Prediction of Health Evaluation Indices for Aquatic Ecosystem using Extreme Gradient Boosting Tree (XGBoost) and SWAT

Chung Gil Jung
(wjd0823@konkuk.ac.kr)
So Young Woo and Seong Joon Kim

School of Civil and Environmental Engineering, College of Engineering, Konkuk University, South Korea
In South Korea, since 1990s, the people became interested in restoration of river environmental function for their life quality with water.

Since 2007, the Ministry of Environment has monitored the aquatic ecosystem health (AEH) and evaluated the stream and river AEH.

By the limitations of monitoring sites, we need some technique to develop AEH evaluation for the whole country streams.

This study is to predict the AEH indices (FAI, TDI, and BMI) of ungauged streams for a standard watershed scale (about 500 km²) using SWAT results and Ensemble Machine Learning algorithm.
Assessment Procedure

Ecological Health Monitoring Data (2008-2015) at 320 locations in Han-river basin

- Fish Assessment Index (FAI)
- Trophic diatom index (TDI)
- Benthic macroinvertebrate index (BMI)
- Total P
- Phosphate (PO$_4^{-}$)
- Total N, Nitrate (NO$_3^{-}$)
- Ammonium (NH$_4^{+}$)
- Water temperature (WT)
- Stream discharge (SD)

Correlation Analysis Between data

- FAI, TDI, BMI versus T-P, PO$_4^{-}$, T-N, NO$_3^{-}$, NH$_4^{+}$
- T-P, PO$_4^{-}$, T-N, NO$_3^{-}$, NH$_4^{+}$ versus WT, SD

Prediction of Ecological Health Indices

- Development of XGBoost algorithm
- Input variables
  (SWAT) WT, SD, T-P, PO$_4^{-}$, T-N, NO$_3^{-}$, NH$_4^{+}$
  (monitoring) FAI, TDI, BMI

SWAT Model set up

- Calibration and validation of Streamflow, ET, SM, T-N, T-P using 11 years (2005-2014) observed data

SWAT Model set up

- Calibration and validation of Streamflow, ET, SM, T-N, T-P using 11 years (2005-2014) observed data

SWAT Model set up

- Calibration and validation of Streamflow, ET, SM, T-N, T-P using 11 years (2005-2014) observed data
The largest river basin in South Korea (Han, Geum, Yeongsan, Seomjin, Nakdong)

Han River basin (34,148 km²)

- Average precipitation 1254 mm
- Average temperature 11.5°C
Data for SWAT model evaluation

GIS data

Elevation: 0 - 1650m (SRTM 90m grid size)

Soil: Loam (24%) and sandy loam (58%)

Land cover (2008): Forest (73%) and paddy rice (6%)

(a) Elevation map
(b) Soil texture map
(c) Land cover map
(d) Han River Basin map
Data for SWAT model evaluation

4 Multipurpose dam data (area-level and storage-level relationship curve)

- **Soyang dam (SYD)**
  - Total storage: 2.9 billion m³
  - Sub-basin area: 2,694 km²
  - (the largest in South Korea)

- **Hoengseong (HSD)**
  - Total storage: 87 million m³
  - Sub-basin area: 209 km²

- **Chungju dam (CJD)**
  - Total storage: 2.8 billion m³
  - Sub-basin area: 6,662 km²
  - (the second largest in South Korea)

- **Paldang dam (PDD)**
  - Total storage: 244 million m³
  - Sub-basin area: 23,539 km²
3 Multifunction weir data (area-level and storage-level relationship curve)

- **Ipo weir (IPW)**
  - Total storage: 17 million m³

- **Yeoju weir (YJW)**
  - Total storage: 13 million m³

- **Kangcheon weir (KCW)**
  - Total storage: 11 million m³
Data for SWAT model evaluation

4 Multipurpose dam data (release and storage: 1984-2014)

- **Soyang dam (SYD)**
  - Data for SWAT model evaluation
  - Precipitation (mm)
  - Total Release
  - Storage
  - Volume of flood water level (2,900 $10^3$ m$^3$)
  - Volume of full water level (2,504 $10^3$ m$^3$)

- **Hoengseong (HSD)**
  - Data for SWAT model evaluation
  - Precipitation (mm)
  - Total Release
  - Storage
  - Volume of flood water level (87 $10^3$ m$^3$)
  - Volume of full water level (79 $10^3$ m$^3$)

- **Chungju dam (CJD)**
  - Data for SWAT model evaluation
  - Precipitation (mm)
  - Total Release
  - Storage
  - Volume of flood water level (2,750 $10^3$ m$^3$)
  - Volume of full water level (2,252 $10^3$ m$^3$)

- **Paldang dam (PDD)**
  - Data for SWAT model evaluation
  - Precipitation (mm)
  - Total Release
  - Storage
  - Volume of flood water level (244 $10^3$ m$^3$)
  - Volume of full water level (226 $10^3$ m$^3$)
Data for SWAT model evaluation

3 Multifunction weir data (release and storage: 2012-2014)

Ipo weir (IPW)

Yeoju weir (YJW)

Kangcheon weir (KCW)
SWAT Model calibration and validation

- Hydrology Results

Observed vs. simulated streamflow results of model calibration and validation

- Calibration: 5 years (2005-2009) / Validation: 5 years (2010-2014)

Ahn et al. (2015)

Calibration period
Validation period

SYD

HSD

CJD

PDD

R²: 0.90
NSE: 0.78
PBIAS(%): 12

R²: 0.83
NSE: 0.59
PBIAS(%): 14

R²: 0.78
NSE: 0.61
PBIAS(%): 9

R²: 0.90
NSE: 0.80
PBIAS(%): 5
SWAT Model calibration and validation

- Hydrology Results

Ahn et al. (2015)

Observed vs. simulated streamflow results of model calibration and validation

- Calibration: 2 years (2012-2013) / Validation: 1 year (2014)

\[ \text{R}^2: 0.76 \quad \text{NSE: 0.88} \quad \text{PBIAS(\%): 21} \]

\[ \text{R}^2: 0.77 \quad \text{NSE: 0.77} \quad \text{PBIAS(\%): 20} \]

\[ \text{R}^2: 0.77 \quad \text{NSE: 0.79} \quad \text{PBIAS(\%): 11} \]
Ahn et al. (2015)

SWAT Model calibration and validation

- Hydrology Results

Observed vs. simulated ET & SM results of model calibration and validation

Calibration: 3 years (2009-2011) / Validation: 2 years (2012-2013)

Ahn et al. (2015)
Ahn et al. (2015) observed vs. simulated sediment results of SWAT model calibration and validation.

Calibration: 5 years (2005-2009) / Validation: 5 years (2010-2014) at 7 stations.
SWAT Model calibration and validation

- Waterquality

Observed vs. simulated T-N results of SWAT model calibration and validation

- Calibration: 5 years (2005-2009) / Validation: 5 years (2010-2014) at 7 stations

Ahn et al. (2015)
SWAT Model calibration and validation

- Waterquality

Ahn et al. (2015)

Observed vs. simulated T-P results of SWAT model calibration and validation

✅ Calibration: 5 years (2005-2009) / Validation: 5 years (2010-2014) at 7 stations

R²: 0.77
R²: 0.90
R²: 0.68
R²: 0.57
R²: 0.58
R²: 0.58
R²: 0.71
South Korea has the **National Aquatic Ecological Monitoring Program (NAEMP)** operated by the Ministry of Environment and the National Institute of Environmental Research, Korea.

- since 2007 for entire country
- spring (April to May) and autumn (September to October) in twice a year
- Measurement components: water temperature, pH, DO, BOD, NH$_4$, NO$_3$, T-N, T-P, PO$_4$, Chlorophyll-a, and so on.
- from that components, TDI (Trophic Diatom Index), FAI (Fish Assessment Index), and BMI (Benthic Macroinvertebrate Index) have been estimated

Locations of the sampling sites in five major watersheds in South Korea

Int. J. Environ. Res. Public Health 2012, 9, 3629-3653
Aquatic Ecology Health Index

- **Fish Assessment Index, FAI (U. S. EPA, 1993)**
  - Calculate the score for 4 metrics \(M1, M2, M3\) and \(M7\) that depends on the stream order of Korea and the other 4 metrics \(M4, M5, M6\) and \(M8\), using class division “0”, “6.25”, “12.5”

\[
\text{FAI} = M1 + M2 + M3 + M4 + M5 + M6 + M7 + M8
\]

- **Trophic Diatom Index, TDI (Kelly and Withon, 1995)**
  - Calculate the score using relative density, contamination sensitivity and indicative value of emerging species for about 400 species of Trophic Diatom

\[
\text{TDI} = 100 - (25\left(\sum a_i s_i v_i / \sum a_i v_i\right) - 25)
\]

- **Benthic Macroinvertebrate Index, BMI (Merritt and Cummins 1984)**
  - Calculate the score using benthic macroinvertebrate organisms with contaminate and indicative weighted value

\[
\text{BMI} = (4-\sum s_i h_i g_i / \sum h_i g_i) * 25
\]

<table>
<thead>
<tr>
<th>Index</th>
<th>A (Very Good)</th>
<th>B (Good)</th>
<th>C (Fair)</th>
<th>D (Poor)</th>
<th>E (Very Poor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAI</td>
<td>100 ≥ ~ ≥ 80</td>
<td>80 &gt; ~ ≥ 60</td>
<td>60 &gt; ~ ≥ 40</td>
<td>40 &gt; ~ ≥ 20</td>
<td>20 &gt; ~ ≥ 0</td>
</tr>
<tr>
<td>TDI</td>
<td>100 ≥ ~ ≥ 90</td>
<td>90 &gt; ~ ≥ 70</td>
<td>70 &gt; ~ ≥ 50</td>
<td>50 &gt; ~ ≥ 30</td>
<td>30 &gt; ~ ≥ 0</td>
</tr>
<tr>
<td>BMI</td>
<td>100 ≥ ~ ≥ 80</td>
<td>80 &gt; ~ ≥ 65</td>
<td>65 &gt; ~ ≥ 50</td>
<td>50 &gt; ~ ≥ 35</td>
<td>35 &gt; ~ ≥ 0</td>
</tr>
</tbody>
</table>
Fish Assessment Index, FAI:

- It refers to the organism at the top level of the food chain in the water body that represents omnivorous, herbivorous, inflorescence, and carnivorous at various nutritional stages.
Trophic Diatom Index, TDI:

- As the primary producer of the river ecosystem food chain, it refers to the diatom attached to the stone (substrate) such as gravel or cobble stone in the bottom.

- It is responsible for energy transfer in the ecosystem.

- Also, it is sensitive to changes in water quality (TN, TP) and environment.
Benthic Macroinvertebrate Index, BMI:

- It is a biologic indicator that reflects local characteristics as a primary or secondary consumer of river ecosystem.

- It refers to a group of aquatic insects.

- As a sub-consumer linking producers and upper consumers in aquatic ecosystem food chain, they are sensitive to environmental changes and have excellent indicators and are used as indicators of water quality evaluation.
Correlation Analysis

- The relationship between the score and the water quality is not clear
Machine Learning

General Machine Learning

- Supervised Learning
  - Regression
- Classification
  - Clustering
- Unsupervised Learning
  - Density estimation

Classification
- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-nearest Neighbors Classifier (KNN)
- Classification And Regression Tree, Decision Tree (CART)
- Gaussian Naive Bayes (NB)
- Random Forest Classifier (RF)

Ensemble Machine Learning

Bootstrap sampling (복원추출 샘플링)

Step 1: Create Multiple Data Sets

Step 2: Build Multiple Classifiers

Step 3: Combine Classifiers
**Machine Learning**

**eXtreme Gradient Boosting tree (XGBoost)**

- Algorithm to make decision by mixing several models (tree boosting model)

- The algorithm that learns results sequentially, with previous result and previous result affect the next model result in the current stage (additive training)

- This is a system in which the predicted value ($Y$) gradually approaches the target value ($\hat{Y}$) as the stage progresses for next stage

- So, this learn weakly learning, and gradually get closer to actual value unlike random forest model

- The method that weakly fit current data has a high bias but low variance.

- The high bias can be improved sufficiently by sequentially learning data weakly.
Machine Learning

Input variables

Flow_Spring

PCP_Spring

T-N

NH4

T-N (mg/L)_Spring

NH4 (mg/L) _Spring
Machine Learning

Input variables

<table>
<thead>
<tr>
<th>NO3(mg/L)_SPRING</th>
<th>T-P(mg/L)_SPRING</th>
<th>PO4(mg/L)_SPRING</th>
<th>WT_SPRING</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>2009</td>
<td>2009</td>
<td>2009</td>
</tr>
<tr>
<td>2010</td>
<td>2010</td>
<td>2010</td>
<td>2010</td>
</tr>
<tr>
<td>2011</td>
<td>2011</td>
<td>2011</td>
<td>2011</td>
</tr>
<tr>
<td>2012</td>
<td>2012</td>
<td>2012</td>
<td>2012</td>
</tr>
<tr>
<td>2013</td>
<td>2013</td>
<td>2013</td>
<td>2013</td>
</tr>
<tr>
<td>2015</td>
<td>2015</td>
<td>2015</td>
<td>2015</td>
</tr>
</tbody>
</table>

(Water temperature, °C)

- 16.30491447 - 16.5
- 16.50000001 - 17
- 17.00000001 - 17.5
- 17.50000001 - 18
- 18.00000001 - 19.5
**Training technique (Random forest)**

**K-fold Cross validation**

- Generally, the data divided into **training (70%)** and test (30%).
- To reduce variability, cross-validation are performed using different partitions and results of the different partitions are combined (e.g. averaged) to estimate a final predictive model.
- The k-fold cross validation separate training data into k folds without overlap.
- One set of the k folds is separated by training and validation.

- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-nearest Neighbors Classifier (KNN)
- Classification And Regression Tree, Decision Tree (CART)
- Gaussian Naive Bayes (NB)
- Random Forest Classifier (RF)

**Average accuracy**

\[
\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)
\]

**Average accuracy of verification**:
- FAI: 0.43
- TDI: 0.45
- BMI: 0.58
Training technique (XGBoost)

Training & Verification

- The k-fold cross validation & Grid search
- Feature importance (parameter tuning)
  - Max-depth: Maximum tree depth for base learners
  - Gamma: Minimum loss reduction required to make a further partition on a leaf node of the tree
- For optimization of parameters, max_depth applied from 1 to 10 with 1 intervals and Gamma applied from 0 to 4.5 with 0.5 intervals

Average accuracy of verification:

- FAI: 0.72
- TDI: 0.66
- BMI: 0.80
Watershed Health (FAI_spring)

2008: Obs. (first), Predicted (second)
2009: Obs. (first), Predicted (second)
2010: Obs. (first), Predicted (second)
2011: Obs. (first), Predicted (second)
2012: Obs. (first), Predicted (second)
2013: Obs. (first), Predicted (second)
2014: Obs. (first), Predicted (second)
2015: Obs. (first), Predicted (second)

- 36 watersheds that have failed to predict.
- TP and NO$_3$ in failed watersheds were relatively greater than the overall average TP and NO$_3$ values.
- TP was 19.3% greater than whole.
- PO$_4$ was 14.4% greater than whole.
Watershed Health (TDI_spring)

- 2008: Obs. (first), Predicted (second)
- 2009: Obs. (first), Predicted (second)
- 2010: Obs. (first), Predicted (second)
- 2011: Obs. (first), Predicted (second)
- 2012: Obs. (first), Predicted (second)
- 2013: Obs. (first), Predicted (second)
- 2014: Obs. (first), Predicted (second)
- 2015: Obs. (first), Predicted (second)

- 41 watersheds that have failed to predict.
- TP and PO4 in failed watersheds were relatively smaller than the whole average TP and PO4 values.
- TP was -30.1% smaller than whole.
- PO4 was -29.6% smaller than whole.
25 watersheds that have failed to predict.

- TP and PO₄ in failed watersheds were relatively greater than the overall flow and PO₄ values.
- TP was 66.1% greater than whole.
- PO₄ was 81.8% greater than whole.
19 watersheds that have failed to predict.

- NH$_4$ and NO$_3$ in failed watersheds were relatively greater than the overall values.
- NH$_4$ was 25.9% greater than whole.
- NO$_3$ was 28.0% greater than whole.
- Flow was -66.3% smaller than whole.
45 watersheds that have failed to predict.

TP and NH$_4$ in failed watersheds were relatively smaller than the overall values.

TP was -23.4% smaller than whole.

NH$_4$ was -31.6% smaller than whole.

Flow was -38.6% smaller than whole.
Watershed Health (BMI_autumn)

2008: Obs. (first), Predicted (second)
2009: Obs. (first), Predicted (second)
2010: Obs. (first), Predicted (second)
2011: Obs. (first), Predicted (second)
2012: Obs. (first), Predicted (second)
2013: Obs. (first), Predicted (second)
2014: Obs. (first), Predicted (second)
2015: Obs. (first), Predicted (second)

✓ 23 watersheds that have failed to predict.
✓ NO₃ and NH₄ in failed watersheds were relatively greater and smaller than the overall values.
✓ NO₃ was 50.6% greater than whole.
✓ NH₄ was -29.8% smaller than whole.
✓ Flow was -35.3% smaller than whole.
Findings and Future Researches

- This study was to develop XGBoost which is one of ensemble machine learning algorithms (Random forest vs. XGBoost) for AEH indices prediction using SWAT results.

- We could predict AEH indices of ungauged streams via XGBoost with SWAT water quality results at a standard watershed scale.

- For further research, we will include environmental components such as river width, bankside cover treatment with porous material or vegetation cover etc..