSE

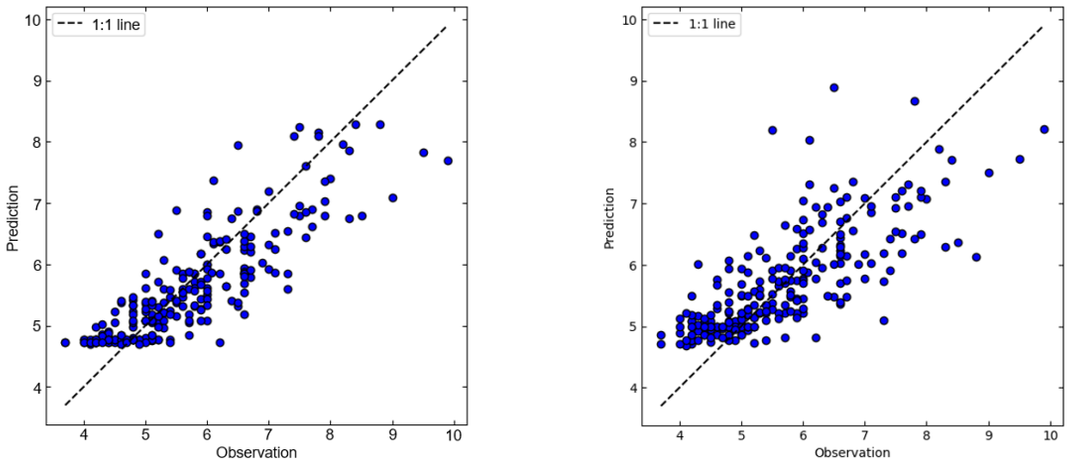


(b) RMSE

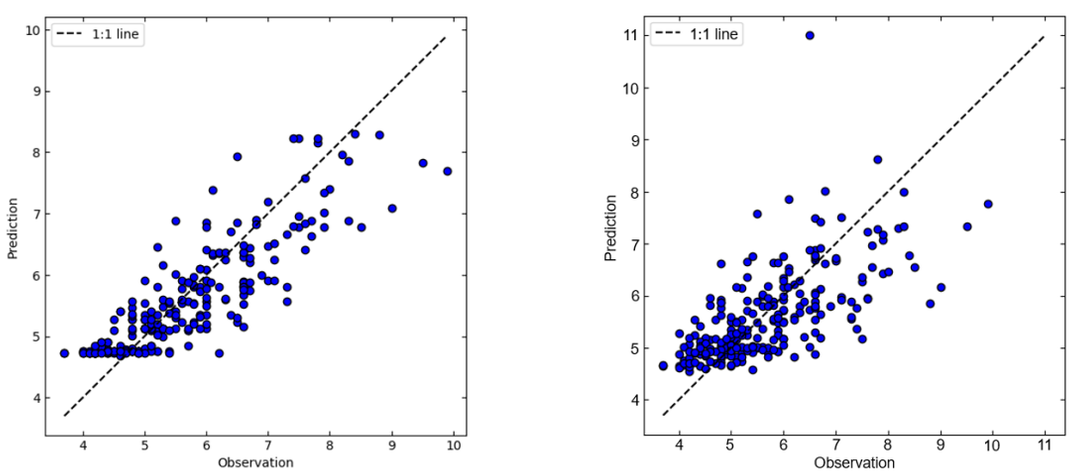


(c) RSR

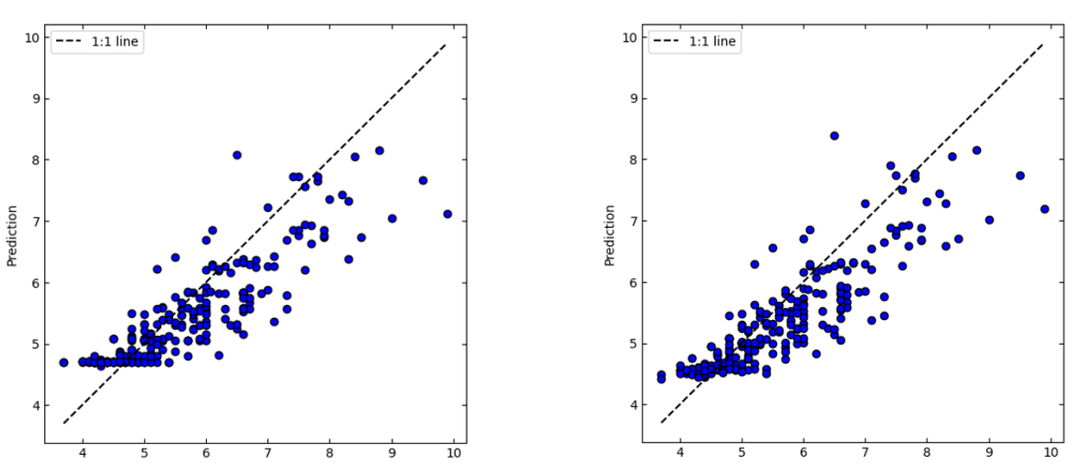
Fig. 5. Comparison of model performance with varying input variable selectin considering field applicablity.



(a) M2_1 (b) M2_2

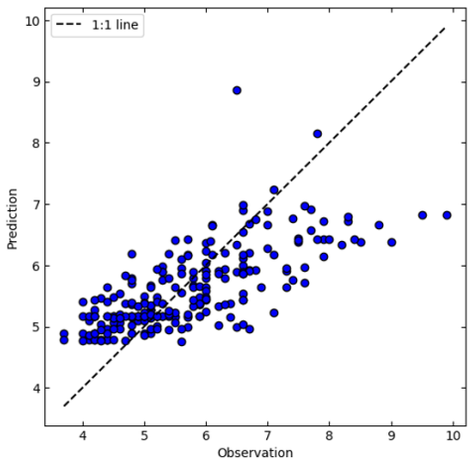


(c) M2_3 (d) M2_4

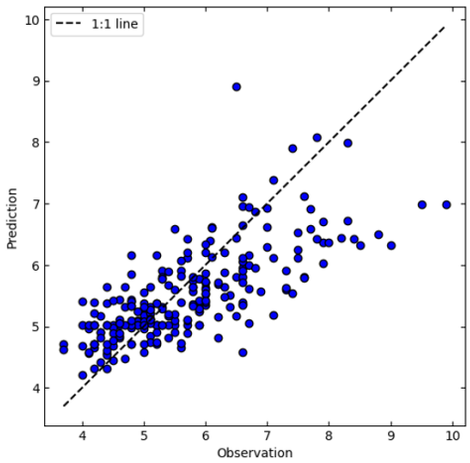


(e) M2_5 (f) M2_6

Fig. 6. Comparison of model predictions with observations with varying input variable composition considering field applicablity.



(g) M2_7



(h) M2_8

Fig. 6. Continued.

측정 자료의 취득과 결측치가 없이 많은 양과 정확한 데이터가 필요하다. 수질 항목으로 측정이 되는 BOD, COD, TN, TP는 실험을 통해 취득되는 항목으로 분석을 위해 숙련된 전문인력이 필요하며 많은 시간과 비용이 소요되는 항목이다.

반면 TEMP, PH, EC, DO는 일반적으로 현장에서 센서를 통한 측정이 가능한 항목으로 상대적으로 분석 및 취득이 용이하다. 본 연구에서는 구축된 머신러닝 모형의 현장 적용성을 확인하기 위해 현장 측정이 용이한 항목 중심으로 변수를 선정해서 모형을 구축하고 (M2_2~M2_8) 성능을 비교하였다 (Table 3).

구축된 모형(M2_2~M2_8)의 성능분석결과 NSE는 0.48~0.72, RMSE는 0.61~0.84, RSR은 0.53~0.72로 산정되었다. 전체 항목을 적용한 모형(M1_1)에 비해 다소 성능이 낮아졌으나 M2_2~M2_6의 경우 RSR 0.53~0.70으로 안정적인 성능을 보였으며 BOD, COD 등을 제외한 모형인 M2_7 및 M2_8의 경우에도 NSE가 각각 0.50 및 0.48, RSR이 각각 0.71 및 0.72로 안정적인 성능확보가 가능한 것으로 분석되었다. 구축된 모형의 예측값과 실측값을 비교하여 Fig. 5에 제시하였다.

4. 결 론

하천 및 저수지 등의 유기물 양은 수질 오염의 척도로 중요한 의미를 가지고 있으며 TOC는 이러한 유기물량을 나타내는 대표적인 정량지표이다. 본 연구에서는 우수한 성능으로 다양한 분야에서 널리 활용되는 대표적인 앙상블 머신러닝 모형인 XGB를 이용하여 수중의 TOC 농도를 예측하는 모형을 구축하였다. 데이터 기반 모형인 머신러닝 모형은 입력자료의 구성과 특성에 따라 성능에 많은 영향을 받으며 본 연구에서는 다양한 입력변수의 구성이 XGB 모형의 성능에 미치는 영향에 대한 세부적인 분석을 수행하였다. 우선 XGB 모형의 구축에 활용된 입력변수의 중요도를 산정하고 중요도가 낮은 항목부터 순차적으로 제외하여 10개의 세부 모형을 구축하여(M1_1~M1_10) 모형의 성능을 비교하였다. 변수중요도 산정결과 TOC와 함께 유기물량을 나타내는 대표적인 정량지표인 COD가 모형의 성능에 미치는 영향이 가장 높은 것으로 분석되었다. 구축된 모형의 성능 분석 결과 전체 항목을 모두 적용하는 경우 가장 우수한 성능을 보였으며, 전체적으로 입력변수를 순차적으로 제외한 모형들도 모두 안정적인 성능을 보였는데 이는 유기물의 정량지표인 COD를 포함하여 모형의 변수를 구성한 결과인 것으로 판단된다.

머신러닝 모형의 구축을 위해서는 양질의 현장 측정자료의 확보가 필요하며, 특히 실험실 분석 등을 수행하기 위해서는 많은 전문인력과 시간이 소요된다. 본 연구에서는 머신러닝 모형의 현장 적용성을 고려하여 상대적으로 센서 등을 활용한 현장 측정이 용이한 수질항목(TEMP, PH, DO 등)을 중심으로 모형을 구축하고(M2_2~M2_8) 그 성능을 비교하였다. 분석 결과 상대적으로 측정이 용이한 TEMP, PH, EC, DO, SS만을 입력변수로 사용하는 경우에도 어느 정도 안정적인 모형의 성능확보가 가능한 것을 확인하였다.

본 연구의 결과를 통해 TOC 예측을 위한 머신러닝 모형의 적용 가능성을 확인하였다. 향후 지속적인 관련 분야 연구를 통해 모형구축에 필요한 현장 측정 자료의 확보를 위한 인력 및 비용의 절감 등 모형의 현장 적용성 및 현장 수질관리 효율을 높이는데 기여할 수 있을 것으로 판단된다.

사 사

이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1F1A1065518). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022 R1F1A1065518).

References

1. Ministry of Environment (ME), Introduction of Total Organic Carbon Management in Nakdonggang River Water System, Ministry of Environment, 1~5, (2022).
2. Park, S. R., Son, S. H., Bae, J. G., Lee, D., Seo, D. I., and Kim, J. S., “Estimation of Chlorophyll-a Concentration in Nakdong River Using Machine Learning-Based Satellite Data and Water Quality, Hydrological and Meteorological Factors”, Korean Journal of Remote Sensing, 39(5), pp. 655~667. (2023).
3. Lee, S. M., Park, K. D., and Kim, I. K., “Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong River (focusing on water quality and quantity factors)”, Journal of Korean Society of Water and Wastewater, 34(4), pp. 277~288. (2020).
4. Jun, G., Kwon, D., and Ki, S., “Comparing the Performance of Machine Learning Algorithms in Predicting River Water Quality and Quantity”, Journal of Korea Society of Water Science and Technology, 28(1), pp. 49~57. (2020).
5. Nafsin, N., and Li, J., “Prediction of total organic carbon and E. coli in rivers within the Milwaukee River basin using machine learning methods”, Environmental Science: Advances, 2(2), pp. 278~293. (2023).
6. Chen, T., and Guestrin, C., “Xgboost: A scalable tree boosting system”, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785~794. (2016).
7. Water Environment Information System (WEIS), https://water.nier.go.kr/web (Accessed date: April 23, 2024).
8. Choi, B. D., “The Function or urban river and sustainable regional development: The case of Kumho river”, Journal of the Korean Association of Regional Geographers, 10(4), pp. 757~774. (2004).
9. Yang, D. S., and Bae, H. K., “The effect of branches on Kumho River's water quality”. Journal of Environmental Science International, 21(10), pp. 1245~1253. (2012).
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É., “Scikit-learn: Machine learning in Python”, the Journal of machine Learning research,

12, pp. 2825~2830. (2011).
11. Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., and Xue, K., “A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes”, Remote Sensing of Environment, 248, pp. 111974. (2020).
12. XGBoost, https://pypi.org/project/xgboost/ (Accessed date: November 21, 2023).
13. Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Northon, J. P., Perrin, C., Pierce, S. A., Robson, B. J., Seppelt, R., Voinov, A., Fath, B. D., and Andreassian, V., “Characterising performance of environmental models”, Environmental Modelling & Software, 40, pp. 1~20. (2013).
14. Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L., “Model evaluation guidelines for systematic quantification of accuracy in watershed simulations”, Transactions of the ASABE, 50(3), pp. 885~900. (2007).

Table S1. Performance of XGBoost Model Trained with Various Input Variable Compositions Selected based on Feature Importance Analysis

	NSE	RMSE	RSR
M1_1	0.7227	0.6143	0.5266
M1_2	0.7227	0.6143	0.5266
M1_3	0.7227	0.6143	0.5266
M1_4	0.7226	0.6144	0.5266
M1_5	0.7210	0.6161	0.5281
M1_6	0.7155	0.6222	0.5333
M1_7	0.7159	0.6218	0.5330
M1_8	0.7073	0.6311	0.5410
M1_9	0.6974	0.6417	0.5501
M1_10	0.7119	0.6261	0.5367