Journal of Korean Society on Water Environment, Vol. 35, No. 5, pp. 432-438 (September, 2019)
pISSN 2289-0971 eISSN 2289-098X https://doi.org/10.15681/KSWE.2019.35.5.432

Correlation-based Feature Selection 기법과 Random Forest 알고리즘을 이용한 한강유역 지류의 TDI 예측 연구

김민규[1a] · 윤춘경[2a] · 이한필[1b,†] · 황순진[2b] · 이상우[3]

[1](주)이티워터 · [2]건국대학교 환경보건과학과 · [3]건국대학교 산림조경학과

A Study on Predicting TDI(Trophic Diatom Index) in tributaries of Han river basin using Correlation-based Feature Selection technique and Random Forest algorithm

Kim Minkyu[1a] • Yoon Chun Gyeong[2a] • Rhee Han-Pil[1b,†] • Hwang Soon-Jin[2b] • Lee Sang-Woo[3]

[1]ETWaters Inc.
[2]Department of Environmental Health Science, Konkuk University
[3]Department of Forestry and Landscape Architecture, Konkuk University
(Received 18 June 2019, Revised 20 September 2019, Accepted 25 September 2019)

Abstract

The purpose of this study is to predict Trophic Diatom Index (TDI) in tributaries of the Han River watershed using the random forest algorithm. The one year (2017) and supplied aquatic ecology health data were used. The data includes water quality(BOD, T-N, NH_3-N, T-P, PO_4-P, water temperature, DO, pH, conductivity, turbidity), hydraulic factors(water width, average water depth, average velocity of water), and TDI score. Seven factors including water temperature, BOD, T-N, NH_3-N, T-P, PO_4-P, and average water depth are selected by the Correlation Feature Selection. A TDI prediction model was generated by random forest using the seven factors. To evaluate this model, 2017 data set was used first. As a result of the evaluation, R^2, % Difference, NSE(Nash-Sutcliffe Efficiency), RMSE(Root Mean Square Error) and accuracy rate show that this model is compatible with predicting TDI. To be more concrete, R^2 is 0.93, % Difference is －0.37, NSE is 0.89, RMSE is 8.22 and accuracy rate is 70.4%. Also, additional evaluation using data set more than 17 times the measured point was performed. The results were similar when the 2017 data set were used. The Wilcoxon Signed Ranks Test shows there was no statistically significant difference between actual and predicted data for the 2017 data set. These results can specify the elements which probably affect aquatic ecology health. Also, these will provide direction relative to water quality management for a watershed that must be continuously preserved.

Key words : Aquatic ecology health, Correlation-based Feature Selection, Random forest, TDI Prediction

[1a] 연구원(Researcher), mkkim@etwaters.co.kr, 0000-0002-0930-2481
[2a] 교수(Professor), chunyoon@konkuk.ac.kr, 0000-0003-2942-2197
[1b,†] Corresponding author, 대표(President), hprhee@etwaters.co.kr, 0000-0003-2519-1547
[2b] 교수(Professor), sjhwang@konkuk.ac.kr, 0000-0001-7083-5036
[3] 교수(Professor), swl7311@konkuk.ac.kr, 0000-0002-3275-7564

1. Introduction

우리나라는 지형적·기후적인 영향으로 안정적인 수자원 확보가 쉽지 않은 특성이 있다. 또한, 급격한 산업화와 인구증가는 물수요를 가파르게 증가시켰고, 그에 따라 오염원도 증가하여 수질관리의 필요성이 대두되었다. 과거 우리나라에서는 유기물 오염 지표인 DO(용존산소량), BOD(생물화학적 산소요구량), COD(화학적 산소요구량)와 부영양화의 지표인 T-N(총질소), T-P(총인)과 같은 이화학적 항목들로 하천 수질을 평가해왔다(An et al., 2005).

하지만 BOD를 포함하는 화학적 요인만으로는 생태계의 건강성을 총체적으로 판정하기에는 크게 미흡하다(Hwang et al., 2006). 이에 반해 지표생물(indicator organisms)을 이용한 생물학적 방법은 연간의 평균적인 수질을 대변하고 과거 오염물질의 임의적 유출에 대한 추정을 가능케 해줌과 동시에 오염물질의 복합효과 등에 따른 종합적 영향을 반영해 준다(Kong, 2002). 이러한 이유로 선진국에서는 이미 생물학적 평가기법을 이용하여 수생태계 환경을 평가할 수 있는 지수를 개발(Kelly and Whitton, 1995; Prygiel and Coste, 1993)하여 이용하고 있으며, 우리나라에서도 Trophic Diatom Index (TDI), Benthic Macroinvertebrate Index(BMI), Fish Assessment Index (FAI) 등 생물학적인 평가 기준을 마련하여 시행하고 있다(NIER, 2017).

이와 더불어 과거에는 본류의 수질관리에 집중해왔으나, 최근에는 본류뿐만 아니라 지류와 지천의 수질관리에 관한 관심이 높아지는 추세이다. 일반적으로 지류구간은 본류에 비해 유량변동성이 높아 건천화에 취약하고, 특히 도심의 경우는 인공구조물로 인한 영향으로 수생태계 및 수질이 불량할 뿐만 아니라, 오염원의 영향권 내 지류의 수질이 매우 나쁘기 때문에 이에 대한 지속적 관리가 필요한 실정이다(Kal et al., 2017). 또한, 지류·지천에 해당하는 소유역으로부터 발생하는 비점오염물질, 생활하수, 축산폐수 등이 하천으로 유입되면 일부는 하천의 자정작용에 의해 저감될 수 있으나, 일부는 본류에 유입되어 수질오염을 가중시킬 수 있으므로 지류·지천의 관리는 점점 중요해지고 있다.

국내에서는 낙동강 지류·지천의 유량·수질 특성 및 하천관리를 위한 등급화 방안 연구(Na et al., 2015)에서 지류·지천의 효율적인 수질 개선방안을 모색하였고, 금강권역 주요 하천의 돌부착돌말류 분포 및 출현예측(Cho et al., 2015)에서 금강권역의 돌부착돌말류 분포와 환경과의 관계를 파악하였다. 또한 낙동강 상류 수계인 내성천의 부착돌말 군집과 부착돌말지수를 이용한 생물학적 수질평가(Choi et al., 2017)에서 내성천의 생물학적 수질을 평가하였으며, SWAT 및 random forest를 이용한 기후변화에 따른 한강유역의 수생태계 건강성 지수 영향 평가(Woo et al., 2018)에서 수질과 수생태계 건강성 지수 간 상관성 확인 및 한강 유역의 미래 수생태 영향을 SWAT 모형과 random forest 모형을 이용하여 평가하였다.

효율적으로 지류·지천의 수질을 관리하기 위해서는 미래에 유역환경의 변화에 따른 수문, 수질 및 수생태계의 변동성을 예측하고 그에 따른 관리방안 마련이 필요하다. 대부분의 연구에서 현재 혹은 과거 몇 년간의 수생태계 건강성 혹은 생물학적 수질에 대해 평가해왔으나, 수생태계 건강성 평가 지수의 미래 예측 부분에 관한 연구는 아직 초기단계이다. 이미 선행된 연구(Woo et al., 2018)에서도 봄철, 수질인자에 국한되어 예측하여 계절적, 수리적 영향이 반영되지 않은 한계점이 있다.

따라서 본 연구에서는 수생태계 건강성 지표 중 하나인 TDI의 예측 모형을 구축하여 실측지수와 예측된 건강성 진단결과를 비교 및 평가하고, 이를 관리방안 마련에 근거자료로 사용할 수 있도록 하였다. 이를 위해 수질・수리・수문 등 다양한 하천환경 인자와 수생태 지표 간 상관분석을 하고, 환경인자와 생물 지표 간 피어슨 상관관계 분석(Pearson correlation analysis) 및 Correlation-based Feature Selection(CFS) 기법을 통한 주요 속성 집합 탐색을 적용하였다. 또한 CFS 탐색 결과로부터 인자의 우선순위를 선정하고, 기계학습 기법 중 랜덤 포레스트 알고리즘을 적용하여 TDI 예측 모형을 구축 및 평가하였다.

2. Materials and Methods

2.1 연구대상하천

본 연구는 한강권역 내의 지류를 대상으로 하였다. 한강권역은 강원도와 경기도, 서울과 충청도 일부를 포함하며 면적은 약 31,648 km^2이다. 하천의 개소수는 국가 하천 19개소와 지방하천 895개소를 포함 총 914개이다. 하천 연장의 경우 국가 하천이 약 917 km, 지방하천이 약 7,662 km이다(MOLIT, 2014). 한강수계의 생물측정망 조사지점(2016~2018)은 총 907개이며, 상시 조사지점은 1년 주기, 일반 조사지점은 3년 주기로 봄・가을 2회에 걸쳐 수생태계 건강성을 조사 및 평가하고 있다. 본 연구에서는 2008년~2017년 한강권역 수생태계 건강성 데이터 중 대상 하천의 분류가 지류로 구분되어있는 지점의 1,312개 자료를 한강유역환경청으로부터 제공받아 활용하였다.

2.2 피어슨 상관관계 분석

상관분석은 유의미한 연관성이 예상되는 두 요소 간 구간・비율척도 변수에 대하여 선형적으로 연관성이 얼마나 있는지 확인하는 분석방법이다. 변수 간 관계의 방향에 따라 양의 상관관계, 음의상관관계로 나뉘며, -1에서 1의 값을 갖는다.

피어슨 상관분석은 일반적으로 표본수가 많은 경우와, 각 변수의 모집단 분포가 정규분포에 근접한 경우에 사용할 수 있다. 피어슨 상관계수는 보통 r로 표현하는데, 여기서 X, Y의 상관계수는 두 변수의 값이 (x_1, y_1), (x_2, y_2), …,(x_n, y_n)일 때 식 (1)~(3)과 같이 계산할 수 있다.

$r = \frac{S_{xy}}{S_x \cdot S_y}$일 때,

$$S_{xy} = X\text{와 } Y\text{의 공분산} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad \text{- (1)}$$

$$S_x = X\text{의 표준편차} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \quad - (2)$$

$$S_y = Y\text{의 표준편차} = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} \quad - (3)$$

2.3 CFS(Correlation-based Feature Selection) 기법

CFS 기법은 상관관계에 기반하여 휴리스틱(heuristic) 평가함수에 따라 속성들의 부분집합에 순위를 부여하는 간단한 필터 알고리즘이다(Hall, 1999). CFS는 추정해야 할 결과값과 상관성이 큰 속성이면서, 동시에 다른 속성들과의 상관성은 낮은 속성을 탐색하는 방법으로 알려져 있다. 불필요한 속성은 제외시키고 의미 있는 속성만을 선택적으로 추출한다는 점에서 큰 이점을 갖는다. 특정 분류 작업과 가장 관련성이 높은 기능 하위 집합을 찾는 데 사용된다. CFS 알고리즘의 주요 부분은 방정식에서 제공한 대로 기능의 하위 집합의 유용성 또는 장점을 평가하는 경험적 방법이다. 이 경험적 방법은 클래스 레이블을 예측하기 위한 개별 기능의 유용성과 그 클래스 간의 상호 상관 수준을 보여준다. 동일한 원리가 예측 및 외부 관심 변수에 대한 복합 테스트(개별 테스트의 합 또는 평균)를 디자인하는 데 사용된다. 이 상황에서 feature는 관심 변수(class)와 관련된 특성을 측정하는 개별 테스트이다(식 (4)).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad - (4)$$

여기서 Merit$_s$는 k개의 feature를 포함한 feature subset의 휴리스틱 요소이다. $\overline{r_{cf}}$는 평균 feature = class correlation이고 $\overline{r_{ff}}$는 평균 feature-feature 급간 상관이다. 위의 식은 실제로 모든 변수가 표준화된 피어슨 상관관계이다. 분자는 feature group이 얼마나 예측 가능한지를 나타내고, 분모는 얼마나 많은 중복성이 있는지를 나타낸다. 중복 feature는 하나 이상의 다른 기능과 높은 상관관계가 있으므로 구별된다. CFS는 먼저 학습 데이터로부터 feature class 및 feature간 상관관계의 행렬을 계산한 후 Best first search를 사용하여 feature subset space를 검색한다.

2.4 TDI(Trophic Diatom Index)

부착돌말류는 하천생태계의 1차 생산자로서 돌, 식물, 모래, 진흙 등 다양한 기질에 부착하여 서식하는데, 수질 영양상태 및 환경변화에 민감하며 기질에 장기간 부착하여 서식함으로써 수생태계 건강성을 판단하는 생물로 활용하며, 이를 TDI라는 부착돌말 영양지수로 나타낸다(NIER, 2011).

부착돌말 영양지수는 각 구간에서 출현한 종의 상대밀도, 종의 오염민감도 및 지표값을 사용하여 계산한다(NIER, 2013).

$$TDI = 100 - (WMS \times 25) - 25 \quad - (5)$$

WMS : 가중평균민감도(Weighted Mean Sensitivity)

$$WMS = \frac{\sum A_i \times S_i \times V_i}{\sum A_i \times V_i} \quad - (6)$$

A_i : 표본 내 종의 상대풍부도(abundance(proportion) of ith species in sample, %)

S_i : 종의 오염 민감도(pollution sensitivity of ith species, $1 \leq S \leq 5$)

V_i : 종의 지표 값(indicator value of ith species, $1 \leq V \leq 3$)

계산된 값이 90점 이상이면 A, 70점 이상이면 B, 50점 이상이면 C, 30점 이상이면 D, 30점 미만은 E로 등급을 부여한다. 본 연구에서는 등급 분류를 하기 전의 점수를 예측하고, 그 점수를 등급화하여 실제 부여된 등급과 예측치를 비교하여 그 정확성을 검증하였다.

2.5 Weka

뉴질랜드의 The University of Waikato에서 개발한 프로그램인 Weka (Waikato Environment for Knowledge Analysis)는 데이터 마이닝 작업에 필요한 기계학습 알고리즘 모음이다. 여기에는 데이터 준비, 분류, 회귀, 클러스터링, 연관 마이닝 및 시각화를 위한 도구가 포함되어 있다. Weka는 사용자가 빠르고 유연한 방법을 통해 새로운 데이터 집합을 대상으로 기존 알고리즘을 적용할 수 있게 설계됐다. 다양한 학습 알고리즘, 광범위한 전처리 도구 등 다양한 툴이 보편적인 인터페이스로 접근할 수 있게 만들어져 있어 사용자들은 여러 가지 알고리즘을 비교하고 해결하려는 문제에 어떤 알고리즘이 가장 적합한지 직접 확인할 수 있다. 여기에 데이터 집합에 학습 알고리즘을 적용하고, 그 처리 결과를 분석해서 데이터의 본질에 대해 더욱 많은 정보를 알아낼 수 있다. 그뿐만 아니라, 알아낸 모형을 이용해 새로운 인스턴스들에 대한 예측결과를 생성하는 것이 가능하다. 본 연구에서는 Weka의 기능 중 피어슨 상관계수 분석, CFS 기법과 랜덤포레스트 알고리즘을 사용하여 데이터 분석 및 예측을 수행하였다.

2.6 랜덤포레스트(Random Forest)

랜덤포레스트(random forest)는 의사결정나무분석 중 CART 알고리즘과 앙상블 모형 중 배깅 알고리즘을 적용한 알고리즘으로, 2001년 Leo Breiman이 제안하였다(Breiman, 2001). 랜덤포레스트는 각각의 트리가 독립적으로 추출된 임의의 벡터값과 포레스트의 모든 트리에 대해 동일하게 분포 의존하는 트리 예측기의 조합이다.

랜덤포레스트는 예측력이 우수하며 의사결정나무와는 달리 과대적합하지 않는다. 적당한 임의성을 부여하면 정확한 분류기 및 회귀 변수가 될 수 있다. 의사결정나무는 규칙을 만드는 조건을 조금 바꾸게 되면 모형이 크게 변하지만, 랜덤포레스트는 여러 가지의 작은 부트스트랩 및 변수 샘플 데이터셋에서 모형을 만들어 결합하므로 생성된 모형이 안정적인 예측결과를 산출할 수 있다.

2.7 예측 모형의 평가

일반적으로 모형의 정확도를 검증할 때 결정계수(R^2)와

Nash-Sutcliffe Efficiency(Nash and Sutcliffe, 1970)를 사용한다. 결정계수의 경우 0부터 1사이의 값이며, 1에 가까울수록 정확도가 높은 것으로 알려져 있다. NSE는 $-\infty$에서 1까지의 값을 가지는데, 1에 근접할수록 정확도가 높은 모형으로 평가할 수 있다. % Difference의 경우 실제값의 평균과 예측값의 평균을 비교할 때 사용하며, 0에 근접할수록 예측이 잘 된 것으로 판단할 수 있다. Root Mean Square Error는 환경에서 측정된 값과 모형에서 예측된 값의 차이를 비교하는데 널리 사용된다. 그리고 TDI의 예측 등급과 실제 등급의 비교를 통해 예측 모형의 Accuracy Rate를 평가하였다. 또한 실제값과 예측값을 통계적인 분석을 통해 검증하였다.

3. Results and Discussion

3.1 피어슨 상관계수 분석을 통한 주요 인자 도출

TDI에 영향이 있는 인자를 도출하기 위해, 2017년 수생태 측정망 자료와 수리・수문・수질 자료를 Weka의 입력자료로 생성하고, Weka를 이용하여 피어슨 상관계수를 분석하였다. 수온, 용존산소, pH, 전기전도도, BOD, NH_3-N, T-N, PO_4-P, T-P, 수폭, 수심, 유속, TDI값 등 총 14개의 속성값을 사용하였고, 결과는 Table 1과 같다.

Table 1. Correlation analysis using Pearson-R

Factor	Value
Water temperature	-0.2457
DO	0.0892
pH	-0.028
Conductivity	-0.3909
Turbidity	-0.1071
BOD	-0.3653
NH_3-N	-0.3213
T-N	-0.4008
PO_4-P	-0.339
T-P	-0.3482
Water width	-0.0868
Average water depth	0.0467
Average flow velocity	0.1295

결과값을 봤을 때 $-0.4008 \sim 0.1295$의 범위에 있어 몇 가지 인자는 약한 상관관계가 있다고 할 수 있지만 그 수준이 높지 않고, 그 외 항목은 상관관계가 매우 낮기에 개별적인 인자만으로는 TDI와의 높은 상관성을 탐색하기엔 한계가 있는 것으로 판단된다.

3.2 CFS 기법을 통한 주요 인자 도출

앞서 사용한 피어슨 상관계수의 한계와 하나의 변수를 통한 예측 한계를 극복하기 위해 CFS 기법을 통한 탐색을 적용하였다. 피어슨 상관계수 분석을 할 때와 동일한 14개의 속성값을 사용하였고, Weka를 이용하여 CFS 기법을 사용하여 인자를 탐색한 결과 TDI값을 제외한 13개의 인자들 중 TDI와 상관성이 높은 인자는 수온, BOD, NH_3-N, T-N, PO_4-P, T-P, 수심 등 총 7개의 인자가 선정되었다.

3.3 랜덤포레스트 알고리즘을 이용한 예측

CFS 기법을 통해 선정된 7개의 인자들과 확보된 1,312개의 데이터로 랜덤포레스트 알고리즘을 이용하여 예측모형을 생성하였다. 생성된 예측모형을 평가하기 위해 2017년 한강유역 지류의 TDI 데이터를 이용하여 예측을 하고 그 결과를 실제값과 비교하였다. 예측 정확도를 평가하는 지수로 R^2와 % Difference, Nash-Sutcliffe Efficiency(NSE), Root Mean Square Error(RMSE) 값 등을 사용한 결과를 Table 2에 나타내었다. R^2와 % Difference(Duda et al., 2012), NSE(Moriasi et al., 2007)는 각 기준에 근거하여 평가하였다.

Table 2. Evaluation Index of TDI prediction model made by random forest algorithm - 1

Index	Value	Evaluation
R^2	0.93	Very Good
% Difference	-0.37	Very Good
Nash-Sutcliffe Efficiency	0.89	Very Good
Root Mean Square Error	8.22	-
Accuracy Rate	70.4 %	-

평가 결과 R^2는 0.93으로 매우 유의한 상관관계를 나타내었고, 이는 Figure 1의 그래프를 통해 확인할 수 있었다. % Difference 또한 -0.37로 나타났으며, NSE가 0.89, RMSE가 8.22로 나타나 예측값이 실제값을 잘 반영하는 것으로 나타났다. Accuracy Rate는 70.4 %로 나타났고, 일부 값은 등급 점수의 경계에서 분류되는 과정에서 약간의 오분류가 발생하였으나 전반적으로 예측에 근소한 영향을 주는 것으로 판단된다.

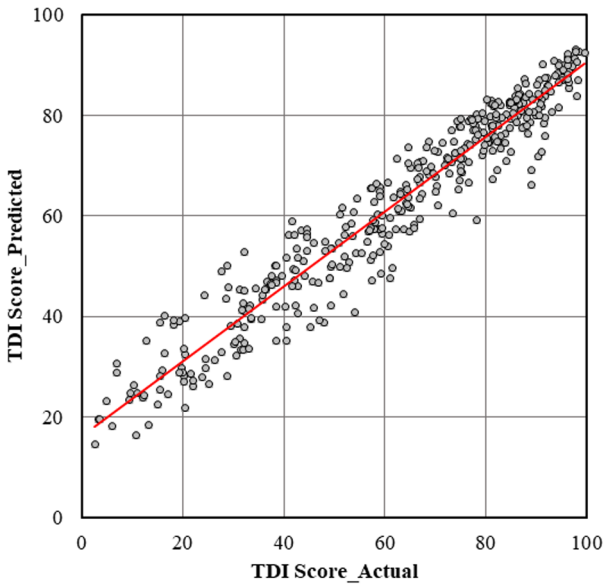

Fig. 1. Evaluation of TDI prediction model using 2017 dataset.

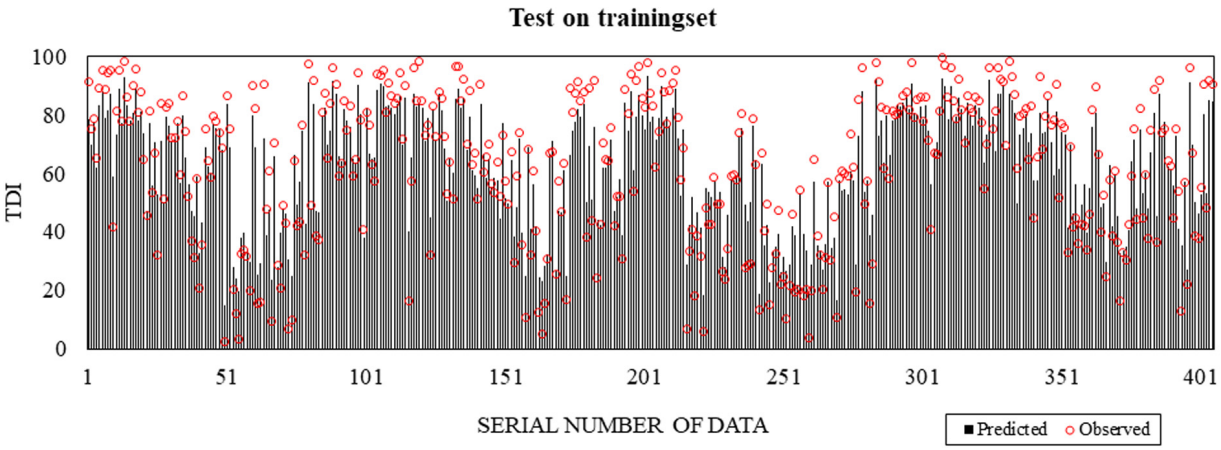

Fig. 2. Reproducibility examination result of TDI prediction model using 2017 dataset

Table 3. Evaluation Index of TDI prediction model made by random forest algorithm - 2

Index	Value	Evaluation
R^2	0.93	Very Good
% Difference	-2.37	Very Good
Nash-Sutcliffe Efficiency	0.89	Very Good
Root Mean Square Error	8.53	-
Accuracy Rate	69.7 %	-

2017년에 국한되지 않고, 예측 모형의 연속적 예측력을 평가하기 위해, 특정 지점에서 장기 시계열 예측을 적용하였다. 대상 지점은 TDI를 측정하는 한강유역 지류 중에 측정된 횟수가 17회 이상인 지점을 선별하여 적용함으로써 예측 모형을 검증 및 평가하였다. 이는 Table 3～4와 Figure 3～5에 평가결과를 나타내었다.

평가 결과 R^2가 0.93으로 강한 상관관계를 나타내었고, % Difference는 －2.37로 나타났으며, NSE가 0.89, RMSE가 8.53으로 나타나 2017년 데이터를 사용했을 때와 마찬가지로 예측값이 실제값을 잘 반영하는 것으로 나타났다. Accuracy Rate는 69.7%로 나타났는데, 2017년과 유사한 예측 정확도를 보였고 마찬가지로 등급 점수 경계의 값들로 인한 일부 값들

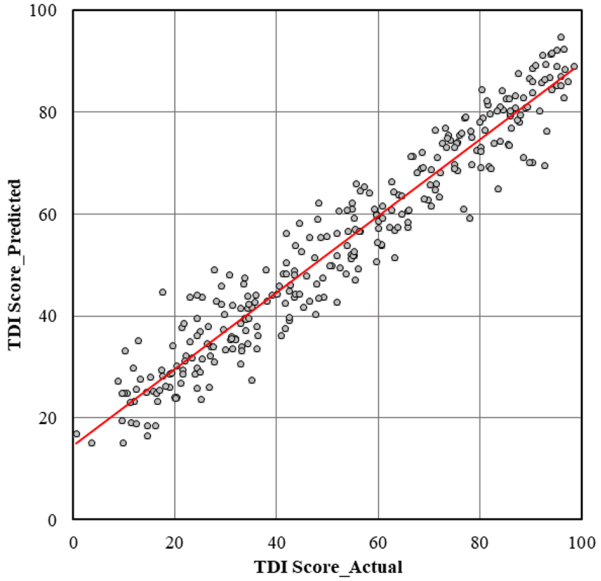

Fig. 3. Evaluation TDI prediction model using data at least 17 times measured

의 예측 오류인 것으로 판단되나 그 값들은 예측력에 근소한 영향을 준 것으로 사료된다. 또한, 지점별 시계열 분석 결과도 대체로 예측값이 실제값을 잘 반영하는 것으로 나타났다.

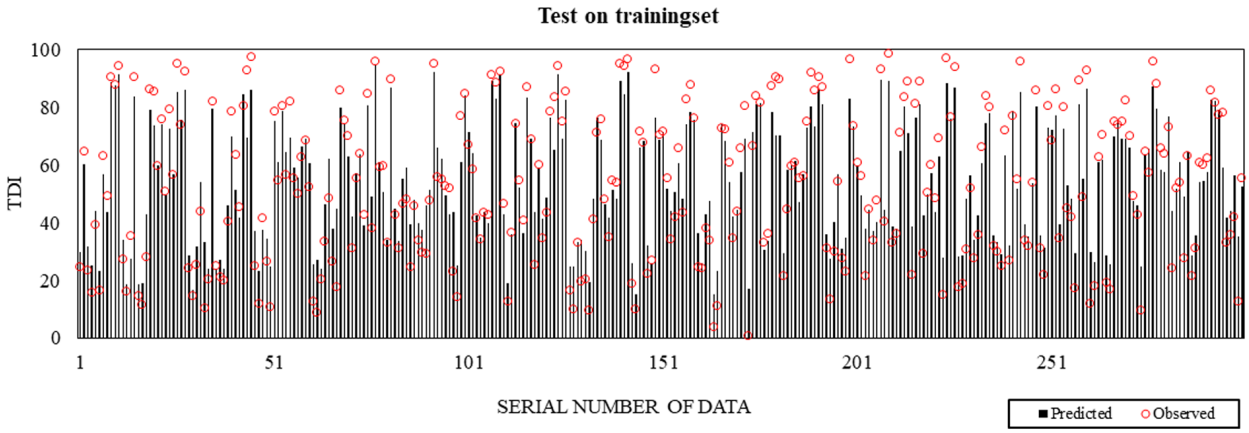

Fig. 4. Reproducibility examination result of TDI prediction model using data at least 17 times measured

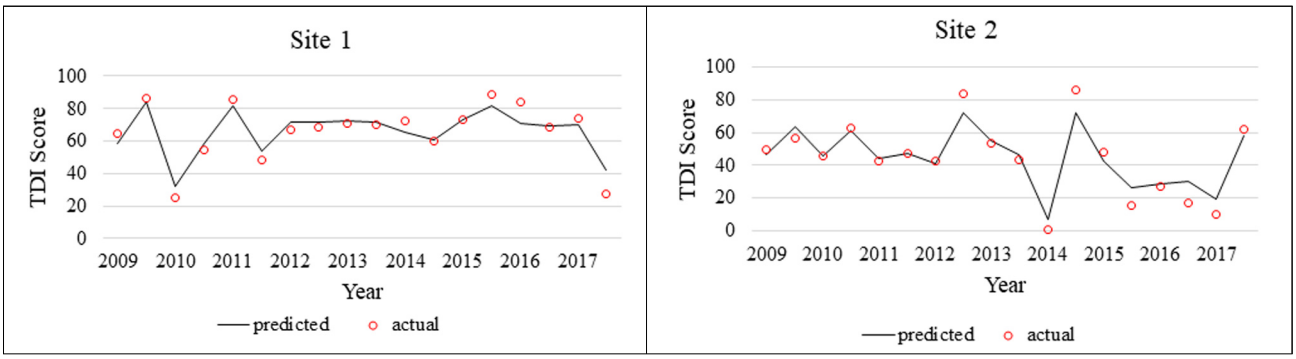

Fig. 5. Time series analysis of TDI prediction model using data at least 17 times measured

Table 4. Evaluation Index of TDI prediction model made by random forest algorithm - 3

Site \ Index	R^2	% Difference	Nash-Sutcliffe Efficiency	Root Mean Square Error	Accuracy Rate
1	0.9182	0.007	0.992	6.28	61.1 %
2	0.9469	1.767	0.981	8.61	66.7 %
3	0.9011	-4.059	0.979	7.13	83.3 %
4	0.9323	-3.628	0.977	7.53	83.3 %
5	0.8455	-1.148	0.982	6.92	72.2 %
6	0.8457	3.833	0.981	8.73	47.1 %
7	0.9544	-2.015	0.980	6.93	100.0 %
8	0.8928	6.050	0.992	7.68	44.4 %
9	0.9295	-9.557	0.952	9.01	72.2 %

Table 5. Results of Normality tests

	Kolmogorov-Smirnov			Shapiro-Wilk		
Index	Statistic	Degree of freedom	P-Value	Statistic	Degree of freedom	P-Value
Value	0.068	405	0.000127	0.983	405	0.000125

Table 6. Results of Wilcoxon Signed Ranks Test

	Negative Ranks	Positive Ranks	Total	Z	Asymp. sig. (2-tailed)
Value	218[a]	187[b]	405	-0.379[c]	0.705

a : predicted < actual
b : predicted > actual
c : Based on negative ranks

3.4 통계분석을 이용한 모형 검증

랜덤포레스트를 이용하여 생성된 모형의 검증을 위해 통계적 분석을 수행하였다. 2017년을 제외한 데이터로 학습을 한 모형에 2017년 데이터를 검증 데이터로 활용해 예측값을 생성하였다. 먼저 데이터의 정규성 검증을 위해 예측된 2017년의 데이터와 실측데이터간의 차이값을 이용하여 대응표본 t 검정을 수행하였다. 그 결과는 Table 5에 나타냈으며, 결과적으로 Kolmogorov-Smirnov 검정에서 P-value가 0.000127으로 0.05 이하의 값이므로 정규성을 만족한다는 귀무가설이 기각되고, 정규성을 갖지 않는다고 할 수 있다.

정규성을 갖지 않으므로 비모수 검정을 수행하였다. 실측값과 예측값을 변수로 하여 Wilcoxon Signed Ranks 검정을 수행한 결과를 Table 6에 나타내었다.

Wilcoxon Signed Ranks 검정 결과 근사 유의확률이 0.705로 0.05를 초과하는 값을 가지기 때문에 실측값과 예측값이 유의한 차이가 없다는 귀무가설을 채택한다. 즉 통계적으로 실측값과 예측값의 크기에 유의한 차이가 없다는 결과가 도출되었다.

4. Conclusion

본 연구에서는 수생태지표 중 하나인 TDI를 예측하기 위해 수질·수리수문 데이터를 구축하여 피어슨 상관계수 분석 및 CFS기법을 이용해 우선순위 인자를 도출하였다. 수온, BOD, T-N. NH_3-N, T-P, PO_4-P, 수심 등 총 7개의 선정된 인자들로 기계학습을 이용하여 예측 모형을 생성하였고, 생성된 모형을 평가하기 위하여 수생태 측정망 중 2017년 데이터셋과 17회 이상 측정된 지점들의 자료들로 모형으로 예측

된 값과 실제값을 비교 및 평가하였다. 평가 결과 모형의 예측치가 실측치를 비교적 합리적으로 재현할 수 있는 것으로 분석되었다.

이 연구 결과는 향후 수생태 건강성 평가 지수 및 그에 영향을 미치는 인자를 특정하고, 미래에 지속적 관리가 필요한 유역의 선정이나 수질 관리의 방향성을 제시하는 데 그 기초의 제공이 가능할 것으로 판단된다. 향후 연구에서는 다량의 관측 자료를 이용하여 기계학습을 시키고 모형을 검증하는 것이 필요할 것으로 사료된다.

머신러닝 기법을 활용한 예측 모형의 활용성으로 우선 수생태 모니터링이 이루어지지 않는 지점에 대해 Screening 단계에서 훼손지역을 검토해볼 수 있다. 그리고 기존 유역 수질 예측 기법과의 연계를 통해 수생태 건강성을 1차적으로 예측할 수 있을 것으로 사료된다. 이 모형의 경우 데이터의 지속적인 학습을 통해 예측력을 더 강화할 필요가 있다. 하지만 다양한 내·외부 환경 요소의 고도로 복잡한 영향 관계에 따라 달라질 수 있는 생태 특성상 Screening 단계에서 예측 및 대상지 사전검토 등에 국한해서 사용이 가능할 것이며, 한강수계의 지류만으로 학습된 모형이므로 금강, 영산강, 낙동강 등 특성이 다른 수계나 호소구간, 상류 영향이 지대한 본류 구간 등에는 적합하지 않을 것으로 사료된다. 따라서 향후 다양한 환경에 적용할 수 있는 예측 모형이 개발되면 활용성이 증대될 것으로 기대된다.

Acknowledgement

본 논문은 한강수계관리위원회 환경기초조사사업의 지원을 받아 수행되었습니다.

References

An, K. G., Lee, J. Y., and Jang, H. N. (2005). Ecological health assessments and water quality patterns in Youdeung stream, *Korean Journal of Limnology*, 38(3), 341-351. [Korean Literature]

Breiman, L. (2001). Random forests, *Machine Learning*, 45(1), 5-32.

Cho, I. H., Kim, H. K., Choi, M. Y., Kwon. Y. S., Hwang, S. J., Kim, S. H., and Kim. B. H. (2015). Distribution and species prediction of epilithic diatom in the Geum river basin, South Korea, *Korean Journal of Ecology and Environment*, 48(3), 153-167. [Korean Literature]

Choi, J. S., Lee, J. H., and Kim, H. S. (2017). The epilithic diatom community and biological water quality assessment of Naeseongcheon located at the upper region of Nakdong river, *Korean Journal of Ecology and Environment*, 50(4), 470-477. [Korean Literature]

Duda, P. B., Hummel, P. R., Donigian, Jr, A. S., and Imhoff, J. C. (2012). BASINS/HSPF: Model use, calibration and validation, *Transactions of the American Society of Agricultural and Biological Engineers*, 55(4), 1523-1547.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*, PhD Thesis, Department of Computer Science, The University of Waikato, New Zealand.

Hwang, S. J., Kim, N. Y., Won, D. H., An. K. K., Lee, J. K., and Kim, C. S. (2006). Biological assessment of water quality by using epilithic diatoms in major river systems (Geum, Youngsan, Seomjin River), Korea, *Journal of Korean Society on Water Environment*, 22(5) 784-795. [Korean Literature]

Kal, B. S., Park, J. B., Kim, S. H., and Im, T. H. (2017). Assessment of tributary water quality using integrated water quality index, *Journal of Wetlands Research*, 19(3), 311-317. [Korean Literature]

Kelly, M. G. and Whitton, B. A. (1995). The trophic diatom Index: a new index for monitoring eutrophication in rivers, *Journal of Applied phycology*, 7, 433-444.

Kong, D. S. (2002). Necessity and approach of establishing biological water quality standards, *Korean Journal of Environmental Biology*, 20(Special issue), 38-49. [Korean Literature]

Ministry of Land, Infrastructure and Transport (MOLIT). (2014). *Korea river catalog*, Ministry of Land, Infrastructure and Transport, 3-5. [Korean Literature]

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the American Society of Agricultural and Biological Engineers*, 50(3), 885-900.

Na, S. M., Lim, T. H., Lee, J. Y., Kwon, H. G., and Cheon, S. U. (2015). Flow rate · water quality characteristics of tributaries and a grouping method for tributary management in Nakdong river, *Journal of Wetlands Research*, 17(4), 380-390. [Korean Literature]

Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I - A discussion of principles, *Journal of hydrology*, 10(3), 282-290.

National Institute of Environmental Research (NIER). (2011). *National aquatic ecological monitoring program*, National Institute of Environmental Research, 7-8. [Korean Literature]

National Institute of Environmental Research (NIER). (2013). *Nationwide aquatic ecological monitoring program*, National Institute of Environmental Research, 329-350. [Korean Literature]

National Institute of Environmental Research (NIER). (2017). *Biomonitoring survey and assessment manual*, National Institute of Environmental Research, 1-75. [Korean Literature]

Prygiel, J. and Coste, M. (1993). The assessment of water quality in the Artois-Picardie water basin(France) by the use of diatom indices, *Hydrobiology* 269(270), 343-349.

Woo, S. Y., Jung, C. G., Kim, J. U., and Kim, S. J. (2018). Assessment of climate change impact on aquatic ecology health indices in Han river basin using SWAT and random forest, *Journal of Korea Water Resources Association*, 51(10), 863-874. [Korean Literature]