Journal of Korean Society on Water Environment, Vol. 35, No. 5, pp. 425-431 (September, 2019)
pISSN 2289-0971 eISSN 2289-098X https://doi.org/10.15681/KSWE.2019.35.5.425

머신러닝 기반 CFS(Correlation-based Feature Selection)기법과 Random Forest모델을 활용한 BMI(Benthic Macroinvertebrate Index) 예측에 관한 연구

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

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

A Study on the prediction of BMI(Benthic Macroinvertebrate Index) using Machine Learning Based CFS(Correlation-based Feature Selection) and Random Forest Model

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

[1]ETWATERS Co., Ltd.
[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

Recently, people have been attracting attention to the good quality of water resources as well as water welfare. to improve the quality of life. This study is a papers on the prediction of benthic macroinvertebrate index (BMI), which is a aquatic ecological health, using the machine learning based CFS (Correlation-based Feature Selection) method and the random forest model to compare the measured and predicted values of the BMI. The data collected from the Han River's branch for 10 years are extracted and utilized in 1312 data. Through the utilized data, Pearson correlation analysis showed a lack of correlation between single factor and BMI. The CFS method for multiple regression analysis was introduced. This study calculated 10 factors(water temperature, DO, electrical conductivity, turbidity, BOD, NH_3-N, T-N, PO_4-P, T-P, Average flow rate) that are considered to be related to the BMI. The random forest model was used based on the ten factors. In order to prove the validity of the model, R^2, %Difference, NSE (Nash-Sutcliffe Efficiency) and RMSE (Root Mean Square Error) were used. Each factor was 0.9438, -0.997, and 0,992, and accuracy rate was 71.6% level. As a result, These results can suggest the future direction of water resource management and Pre-review function for water ecological prediction.

Key words : Aquatic ecology, BMI, CFS(Correlation-based Feature Selection), Machine Running, Random Forest

[1a] 연구원(Researcher), wsgo@etwaters.co.kr, 0000-0002-3068-6379
[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

현재 하천은 인간에게 문화, 사화, 경제 등 다양한 가치를 지니게 되었고, 국민의 일상생활 및 휴식과 밀접한 공간이 되었다. 국민에게 직간접적인 영향을 미치는 요소로 수생태계 건강성에 대한 중요성이 증대되고 있으며(KISTEP, 2019) 특히 국민의 관심은 본류에서 생활권에 밀접한 지류로 이동하고 있는 추세이다.

이에 환경부는 수생태계 건강성을 지속적으로 확보하고 증진시키기 위해 수생태계 건강성 증진을 목표로 물 환경의 보전과 개선을 위하여 수질 및 수생태계 보전을 위한 법률을 물 환경보전법으로 개정함으로써 물 환경의 통합관리를 추구하고 있으며(Park, 2018) 생물지표를 이용한 수질 관리방안을 마련하고 수생태 건강성 측정망 운영을 통해 전국 단위의 기초자료 확보 및 축적을 위해 노력하고 있다.

하지만 자연적 요인의 변화, 과도한 개발, 산업활동으로 인한 오염원의 영향 등 수생태 건강성의 위협요인들은 다양화·다변화되고 있다. 특히, 환경 변화에 따른 민감도는 본류에서보다 지류에서 더 크지만, 국가 차원에서의 관리는 4대강 본류 위주로 국한되어 있어 이에 대한 대응은 미흡한 실정이며 지류의 수자원 관리를 위한 활용할 수 있는 데이터 활용방안이 한계가 있다. 이에 따라 국민의 물복지를 향상할 수 있는 지류에 건강한 수생태계 환경을 제공할 필요가 있다. 그러나 이를 위한 수생태 변동성 측정은 현장 모니터링에 의존, 진단하므로 예측·예방에는 미흡하며 이에 대한 편의성을 위한 대응책이 필요한 실정이다.

이에 활용하려는 방안으로 현재 축적된 데이터베이스를 활용하여 지류·지천 수생태계 건강성 측정에 사용할 방법을 분석하고자 머신러닝(Machine Learning)을 활용하였다. 최근 직접적인 조작 없이 컴퓨터 스스로 데이터를 수집, 학습하여 예측하는 인공신경망 기반의 머신러닝을 활용한 연구가 많이 이루어지고 있다. 머신러닝은 학습된 데이터를 바탕으로 매번 발전된 결과를 활용하여 예측하는 장점이 있는데, 아직 수자원 분야에 적용 연구는 미비한 실정이다(Kum et al., 2017). 수생태 측정망 자료 분석에 몇 가지 하천 환경 요소는 일정 수준 이상 상관성이 있는 것으로 판단할 수는 있으나 선형관계에서 그 수준이 높지 않고, 그 외 항목은 매우 낮아 각각의 개별적 요소로는 수생태 지표와 연관성이 없다고 할 수 있다.

따라서 본 연구에서는 한강의 지류·지천을 대상으로 머신러닝을 도입하였으며 Correlation based Feature Selection (CFS) 기법을 통해 수생태 건강성 지수와 연관이 있는 인자들을 도출하고, 관련 인자로부터 랜덤 포레스트(Random Forest) 모델을 적용하여 수생태 건강성 지수중 저서성 대형무척추 동물지수(BMI)의 실측치와 예측치를 비교하였고 예측된 BMI를 실측치와 비교하여 그 예측력을 평가하였다.

2. Materials and Methods

2.1 대상 자료 선정

본 연구는 한강 유역을 대상으로 하였으며 한강은 총 면적

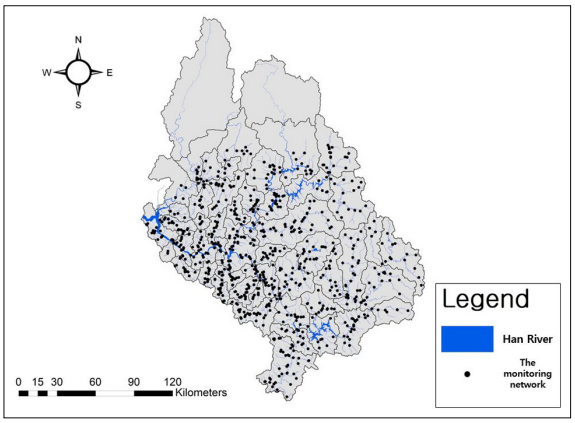

Fig. 1. Status of the ecological monitoring network in Han River.

은 약 31,648 km^2, 하천 연장은 약 917 km의 국가하천, 약 7,662 km의 지방하천으로 이루어져있다(MOLIT, 2014). 또한 국가 하천 19개소와 지방하천 895개소로 분류할 수 있으며 한강 수계 하천의 개소수는 총 914개이다.

한강수계의 2016년~2018년 3년간 축적된 데이터의 생물측정망 조사지점은 총 907개이며, 본 연구에서 활용된 자료는 전체 데이터 중 '지류'로 구분되어있는 157개 지점이다. 10개 지점(섬강A, 섬강B, 경안A, 경안B, 북한D, 한강F, 양화A, 청미A, 복하A, 홍천A)은 2009년~2017년, 10년간 축적된 자료를 활용하였고, 나머지 지점은 2016년~2018년 3년간 축적된 자료를 바탕으로 이루어져 있다. 축적된 데이터는 총 1312개이다(Fig. 1.).

각 지점의 수리·수질적 하천환경요소는 향후 머신러닝 학습 결과와 연계하여 미래의 BMI를 예측할 수 있어야 한다. 즉, 모형을 통해 분석의 대상이 되는 속성항목은 모형으로부터 도출된 결과와 상관성이 있어야하고 예측 가능한 항목이어야 하며 BMI와 상관분석이 이루어져야 한다. 따라서 본 연구에서는 생물측정망 조사자료로부터 연관성이 있는 항목들과 BMI를 추출하여 자료를 구축하였으며 정확도 향상을 위해 하천환경요소의 조사결과가 누락되거나 측정 불가능한 항목은 모두 제거하였다.

2.2 하천환경요소 데이터 선정 기준

각 인자와 BMI간 연관성을 파악하기 위한 피어슨 상관계수 산정에 앞서 생물측정망 지점 데이터의 각 속성마다 우선순위를 부여하였다.

1순위 대상으로 수계 내 모든 생물에 영향을 줄 수 있고 모형을 통한 예측결과를 직접적으로 활용할 수 있는 인자를 고려하였다. 이 중 하천의 물리적인 측면인 규모와 하안 부지 등을 결정할 수 있는 인자로서 수폭과 평균수심을 선정하였다. 또한 하천 유량과 하상의 고도지형에 큰 영향을 줄 수 있는 요소로 판단되는 평균 유속을 1순위 항목으로 선정하였다.

이화학적 수질요소로서 수온과 용존산소(DO)는 모든 생물군이 직접적 영향을 받으며 BOD는 분해 가능한 유기물의 지표로서 DO를 고갈시킨다는 부분에서 주요 인자로 판단하

Table 1. The data property selection for the correlation analysis between environmental factors and the aquatic ecosystem

	factort	
	Physical element	Physicochemical Elements
1st priority	Average velocity, Average depth, Width	Water temperature, DO, BOD, NH_3-N, PO_4-P
2nd rank	-	pH, T-N, T-P
3rd rank	-	EC, Turbidity

였다. NH_3-N 및 PO_4-P는 하천 내 수생식물 및 조류가 다른 형태의 영양물질에 비해 우선적으로 흡수할 수 있는 물질로 독립영양생물의 생장과 상위영양단계의 생물에 영향을 미칠 수 있다는 점에서 1순위 항목으로 선정하였다.

2순위 대상으로는 수질관리항목인 T-N, T-P 항목과 pH를 선정하였다. 이는 태별 영양물질에 비해 더욱 취득이 용이하며, 향후 예측결과와 직간접적으로 연계할 수 있는 항목이라는 점에서 2순위 항목으로 선정하였다.

3순위 대상으로 예측결과와 직접적으로 연계할 수는 없으나 환산식 혹은 수질항목간 회귀분석을 통해 간접적으로 활용할 수 있고 하천 환경을 추정하는데 활용할 수 있거나 생물에게 직간접적 영향을 미칠 수 있는 항목으로 선정하였다(Table 1.).

2.3 저서성 대형 무척추 동물 지수(Benthic Macroinvertebrate Index)

저서동물은 수서곤충, 조개류, 갑각류, 거머리 등의 하천바닥에 서식하는 무척추동물을 말하며 생태적 중요성과 환경지표성이 크다. 저서동물지수는 수생태 환경 평가에 가장 폭넓게 활용되며 저서동물지수(BMI지수)는 A-E등급, 5개 등급으로 나눌 수 있고 해당 과정을 다음과 같이 산정할 수 있다(Table 2.).

Table 2. The BMI calculation method

Formula	Index range	Class	State
$BMI = \left(4 - \frac{\sum_{i=1}^{n} s_i h_i g_i}{\sum_{i=1}^{n} h_i g_i}\right) \times 25$	$80 \leqq BMI \leqq 100$	A	very poor
	$65 \leqq BMI < 80$	B	good
	$50 \leqq BMI < 65$	C	medium
s_i: Unit indecision index h_i: Appearance g_i: Indicator weights	$35 \leqq BMI < 50$	D	poor
	$0 \leqq BMI < 35$	E	very poor

s_i는 단위오탁지수, h_i는 출현도, g_i는 지표가중치로서 해당 계산식에 따라 지수를 산정하여 점수에 따라 5개의 등급으로 분류 된다.

2.4 환경인자와 생물지표 간 피어슨 상관관계 분석

수생태 건강성 지표 중 상관성이 높은 항목을 도출하기 위해 통계분석인 피어슨 상관관계 분석을 활용하였다. 수생태 측정망 자료와 수리・수문・수질 자료를 입력 자료로서 총 14개의 항목(수온, 용존산소, pH, 전기전도도, BOD, NH_3-N, T-N, PO_4-P, T-P, 수폭, 수심, 유속, BMI값)을 입력하였고, 각각의 상관계수를 도출하였다. 피어슨 상관계수는 종속변수와 독립변수간 상관성을 파악하기 위해 널리 사용되는 방법이다. 그 값은 ±1 사이로 나타나며 ±0.5 수준 혹은 그 이상의 값이 나타나는 경우 높은 상관성을 나타낸다고 판단할 수 있다. ±0.2 ~ 0.5 수준일 경우 상관성이 없다고 할 수는 없지만 모호한 수준으로 판단할 수 있다. 한편, 이러한 피어슨 상관계수는 두 변수의 상관관계가 서로 선형(1차 함수)으로 표현 가능할 때는 유용하다. 이를 통해 단일 인자와 수생태 지수간 상관관계를 분석을 실시하였다.

2.5 CFS 기법을 통한 주요 속성집합 탐색

선형관계에서 적용되는 피어슨 상관계수를 통해 단일 인자와 BMI 사이의 상관분석을 통한 예측 결과를 통해 BMI와 단일인자 사이의 상관성을 판단하는데 한계가 있다고 판단하였으며 비선형 다중 회귀식을 통한 분석의 필요성을 파악하였다. 다중 인자의 분석에 앞서 수식의 특정 형태를 예측할 수 없다는 점에서 통계적 범주를 벗어나 머신러닝을 도입할 필요가 있을 것으로 판단하였다.

예측에 활용될 주요 인자집합을 찾기 위해 CFS 기법을 적용하였다. CFS 기법은 여러 가지 후보 인자 중 상관성을 기반으로 예측모델에 사용될 대표적인 속성들을 추출하기 위한 머신러닝 기법의 하나다. CFS기법은 Best-first 방법 즉, 추정해야 할 결과와 상관성이 크면서, 동시에 다른 속성들과의 상관성은 낮은 속성을 탐색하는 방법이다(Lee, 2012). 불필요한 변수 집합을 축소하여 속도를 향상시키며 연관성 있는 속성만을 선택적으로 추출한다는 점에서 큰 이점을 갖는다(Hall, 1999). 앞서 시도한 피어슨 상관계수 분석과 비교하여 추후 연구에서 활용하기 위해 동일 프로그램(Weka)을 통해 변수로 작용하는 인자를 탐색하였으며 CFS기법이 사용되었다. 이를 통해 10개의 인자를 도출하였다(Fig. 2.).

Fig. 2. The machine learning software 'weka' start screen.

2.6 랜덤포레스트 모델 적용

Breiman (2001)이 고안한 랜덤포레스트 모델은 데이터를 분류하는 기법으로서 의사 결정 나무(decision tree)에 기반한다. 학습 과정에서 구성한 여러 개의 의사 결정 트리들로부터 분류 또는 평균 예측값(회귀분석)을 출력함으로써 동작하는 알고리즘으로, 학습 단계에서 여러 개의 의사결정트리를 구성하고, 결과를 분류하거나 평균을 예측하는 과정을 거치게 된다(Choi and Seo, 1999). 의사결정트리는 구축된 데이터의 특성과 얻고자하는 결과를 반영하여 분리 기준(split criterion)과 정지 규칙(stopping rule)으로 규칙을 정하여 최종적인 의사결정 트리를 얻게 된다. 해당 모델이 결과값을 도출하는 과정은 다음과 같다(Lee and Chung, 2019)(Fig. 3).

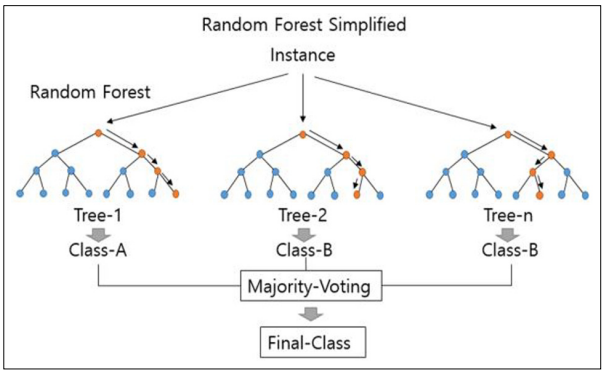

Fig. 3. The random forest structure.

이 기법은 가지치기(pruning)를 통해 오류 야기할 수 있는 데이터의 가지를 제거하고, 교차 타당성 오류(cross validation error)를 확인하여 최종적으로 교차 타당성 오류를 최소화하고 반응변수와 설명변수를 가장 잘 연결하는 트리를 구축하게 된다(Kim and Park, 2019).

본 연구에서 생물측정망에서 축적된 조사자료는 머신러닝 학습에 활용되었기 때문에 개발된 BMI 예측 모델의 재현성 검토를 위한 별도의 검증용 데이터 확보가 어려운 상황이므로 머신러닝 학습에 활용된 동일한 데이터를 이용하였다. 축적된 데이터를 활용하여 회귀분석을 실시하였으며 이를 통해 도출한 예측 모델을 기반으로 2017년 1개년의 하천환경요소를 통한 BMI를 예측하였고 실측값과 비교하였다.

2.7 예측모델 검증

예측 모델 검증을 위해 결정계수(R^2), Root Mean Square Error(RMSE)와 %Difference(%diff.), Nash-Sutcliffe 계수(NSE)를 활용하였다.

R^2는 0~1사이의 값을 가지며, 종속변인과 독립변인의 상관관계가 높을수록 1에 가까워진다. 결정계수가 1에 가까운 값을 가지는 회귀모형은 유용성이 높고, 결정계수의 값이 0에 가까울수록 낮다고 할 수 있으며 R^2의 산정식은 다음과 같으며, 값에 따른 등급은 다음(Donigian, 2000)과 같다(Table 3.).

$$R^2 = \left(\frac{\sum_{i=1}^{n}(O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2}\sqrt{\sum_{i=1}^{n}(P_i - \overline{P})^2}} \right)^2 \quad (1)$$

%diff.와 RMSE를 활용하여 예측모델을 통한 BMI와 실제 측정된 BMI간의 분포를 비교하였다.%diff.는 실측값과 모델 예측값의 적합도를 평가하는 지표이며 해당 식은 다음(2)식과 같다. 이를 통해 도출된 결과값은(Donigian, 2000)에서 제시한 분류 등급을 통해 산출 하였다(Table 3.).

$$\%diff. = \frac{\left| \sum_{i=1}^{n} O_i - \sum_{i=1}^{n} P_i \right|}{\sum_{i=1}^{n} O_i} \times 100 \quad (2)$$

RMSE값은 모의결과를 통해 산출된 BMI와 실제 BMI의 평균 오차로 모델의 정밀도를 나타내며 산정하는 식은 다음 식(3)과 같다.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (3)$$

NSE는 수문학 모델의 정확성을 검증하는데 주로 사용되는 값으로서 다음 식(4)과 같다(Nash and Sutcliffe, 1970).

$$NSE = 1 - \left(\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O_i})} \right) \quad (4)$$

NSE는 O_i는 측정값, $\overline{O_i}$는 측정값의 평균, P_i가 모델값이다. NSE는 선형 회귀분석의 결정계수(R^2)와 같은 개념이며, $-\infty$ ~ 1 범위의 값을 가진다. 해당 값이 1이면 측정값

Table 3. The recommended general performance rating for R^2 and %difference

		Very good	Good	Fair	Poor
R^2	Hydorology/Flow	> 0.8	0.7 ~ 0.8	0.6 ~ 0.7	0.6 >
%Difference.	hydrology/Flow	< 10	10 ~ 15	15 ~ 25	25 <

Table 4. The recommended general performance rating for the NSE

		Very Good	Good	Satisfactory	Usatisfactory
NSE	Hydrology/Flow	> 0.75	0.65 ~ 0.75	0.50 ~ 0.65	0.50 >

과 모델 값의 완전한 일치를, 작을수록 측정값과 모델값이 덜 일치함을 의미한다(Kim et al., 2014). NSE의 값에 따른 등급은(Moriasi et al., 2007)에서 다음과 같이 제시하였다 (Table 4.).

3. Results and Discussion

최종적으로 구축된 1312개의 데이터를 통한 피어슨 상관분석 결과, 다섯 가지 항목(T-N, Turbidity, T-P, BOD 등)은 일정 수준 상관성이 있는 것으로 판단할 수는 있으나 수치가 높지 않으며, 그 외의 항목들은 매우 낮은 수치로서 개별적 요소와 BMI간 높은 상관성이 있다고 판단하기 어려웠다 (Table 5.).

Table 5. Analysis of the individual selection attributes and the BMI correlation coefficient

Factor	Correlation coefficient
Average flow rate	0.20937
Average depth	0.038
Water width	-0.0713
Water temperature	-0.2576
PO_4-P	-0.3849
T-N	-0.45868
Turbidity	-0.50636
DO	0.11798
pH	0.00758
NTU	-0.20246
NH_3-N	-0.37489
T-P	-0.41233
BOD	-0.47057

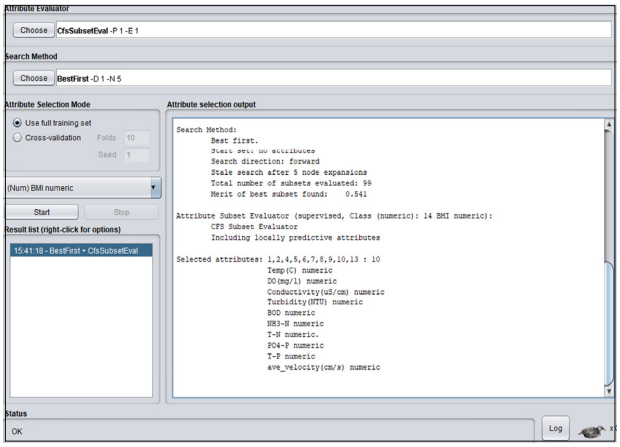

Fig. 4. The deduction of river environmental factors through the CFS explorer technique.

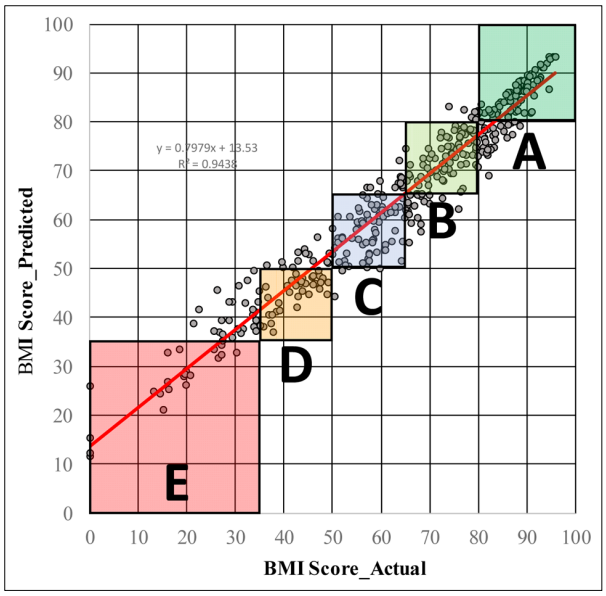

Fig. 5. The linear graph of the measured value and predicted value of the BMI.

이로 인해 다중회귀분석의 필요성을 파악하였고, 머신러닝 기법중 CFS 탐색 기법을 도입하였다. 머신러닝 프로그램 'Weka'를 이용하였으며 이를 통해 BMI와 10가지 인자(수온, DO, 전기전도도, 탁도, BOD, NH_3-N, T-N, PO_4-P, T-P, 평균유속)의 연관성을 확인하였다(Fig. 4.).

10가지 인자를 도출하여 랜덤포레스트 모델을 통해 개발된 학습모델이 타당한지 판단하기 위해 전체 학습데이터의 BMI를 제외한 하천환경요소만을 입력하여 BMI 예측을 실시하였다. 예측 점수를 통한 BMI등급과 실제 등급의 일치율은 80.1 %로 확인하였다.

해당 모델을 통한 예측 가능성을 검증하기 위해 개발된 모델을 통해 17년도의 BMI 측정 결과와 비교하였다. 이를 통한 재현성 검토에서 예측 모델을 통한 등급과 실제 등급과 예측 등급의 일치율은 71.6 %로 나타났으며 R^2, 정확도, %diifrence, NSE, RMSE 값은 다음과 같다(Table 3.).

BMI 예측 모델의 재현성 검토 결과 모델을 통한 실측값과 예측 등급의 일치율은 71.6 %로 나타났다. 점수를 통해 등급 환산하는 과정에서 등급의 경계값과 비슷한 위치에 존재하는 경우로 인하여 등급의 일치율은 다소 낮게 나타났다, 결정계수(R^2)=0.9438로 1과 근사한 값으로 높은 상관성이 나타났다(Fig. 3.). 재현성을 검토하기 위한 값인 %difference, NES, RMSE 값은 각각 -0.097, 0.992, 5.9510으로 실측값과 예측값의 정밀도와 정확도 면에서 긍정적인 수치로 나타났다(Table 6.).

Table 6. The BMI prediction model execution result by the random forest algorithm

	R^2	Accuracy Rate	%diff.	NSE	Root Mean Square Error
Value	0.9438	71.6%	0.997	0.992	5.9510
Raiting	Very Good	-	Very Good	Very Good	Very Good

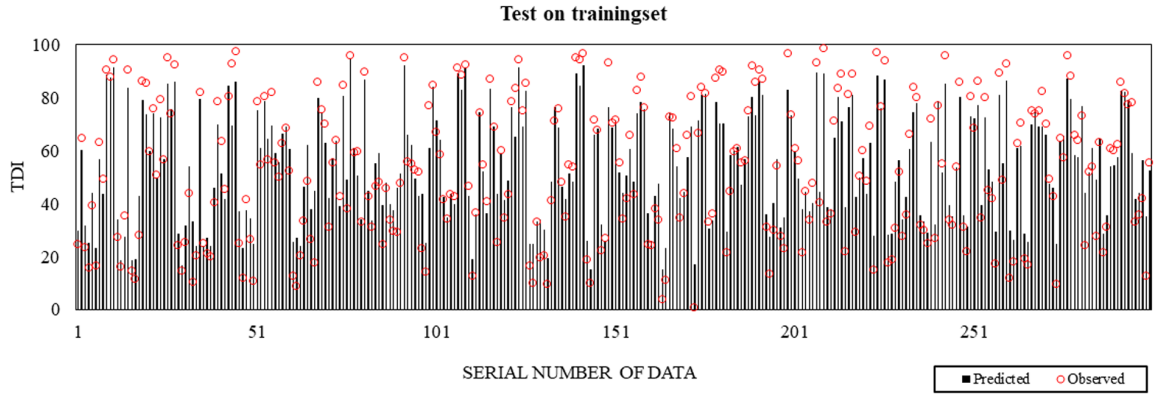

Fig. 6. The BMI prediction model reproducibility review result.

4. Conclusion

현재 국민의 물복지를 실현하기 위한 정책적 노력이 이루어지고 있다. 하지만 일상생활과 밀접한 지류의 에서의 활용 방안은 다소 미흡하며. 현장 모니터링에 의존하는 수생태 건강성 측정에도 적용된다. 수생태지표중 하나인 BMI를 예측하기 위해 앞서 구축된 수리 · 수질 · 수문 요소를 구축, 피어슨 상관계수를 통한 상관계수 분석을 통해 해당 단일 인자와 BMI 간 상관성은 작은 것으로 판단하였다. 이를 통해 머신러닝 도입 필요성을 파악하였으며 CFS기법을 활용하여 다중 인자를 추출하였고 결과를 바탕으로 랜덤 포레스트 모델 적용을 통해 예측모델 검정을 실시하였다. BMI 수치와 관련성이 있는 인자로 수온, DO, pH, 전기전도도, 탁도, BOD, NH_3-N, T-N, PO_4-P, T-P, 평균 하폭, 평균 수심, 평균 유속 10가지 인자를 도출하였고, 이를 바탕으로 하여 Random Forest모델을 적용하였으며 데이터의 예측값(회귀분석)을 출력하여 예측모델을 개발하였다. BMI를 산정하여 예측값과 실측값을 통해 산정된 지수(점수)를 다시 기준에 따라 A ~ E 등급으로 구분하였다.

지점별 BMI 실측값과 머신러닝을 이용한 예측값을 비교했을 때, 결정계수(R^2), %difference, NSE, RMSE값은 각각 0.9438, -0.997, 0,992, 5.9510의 긍정적인 수치로 나타났으며 71.6%의 BMI 등급 적중률을 확인할 수 있었다.

본 연구를 통해 머신러닝을 이용하여 생물측정망을 통한 하천환경요소를 통해 BMI 예측이 이루어졌다. 단일 인자와 BMI 사이의 상관관계는 모호하다고 판단되었고, CFS 분류기법을 통한 다중 회귀분석을 이용하여 BMI와 하천환경요소간 상관관계를 나타내었다. 결과를 통해 도출한 10가지 인자를 이용하여 Random Forest모델을 도입하여 개발된 BMI 예측 모델을 제시해 보았다. 이를 통해 BMI에 영향을 미치는 영향요인을 좀 더 구체화하여 수자원 관리의 방향성을 제시하는 기초자료로서의 활용을 기대할 수 있을것으로 판단되며, 현장 모니터링에 의존하는 수생태 측정의 편의성을 제시해 봄으로써 데이터의 활용 방안이 부족한 지류 수생태계의 건강성 확보를 위한 사전 예측을 위한 자료로서의 기능을 할 수 있을것으로 판단된다. 특히, 본류에 비해 외부 환경요인 변화에 민감한 지류의 특성상 대비 차원의 사전 검토를 위한 사전적 수단으로 활용될 수 있을 것이다.

하지만 CFS 분류 기법을 통해 BMI와 상관성이 있다고 도출한 10가지 인자를 확인함에 있어 각 인자의 상대적 중요도에 대한 부분을 파악하기에 어려움이 있었으며, 추후 해당 부분에 대한 보완을 통해 수자원 관리에 세부적인 관리가 가능할 것으로 판단된다. 또한 예측 모델을 강화, 개발하기 위하여 추후 지속적인 양질의 데이터를 확보를 필요로 하며, 학습횟수를 증가시킨다면 해당 모델의 예측력이 증가할 수 있다고 판단하였다. 해당 모델은 한강 수계의 지류만을 대상으로 학습된 모델로 다른 수계 또는 본류에 적용하기 어려울 것으로 판단 되므로 향후 다양한 범주에 이용될 수 있는 모델의 개발 또는 다양한 수계의 데이터 구축으로 수생태 건강성 확보를 위한 자료로 이용할 수 있을 것으로 기대된다.

Acknowledgement

This research was supported by a grant from Environmental Basic Research program by Commuttee for Management of Han River Basin

본 논문은 한강수계관리위원회 환경기초조사사업의 지원을 받아 수행되었습니다. 이에 감사드립니다.

References

Breiman, L. (2001). Random Forests, *Machine Learning,* 45(1), 5. Available at: http://search.ebscohost.com.proxy.konkuk.ac.kr:8080/login.aspx?direct=true&db=edo&AN=ejs37250840&lang=ko&site=eds-live&scope=site (Accessed: 1 October 2019).

Choi, J. H. and Seo, D. S. (1999). Decision trees and its applications, *Journal of The Korean Official Statistics*, 4 (1), 61-83. [Korean Literature]

Donigian, Jr. A. S. (2000). *HSPF training workshop handbook and CD, Lecture #19, Calibraion and Verification Issures*, Slide #L19-22 EPA Headquarters, Presented and prepared for US EPA.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*, PhD Thesis, Department of Computer

Science, The University of Waikato, New Zealand.

Kim M. R. and Park M. H. (2019). An analysis of the characteristics of college students according to first-time participation in private tutoring using a random forest, *CNU Journal of educational studies*, 40(1), 1-33. [Korean Literature]

Kim S. H., Lee E. J., Na J. S., and Choi J. W. (2014). Calibration of an UV distribution model by Nash-Sutcliffe efficiency coefficient, *Korean Society of Civil Engineers*, 1813-1814. [Korean Literature]

Korea Institute of Science & Technology Evaluation and Planning (KISTEP). (2019). *Technology development project for securing the ecosystem health*, https://www.kistep.re.kr/c3/sub2_4.jsp?brdType=R&bbIdx=12605. 63-67. [Korean Literature]

Kum D. H., Ryu J. C., Sung Y. S., Han J. H.. and Lim G. J. (2017). P-8 : Development and Assessment for extended daily streamflow regression equation of TMDL station using Machine Learning, *Proceedings of the 2017 Spring Co-Conference of the Korean Society on Water Environment and Korean Society of Water and Wastewater*, Korean Society on Water Environment and Korean Society of Water and Wastewater, 289-290. [Korean Literature]

Park, H. J. (2018). *The study of local government organizational reform guideline for integrated water management*, http://www.prism.go.kr/homepage/entire/retrieveEntireDetail.do?research_id=1480000-201800124 (accessed Feb. 2018), Ministry of Environment. 4-19. [Korean Literature]

Lee. H. S. (2012). *A study on CFS-variable subset selection method for classification*, Doctor's Thesis. Sungkyunkwan University, 1-8. [Korean Literature]

Lee H. J. and Chung G. H. (2019). Categorical prediction and improvement plan of snow damage estimation using random forest, *Journal of Wetlands Research*, 21(2), 157-162. [Korean Literature]

Moriasi, D. N., Arniold, 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 ASABE*, 50(3), 885 – 900. doi: 10.13031/2013.23153.

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