Evaluation of algal species distribution and prediction of Cyanophyte cell count using a statistical technique

Seong-Yun Hwang ¹⁺․ Byung-Woong Choi ¹․ Jong-Hwan Park ¹․ Dong-Seok Shin ¹․ Hyeon-Su Chung ¹․ Mi-Sun Son ¹․ Kyung-Lak Lee ³․ Kang-Young Jung ²[[1]](#footnote-1)

*Yeongsan River Environment Research Center, National Institute of Environmental Research* ¹

*Education Planning Division, National Institute of Environmental Human Resources Development* ²

*Water Environmental Engineering Research Division, National Institute of Environmental Research* ³

**ABSTRACT**

Water supply sources are crucial for access to drinking water for local residents. Consequently, the importance of water quality management is widely recognized, including continuous monitoring of water quality at the sources (e.g., monitoring of algae at the sites and measurement of water quality attributes) to ensure safe water for local residents. In the present study, the Juam Lake and Tamjin Lake, which are representative water supply sources in the Yeongsan and Seomjin Rivers, Republic of Korea, were investigated. In addition, the Water Environment Information System data of the National Institute of Environmental Research Water: water quality measurement network data, algae alert system data, and hydraulics and hydrological data measured every seven days from January 2017 to December 2022, were used. Furthermore, an explanatory data analysis and correlation analysis of the measured data from 2017 to 2021 were conducted to examine distributional features of and correlations among measured variables, and factors influencing the distribution of algal species found during the survey via a Redundancy Analysis. The 2017–2021 data were set as training data, and the 2022 data were set as test data. The performances of the corresponding models were compared, and the models were quantitatively evaluated , to identify the prediction performance of seven models for predicting Cyanophyte cell count. The environmental characteristics of algae appearing in water supply sources were observed based on monitoring data, and a prediction model appropriate for the distributional features of the cell count of Cyanophytes, with toxic damage risk. The results of the present study could facilitate water quality management in various water systems, including water supply sources.

**Keywords**: Water quality, Yeongsan River Basin, Seomjin River Basin, Redundancy Analysis, Generalized Linear Model, Zero-Inflated Model, Generalized Additive Model

**1. Introduction**

Water supply sources provide local residents with domestic/drinking water. Consequently, all countries strive to manage water supply source quality within their territories, and Korea is not an exception. Korea established its algae alert system in 1998, and has attempted to prevent toxic damage from harmful blue-green algae; there are several phases of the prevention activity, which is based on the cell count of harmful blue-green algae, including warning (> 1,000 cells/mL, twice in a row), alert (> 10,000 cells/mL, twice in a row), algae outbreak (> 1,000,000 cells/mL, twice in a row), and release (lower than the warning criteria for a cell count of harmful blue-green algae, twice in a row) (Kim, 2017)). The severe drought in the Honam region of Korea in 2022 highlighted the importance of the quality management of water supply sources, since four types of harmful blue-green algae (i.e., *Aphanizomenon, Anabaena, Oscillatoria, and Microcystis*), which are frequently found in summer, release harmful toxins that cause serious diseases such as acute liver disease (Falconer et al., 2005), and threaten aquatic ecosystem stability (Fleming et al., 2002), were observed.

The National Institute of Environmental Research, Korea, collected samples at several points at water supply sources every week to monitor water source quality, investigating diverse water quality characteristics, including the algae present, and has compiled the data collected into a database. Lee et al. (2021) collected seasonally measured data at eight weir sections of the Nakdonggang River in Korea (i.e., Sangju, Nakdan, Gumi, Chilgok, Gangjeong-Goryeong, Dalseong, Hapcheon-Changnyeong, and Changnyeong-Haman) in 2014. Based on the data, they conducted a Redundancy Analysis (RDA) for the major diatom species, used a generalized linear model (GLM) to fit trophic diatom index (TDI) and diatom index for organic pollution (DAIpo) to the model, and investigated which measured variables influenced the distribution of the major diatom species significantly. However, they could not examine the distribution features of various types of algae, such as blue-green algae and green algae, and appropriate distribution characteristics of TDI and DAIpo were not considered even when applying GLM. Seo et al. (2019), like Lee et al. (2021), targeted the same sections, collected monthly data from 2012 to 2016, and used multiple regression analysis to fit a model for predicting the numbers of total coliform and fecal coliform; however, they did not consider the distributional features of the number of *Escherichia coli*, which is a count variable.

In the present study, the aquatic ecosystem study sites were Juam Lake and Tamjin Lake, which are representative water supply sources in the Yeongsan·Seomjin River, Republic of Korea, and data were measured, collected, and monitored every seven days from January 2018 to December 2022. An RDA was conducted for algae found in the corresponding survey periods, and the measured variables influencing algae species emergence were examined. When fitting the prediction model for the cell count of blue-green algae that occurred next, we applied an estimation model based on Poisson distribution and negative binomial distribution, which are appropriate probability distributions for the count variable. Subsequently, we explain how to use a model fitting in consideration of distributional features of response variables, and suggest how to predict the cell count of blue-green algae by combining Random Forest (RF) and extreme gradient boosting statistical machine learning techniques. Furthermore, we suggest strategies for addressing the limitations of the two previous studies.

**2. Materials and Methods**

**2.1 Study Area**

This study investigated the representative water supply sources in Yeongsan and Seomjin River basins, Juam Lake and Tamjin Lake. Yeongsan River Environment Research Center of the National Institute of Environmental Research has attempted to respond to the algae alert system, to monitor the quality of water supply sources providing local residents with domestic water, and to collect and investigate samples obtained from weekly field trips at the Juam Lake dam front point (J1), the Sinpyeong-Gyo (J2) point, the Tamjin Lake’s dam front point of (T1), and the merge section of Yuchicheon (T2). Figure 1 shows the points where samples were collected in detail. Juam Lake is the largest freshwater lake in Jeollanam-do and is an artificial lake that was established during the construction of the Juam Dam. It was completed after eight years of construction from 1984 to 1992 based on the government's long-term multi-purpose dam development plan. The total water storage is 457 × 10⁶ ton, the basin area is 1,029.41 km², and the reservoir area is 33 km².

|  |  |
| --- | --- |
|  | |
|  |  |
| (a) Juam lake (J1, J2) | (b) Tamjin lake (T1, T2) |

Figure 1. Sampling sites in the Juam lake and Tamjin lake

The total water storage is 457 × 10⁶ ton, and it supplies approximately 640×10³ ton of domestic water every day to residents in the western part of Jeollanam-do, including Gwangju, Naju, and Mokpo (Shim et al., 2014). Tamjin Lake is an artificial lake that was created during the construction of the Jangheung Dam; the total basin area is 193 km², and the total water storage is 191 × 10⁶ ton. Consequently, annually, 73 × 10⁶ ton of domestic water of is supplied to residents in nine cities and counties in Jeollanam-do (Lim et al., 2018).

**2.2. Data Collection**

For data collection, we utilized the data of the Environment Information System of the National Institute of Environmental Research Water: water quality measurement network data, algae alert system data, and hydraulics and hydrological data, which were collected every seven days from January to December 2022 at corresponding target points. After comprehensively compiling such data, the data were combined with data from four target points, and refined into analyzable formats. The variables, which are included in the resultant data, are summarized in Table 1.

Table 1. Data variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Response variable**  **(counting variable)** | | **Explanatory variables**  **(continuous)** | |
| **Water quality** | **Hydraulics, hydrological** |
| **For Redundancy Analysis** | **Modeling of cell count** | BOD (mg/L)  COD (mg/L)  TN (mg/L)  TP (mg/L)  TOC (mg/L)  SS (mg/L)  EC (μS/cm)  pH  DO (mg/L)  Temperature (℃)  Turbidity (NTU)  Transparency (m)  Chla (mg/m3) | Low Water Level (cm)  Inflow (cms)  Discharge (cms)  Reservoir (10,000 m³) |
| **Cell counts of all algal species that emerged at the sampling site** | **Cyanophyte**  **cell count Cyanophytes** |

In Table 1, various data, including Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Nitrogen (TN), Total Phosphorus (TP), Total Organic Carbon (TOC), Suspended Solids (SS), and Electrical Conductivity (EC), from water quality measurement network data, were obtained. In addition, data such pH, Dissolved Oxygen (DO), Temperature, Turbidity, Transparency, Chlorophyll-a (Chla), and Cell count from were collected algae alert system data. The remaining variables, low water level (LowWaterLevel), inflow, discharge, and reservoir, were collected from the Water Resources Management Information System website (http://www.wamis.go.kr/). Furthermore, the species of algae found at the survey sites during the data measurement are summarized in Table 2. Figure 2 is a flowchart indicating the methodology of the present research. CANOCO v5.0 (Microcomputer Power, Ithaca, NY, USA), R program v4.2.1 (R Foundation for Statistical Computing, Vienna, Austria), and SigmaPlot v12.5 (Systat Software Inc., Palo Alto, CA, USA) were used for data analysis.

Table 2. Algal species that were observed at the Juam lake and Tamjin lake

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cyanophytes** | | **Diatom species** | **Chlorophytes** | **Others** |
| **Normal** | **Harmful** |
| *Aphanocapsa*  *Chroococcus*  *Merismopedia*  *Phormidium*  *Pseudanabaena*  *Worinochinia* | *Anabaena*  *Aphanizomenon*  *Microcystis*  *Oscillatoria* | *Acanthoceras*  *Achnanthes*  *Asterionella*  *Attheya*  *Aulacoseira*  *Coccoineis*  *Cyclotella*  *Cymbella*  *Fragilaria*  *Gomphonema*  *Melosira*  *Navicula*  *Nitzschia*  *Rhizosolenia*  *Stephanodiscus*  *Surirella*  *Synedra* | *Actinastrum*  *Ankistrodesmus*  *Ankyra*  *Chlamydomonas*  *Chlorella*  *Chodatella*  *Closteriopsis*  *Closterium*  *Coelastrum*  *Coenochloris*  *Cosmarium*  *Crucigenia*  *Dictyosphaerium*  *Dimorphococcus*  *Elakatothrix*  *Euastrum*  *Eudorina*  *Eunotia*  *Gloeocystis*  *Golenkinia*  *Gonium*  *Kirchnerionella*  *Micractinium*  *Monoraphidium*  *Mougeotia*  *Nephrocystium*  *Oocystis*  *Pandorina*  *Pectodictyon*  *Pediastrum*  *Scenedesmus*  *Schroederia*  *Selenastrum*  *Sphaerocystis*  *Spondylosium*  *Staurastrum*  *Tetraedron*  *Tetrastrum*  *Treubaria* | *Ceratium*  *Cryptomonas*  *Dinobryon*  *Euglena*  *Kephyrion*  *Mallomonas*  *Peridinium*  *Phacus*  *Strombomonas*  *Trachelomonas* |

|  |
| --- |
|  |

Figure 2. Methodology flowchart

**2.3. Explanatory Data Analysis**

Explanatory data analysis is a process of performing a broad analysis of the variables included in the data before analyzing the given data, and of determining the general framework of the data to be analyzed (Martinez et al., 2017). The analysis does not have a set procedure but most researchers perform descriptive statistics for each variable (e.g., mean, standard deviation, maximum, and minimum), construct box plots for examining distributions, or density plots. There are potential differences in preferred analysis processes across researchers. In the present study, we conducted a correlation analysis to examine correlations between water quality items, and hydraulics and hydrological variables, which were included in analysis data. In addition, an RDA was conducted to examine relationships between algal species and algal habitats observed at survey sites, and environmental variables.

**2.3.1. Correlation Analysis**

Correlation analysis is a method to analyze the relationship between continuous variables included in the data to be analyzed. In general, Pearson correlation coefficients are calculated, as shown in Equation (1), and corresponding significance tests are performed.

(1)

The method has a constraint whereby normality should be satisfied through a normality test (e.g., Shapiro-Wilk (S-W) test or Kolmogorov-Smirnov (K-S) test) for corresponding variables (Das et al., 2016). However, the two normality test methods have limitations, as they can be applied only when corresponding variables satisfy properties of random variables satisfying independency. The measured variables in the present study are time series variables that are measured over time, so that they do not satisfy independency. Therefore, the Jarque-Bera (J-B) test method is appropriate for such variables; the normality of this method is tested, by comparing whether skewness, which represents the level of asymmetry, and kurtosis, which represents the thickness of the tail of the distribution, are similar to theoretical values of normal distribution (Thadewald & Büning, 2007). In particular, as water quality items, or hydraulics and hydrological variables, which are included in the data of this study, mostly do not satisfy the normality, analyses based on the Pearson correlation coefficients for these variables have lower validity (Cohen et al., 2014). Therefore, the Spearman correlation coefficient, which is a correlation coefficient defined by rank, and which is irrelevant to distribution, as shown in Equation (2), was used in the present study to conduct an analysis.

(2)

**2.3.2. Redundancy Analysis**

Redundancy Analysis is a frequently used method for analyzing ecological data; through the analysis, researchers can use diverse environmental variables, establish a multivariate regression model for a particular species, conduct a principal component analysis on fitted variables with a linear combination of explanatory variables included in the data, and identify relationships among species, habitats, and environmental variables (Capblancq et al., 2021). Such a process enables researchers to identify which factors have greater influence on the habitats of corresponding organisms, and to confirm the distribution of their habitats. Before performing an RDA, variables with a variance inflation factor (VIF) of 20 or lower were selected to eliminate multicollinearity among 17 explanatory variables, which are shown in Table 1 (ter Braak et al., 2018), and 11 measured variables with significant impacts on changes in the phytoplankton communities were eventually confirmed.

Table 3. Selected algal species and code for redundancy analysis

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Species** | | | | | | | | |
| **Code** | **Cyanophytes** | | **Code** | **Diatom species** | **Code** | **Chlorophytes** | **Code** | **Others** |
| **normal** | **harmful** |
| **Chro** | *Chroococcus* |  | **Aste** | *Asterionella* | **Coel** | *Coelastrum* | **Cryp** | *Cryptomonas* |
| **Anab** |  | *Anabaena* | **Aula** | *Aulacoseira* | **Cruc** | *Crucigenia* | **Dino** | *Dinobryon* |
| **Apha** |  | *Aphanizomenon* | **Cycl** | *Cyclotella* | **Mono** | *Monoraphidium* | **Mall** | *Mallomonas* |
| **Micr** |  | *Microcystis* | **Frag** | *Fragilaria* | **Pedi** | *Pediastrum* | **Peri** | *Peridinium* |
| **Osci** |  | *Oscillatoria* | **Syne** | *Synedra* | **Scen** | *Scenedesmus* | **Trac** | *Trachelomonas* |

In the present study, based the data measured from 2017 to 2020, five algal species, which were frequently dominant, were selected, and are shown in Table 2. An RDA was conducted on a total of 20 species. The selected species and Code can be found in Table 3.

**2.4. Modeling for cell count of Cyanophytes**

The model in the present study aims to predict the cell count of Cyanophytes. The cell count of Cyanophytes is an integer greater than 0 and does not have normality because it is a count variable. Therefore, if the generally used multiple regression model based on the normal distribution assumption is utilized, the validity of the analysis would inevitably decrease (Ajiferuke et al., 2015). The probability distribution, which can reflect distributional features of count variables, can be generally divided into Poisson distribution and negative binomial distribution. In that sense, a model for predicting count variables should be based on the two types of distributions. To reduce multicollinearity problems, prior to a model fitting, the standardization process of subtracting the mean of all measured variables used for explanatory variables and dividing by the standard deviation, was performed, as shown in Equation (3).

(3)

Furthermore, when selecting measured variables with great impacts on prediction of blue-green algae cell count, the AIC (Akaike Information Criterion)-based stepwise regression method was used. Smaller AIC values indicate a more accurate prediction model (Cavanaugh et al., 2019).

**2.4.1. Generalized Linear Model**

A GLM is a generalized concept of a multiple regression model, and is the simplest linear model enabling the establishment of a prediction model for response variables based on diverse probability distributions in addition to normal distributions (Dobson et al., 2018). A multiple regression model is also included. In this part, we explained models to which count variables, such as the cell count of Cyanophytes, can be applied.

**2.4.1.1. Poisson Regression Model**

The Poisson regression model (PRM) is a Poisson distribution-based model, as shown in Equation (4), and the distribution is discrete probability distribution expressing how many times events of interest occur within a unit of time (Casella et al., 2021).

(4)

This distribution has a feature in that mean and variance are the same. In other words, . The PRM based on the distribution has the representative form, as shown in Equation (5), with natural logarithms ln as a link. In other words, it can be interpreted as the model for predicting the log value of the mean of the Poisson distribution. However, the characteristic of the Poisson distribution in which the mean and variance are the same is a disadvantage in that there is a limitation in applying it in diverse situations.

(5)

Unlike the multiple regression model utilizing the least square estimation, the estimator of the regression coefficient for this model is repeatedly calculated, until convergence of the Newton-Raphson algorithm, to find the estimator maximizing the Likelihood Function.

**2.4.1.2. Negative Binomial Regression Model**

The negative binomial regression model (NBRM) is a negative binomial distribution-based model, as shown in Equation (6); the distribution is discrete probability distribution expressing how many trials are required to achieve the r-th success (Mukhopadhyay, 2020).

(6)

The mean for the distribution is , and the variance is . Similarly, the NBRM based on the distribution, like the PRM, has the form with natural logarithms ln as a link, as shown in Equation (7). Therefore, the model can be interpreted as a model for predicting log values for the mean of the negative binomial distribution. Furthermore, the estimator of the regression coefficient is identical to the estimation method in the PRM. Only the probability distribution is different. Due to the characteristics of the negative binomial distribution, which has no constraint of similar mean and variance, the NBRM has the advantage of being applicable to a broader range of situations compared to the PRM.

(7)

**2.4.2. Zero-Inflated Model**

The cell count of Cyanophytes to be modeled in the present study tends to increase in early summer and during summer (Seo et al., 2013). Consequently, the period over which the species is observed is limited, leading to numerous points at which the value of cell count is 0. The case where the value of measured variables is extremely high is considered a zero-inflated phenomenon, and the applicable model, in this case, is the zero-inflated model (ZIM) (Zuur et al., 2009). In general, in the ZIM, the following models are simultaneously utilized, PRM or NBRM, which mainly use a log link for modeling count variables, and logistic regression classification model, which uses a logit link for modeling the probability that becomes a classification criterion whether an observed value is 0 or not.

**2.4.2.1. Zero-Inflated Poisson Model**

The zero-inflated Poisson model (ZIPM) can be used in a case where non-zero observed values follow the Poisson distribution, and there is an excess of observed values that are 0 (Loeys et al., 2012), and the model is based on the same distribution as Equation (8).

(8)

Here, is the mean of the Poisson distribution, and is the probability that a random variable has a value of 0 (). The mean of this distribution is ; the variance is . Consequently, and are modeled respectively, in the ZIPM, and the form can be found in Equation (9).

(9)

**2.4.2.2. Zero-Inflated Negative Binomial Model**

The zero-inflated negative binomial model (ZINBM) can be used in a case where non-zero observed values follow the negative binomial distribution, and there are excessive observed values that are 0 (Yusuf et al., 2017), and the model is based on the same distribution as Equation (10).

(10)

Here, is the mean of the negative binomial distribution, and is the over-dispersion parameter; there is the Gamma function in which , the real number greater than 0, is defined as . The mean of this distribution is ; the variance is . Consequently, and are modeled respectively, in the ZINBM, and the form can be found in Equation (11).

(11)

**2.4.3. Generalized Additive Model**

The GLM assumes that there is a linear relationship between the log value of the cell count of blue-green algae, and the measured variables. However, sometimes, there are cases where the relationship is non-linear, and in reality, there are more such cases. The generalized additive model (GAM) can compensate for such limitations, allows non-linear functions for each explanatory variable, and provides an extended form of a GLM (Hastie, 2017). The specific form can be expressed as follows: only the part indicating the linear relationship in the GLM is converted to an additive form of a non-linear function. Therefore, the forms of the generalized additive Poisson model (GAPM) and the generalized additive negative binomial model (GANBM) can be found in Equation (12) and Equation (13), respectively.

(12)

(13)

When assuming a non-linear function showing the relationship between response variables and explanatory variables in the GAM, non-parametric function methods, such as smoothing splines or local regression, are mainly utilized.

**2.5. Prediction of cell count of Cyanophytes using Statistical Machine Learning Technique**

In this section, we explain how to combine and apply two types of statistical machine learning techniques in prediction of the cell count of Cyanophytes. Random forest, which exhibits excellent performance in classification problems, and extreme gradient boosting, which is appropriate for predicting time series variables. In other words, we attempted to first classify whether the cell count is 0 or not, and to predict cell count at the point when the cell count is classified as non-zero, in order to suggest a more accurate method of predicting cell count.

**2.5.1. Random Forest**

Random forest is a method in which a number of samples () for analysis are extracted via sampling with replacement methods, based on that, a great number of decision tree models () can be created, and prediction results obtained through the process are averaged (), or based on classification results, bagging is performed to derive final outcomes through multi-voting (). It is also a method of compensating for the disadvantages of bagging, in which there may be problems due to correlations between samples, by randomly and differently selecting all types of explanatory variables selected for each extracted sample. Through the process, prediction or classification performances can be improved greatly by reducing the variance of the model and the connections among samples (Rigatti, 2017). Figure 3 simply illustrates the principle of bagging, which is the basis of RF.

|  |
| --- |
|  |

Figure 3. Bagging

**2.5.2. Extreme Gradient Boosting**

Extreme Gradient Boosting (XGBoost) is a method of creating a model by using gradients, and repeating a process of remodeling with resultant residuals. By improvement through additional support of parallel learning, the slow performance time and overfitting risks are compensated (Zhang et al., 2021). If given training data are , and a previously created model is , in general gradient boosting, there will be a process of finding the function in which a residual, a difference between actual values and predicted values, is modeled, as shown in Equation (14).

(14)

After clarifying the function through this process, the updated new model through this process can be found using Equation (15).

(15)

Here, parameter serves the role of reducing overfitting risks as a learning rate. XGBoost can perform a cross-validation test by itself, and has an early stopping function that detects in advance when overfitting would emerge Here, the cross-validation test is a method that evaluates the performance of an algorithm for the -divided data, by arbitrarily dividing the data into parts, using the data as new train data, and setting the remaining 1 data as new test data. The principles of gradient boosting and cross-validation test are illustrated in Figure 4. and Figure 5., respectively.

|  |
| --- |
|  |

Figure 4. Gradient boosting

|  |
| --- |
|  |

Figure 5. Cross-validation test

**2.5.3. Extreme Gradient Boosting with Random Forest (XGBRF)**

In the present study, we attempted to predict the cell count of Cyanophytes by simultaneously utilizing RF, which is frequently used for classification problems, and XGBoost, which exhibits good performance in problems of predicting time series variables. In other words, based on RF, it was possible to classify whether cell count is 0 or not, and based on XGBoost, it was possible to predict cell count at surveyed points, demonstrating that the cell count was not 0. In case the cell count was classified as 0, we set the prediction value as 0. The cell count of Cyanophytes is a zero-inflated variable, which has numerous cases of having zero; therefore, before predicting the cell count, pre-classification of cases in which the cell count is highly likely to appear as 0, can facilitate a more accurate prediction.

**3. Results and discussion**

**3.1. Explanatory Data Analysis**

Table 4 presents the descriptive statistics of the measured variables in the data, which were measured from 2017 to 2021. Furthermore, we calculated the p-value for the J-B test to identify the normality of the measured variables, which are time series variables. Figure 6 presents a boxplot with a density plot to visually confirm the distribution of each measured variable. Based on the results, it is possible to confirm the overall distributional features of the measured variables (Kaur et al., 2018). The descriptive statistics encompass the mean, standard deviation, median, minimum, maximum, skewness, and kurtosis of data. Skewness has a positive value with the long tail to the right, and a negative value with the long tail to the left. As for kurtosis, if the value is > 0, the central part of the distribution appears sharp; if it is < 0, the central part appears smooth (Đorić et al., 2009). When it comes to the descriptive statistics as a whole, skewness had positive values for all variables, except for Temperature and LowWaterLevel; kurtosis had positive values for all variables, except for DO, temperature, LowWaterLevel, and reservoir. In other words, most measured variables had long tails to the right, with a sharp distribution in the center. The p-value for the J-B test was much lower than 0.05, which is the general significance level for all measured variables. In other words, since all measured variables violated the normality assumption, a correlation analysis needed to be performed using spearman correlation coefficients rather than Pearson correlation coefficients (Rebekić et al. (2015)).

Table 4. Descriptive statistics

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **statistics** | **BOD**  **(mg/L)** | **COD**  **(mg/L)** | **TN**  **(mg/L)** | **TP**  **(mg/L)** | **TOC**  **(mg/L)** | **SS**  **(mg/L)** | **EC**  **(μS/cm)** | **pH** | **DO**  **(mg/L)** | **Tempe-rature**  **(℃)** | **Turbid-ity**  **(NTU)** | **Trans-parency**  **(m)** | **Chla**  **(mg/㎥)** | **Low Water Level**  **(cm)** | **Inflow**  **(cms)** | **Discha-rge**  **(cms)** | **Reserv-oir**  **(10,000㎥)** |
| **mean** | 1.41 | 4.18 | 1.17 | 0.05 | 3.04 | 5.75 | 127.66 | 7.3 | 8.87 | 14.06 | 2.51 | 2.92 | 4.33 | 8625.58 | 10.68 | 11.36 | 17315.25 |
| **sd** | 0.87 | 1.33 | 0.68 | 0.05 | 1.03 | 8.11 | 61.48 | 0.5 | 2.19 | 5.62 | 2.95 | 0.93 | 2.85 | 1512.88 | 31.76 | 29.94 | 9513.46 |
| **median** | 1.2 | 3.8 | 0.89 | 0.03 | 2.8 | 3 | 98 | 7.3 | 8.7 | 14.4 | 1.7 | 2.9 | 3.6 | 8062 | 2.5 | 7.28 | 13844 |
| **min** | 0.4 | 2.4 | 0.46 | 0 | 1.4 | 0.5 | 62 | 5.6 | 4.1 | 1.5 | 0.1 | 0.7 | 0 | 6167 | 0 | 1.7 | 3552 |
| **max** | 8.5 | 11.7 | 4.94 | 0.23 | 6.7 | 93.2 | 382 | 9 | 13.6 | 28 | 36 | 7.2 | 25.2 | 10704 | 310.63 | 464.6 | 37807 |
| **skewness** | 2.12 | 1.49 | 1.95 | 1.18 | 1 | 5.15 | 1.09 | 0.14 | 0.07 | -0.06 | 5.5 | 0.66 | 2.04 | -0.03 | 5.67 | 10.35 | 0.54 |
| **kurtosis** | 9.16 | 2.87 | 5.19 | 0.67 | 0.51 | 38.92 | 0.47 | 0.4 | -1.15 | -1.13 | 44.55 | 1.13 | 6.87 | -1.67 | 36.96 | 127.82 | -0.97 |
| **J-B test**  **p-value** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0055 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

|  |
| --- |
|  |

Fig. 6. Boxplot with density plot of measured variables

**3.2. Correlation analysis**

The analysis results in Section3.1 indicated that correlation analysis of the measured variables in the data measured from 2017 to 2021, should be performed via Spearman correlation coefficients, a non-parametric method that uses a rank irrelevant to distribution. The results of the correlation analysis to identify associations between measured variables within the analysis data, can be found in Figure 7. The figure on the left indicates the calculation results of the Spearman correlation coefficients, and the figure on the right shows the result of the significance test for this; when there were cases with non-significant correlations, the corresponding parts were marked as X. As a result, Transparency showed a negative correlation with most of the other measured variables, indicating that the results were significant (*P* < 0.05). There were significant positive correlations among water quality items such as BOD, COD, TN, and TP (*P* < 0.05); similar results were observed in hydraulics and hydrological variables, such as LowWaterLevel, inflow, and discharge. Additionally, there was a significant and strong negative correlation between DO and temperature *(P <* 0.05). In other words, it can be numerically confirmed that DO decreases as the temperature of water supply sources increases in summer. However, in the case of such time series data, independency is not generally satisfied. Therefore, if the *P* value for the test confirming the normality for time-dependent measured variables is calculated and confirmed, and based on that, a correlation analysis is performed, there is no major challenge in interpreting corresponding results; however, as there is a clear limitation in the theoretical part, researchers should be careful when analyzing and interpreting the results (Górecki et al., 2018).

|  |  |
| --- | --- |
|  |  |
| **Spearman correlation coefficient** | **Testing significance** |

Figure 7. Correlation analysis

**3.3. Redundancy analysis**

In this section, we explain the RDA results based on the seasonal average of the measured data from 2017 to 2021. In the seasonal average calculations, we classified the period from March to May as spring, the period from June to August as summer, the period from September to November as autumn, and the period from December to February as winter; the results are presented in Figure 8. Figure 8 shows that diatoms such as *Synedra, Aulacoseira, and Asterionella* were mainly found in spring and summer, and harmful Cyanophytes, such as *Aphanizomenon* and *Microcystis*,were found in summer and autumn, even though the cell count was low due to a higher water temperature. In addition, diatoms such as *Aulacoseira* were adapted to areas with a wide range of water temperature regardless of the season, in Juam Lake and Tamjin Lake.

|  |  |
| --- | --- |
|  |  |
|  |  |

Fig. 8. Seasonal variation of dominant Algae species of the study lake

The RDA results are illustrated in Figure 9. The figure on the left presents the distribution of species of algae found during the survey, and the figure on the right indicates the distribution of the seasonal environment. Such figures visually show correlations between each measured variable, and distributions of species of algae and seasonal environments to some extent (Sakizadeh et al., 2021; He et al., 2022). As a result, Figure 9 shows measured variables such as DO, temperature, Chla, and turbidity had high associations with the distributions of species of algae and seasonal environments; in particular, DO, Chla, and pH showed high associations with seasonal environments in spring and winter, whereas temperature and turbidity had high correlations with the seasonal environment in summer and autumn. Furthermore, notably, in the figure regarding seasonal environments (right figure of Figure 9.) , the arrow for DO indicates the direction of spring and winter, and the arrow for temperature presents the direction of summer and autumn. However, the length of the arrow was relatively longer than other measured variables. This is consistent with the strong negative correlation between DO and temperature, which was observed through the correlation analysis. In other words, as summer and autumn are the periods when the water temperature is relatively high, DO decreases during the periods. Since spring and winter are the periods when the water temperature is relatively low, DO increases during the periods. Furthermore, the figure regarding species of algae (left figure of Figure 9.) indicates that there are diatoms, particularly, *Asterionella, and Fragilaria* around the arrow for DO; there are Cyanophytes, particularly *Oscillatoria, and Microcystis*, around the arrow for Temperature. The results indicate that Cyanophytes are mainly found in summer and autumn when the water temperature is relatively high, whereas diatoms are mainly found in spring and winter, when the water temperature is relatively low. RDA has an advantage in that it is possible to clearly classify algae found in water supply sources based on ecological and seasonal characteristics. In a related study, Yu et al. (2014) reported similar results: following an analysis of data measured from 2011 to 2013 at eight weir sections, from Sangju Weir to Hapcheon Changnyeong Weir in the Nakdong River in Korea, diatoms including *Stephanodiscus* were significantly and negatively correlated with water temperature, and harmful Cyanophytes, such as *Microcystis*, were significantly and positively correlated with water temperature.

|  |  |
| --- | --- |
|  |  |
| Algae species | Seasonal environment |

Figure 9. Redundancy analysis ordination showing environmental variables and phytoplankton assemblages on the seasonal samples in Juam and Tamjin Lakes. Circles: spring samples, Squares: summer samples, Diamonds: autumn samples, Up triangles: winter samples.

**3.4. Modeling for cell count of Cyanophytes**

In this section, we present the fitting results of seven models for the cell count of blue-green algae, and compared the prediction performances of the seven models. Prior to model fitting, except for the statistical machine learning technique-based model, as for six probability distribution-based models, we performed standardization for 17 measured variables, which are listed in Table 1, to reduce multicollinearity problems; when fitting models, we used the stepwise regression method, and selected variables assumed to be necessary for predicting the cell count of blue-green algae among measured variables to use them for model fitting. Subsequently, we calculated variance inflation factors (VIF), a statistic that is a criterion for determining multicollinearity problems, and identified how often multicollinearity problems from the model occur; it can be assumed that if the statistics value is < 10, multicollinearity problems are reduced.

**3.4.1. Modeling results**

We conducted a model fitting by setting the measured data from January 2017 to December 2022 as training data. The fitting results of the PRM and NBRM, which correspond to the GLM, can be found in Table 5 and Table 6, respectively.

Table 5. Poisson regression modelling results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Selected variable** | **VIF** | **Estimate** | **Standard**  **Error** | **Z value** | **p-value** |
| **(Intercept)** | X | 2.5660 | 0.0118 | 217.7290 | **0.0000\*\*\*** |
| **BOD** | 2.0610 | 0.2096 | 0.0113 | 18.4690 | **0.0000\*\*\*** |
| **COD** | 9.4699 | 0.5197 | 0.0151 | 34.4250 | **0.0000\*\*\*** |
| **TN** | 3.0730 | 0.0950 | 0.0124 | 7.6600 | **0.0000\*\*\*** |
| **TP** | 9.1553 | -1.1347 | 0.0198 | -57.1920 | **0.0000\*\*\*** |
| **TOC** | 5.8250 | 0.1511 | 0.0097 | 15.6030 | **0.0000\*\*\*** |
| **SS** | 2.6149 | -0.2633 | 0.0143 | -18.3980 | **0.0000\*\*\*** |
| **EC** | 6.2047 | 0.4490 | 0.0151 | 29.817 | **0.0000\*\*\*** |
| **pH** | 1.3401 | 0.0943 | 0.0037 | 25.2790 | **0.0000\*\*\*** |
| **DO** | 1.8163 | -0.3983 | 0.0097 | -41.1760 | **0.0000\*\*\*** |
| **Temperature** | 1.5038 | 1.6676 | 0.0093 | 179.1960 | **0.0000\*\*\*** |
| **Turbidity** | 1.6666 | -0.0705 | 0.0033 | -21.653 | **0.0000\*\*\*** |
| **Transparency** | 1.9554 | -0.4119 | 0.0057 | -72.1130 | **0.0000\*\*\*** |
| **Chla** | 1.5972 | -0.0447 | 0.0039 | -11.5210 | **0.0000\*\*\*** |
| **LowWaterLevel** | **31.9424** | -0.4450 | 0.0221 | -20.0920 | **0.0000\*\*\*** |
| **Inflow** | 1.3606 | -0.0156 | 0.0025 | -6.1440 | **0.0000\*\*\*** |
| **Discharge** | 1.3034 | 0.0439 | 0.0032 | 13.7800 | **0.0000\*\*\*** |
| **Reservoir** | 18.3975 | 0.2860 | 0.0161 | 17.7250 | **0.0000\*\*\*** |
| Significance code: (\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001) | | | | | |

Table 6. Result of Negative Binomial Regression Modelling (NBRM)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Selected variable** | **VIF** | **Estimate** | **Standard**  **Error** | **Z value** | **p-value** |
| **(Intercept)** | X | 2.5974 | 0.2877 | 9.0280 | **0.0000\*\*\*** |
| **BOD** | 2.0837 | 0.1734 | 0.2889 | 0.6000 | 0.5485 |
| **COD** | 9.0067 | 0.5359 | 0.3834 | 1.3980 | 0.1625 |
| **TN** | 3.1230 | 0.0796 | 0.3322 | 0.2400 | 0.8106 |
| **TP** | 8.8756 | -1.1145 | 0.5064 | -2.2010 | **0.0280\*** |
| **TOC** | 5.5053 | 0.1216 | 0.2496 | 0.4870 | 0.6263 |
| **SS** | 2.4888 | -0.1882 | 0.3400 | -0.5530 | 0.5801 |
| **EC** | 5.7872 | 0.4695 | 0.3852 | 1.2190 | 0.2233 |
| **pH** | 1.3059 | 0.0698 | 0.0998 | 0.6990 | 0.4846 |
| **DO** | 1.6948 | -0.3898 | 0.2431 | -1.6040 | 0.1091 |
| **Temperature** | 1.4789 | 1.6661 | 0.2387 | 6.9800 | **0.0000\*\*\*** |
| **Turbidity** | 1.5961 | -0.1049 | 0.0909 | -1.1540 | 0.2487 |
| **Transparency** | 1.5987 | -0.4244 | 0.1382 | -3.0700 | **0.0022\*\*** |
| **LowWaterLevel** | **27.7454** | -0.3653 | 0.5534 | -0.6600 | 0.5093 |
| **Discharge** | 1.2725 | 0.0411 | 0.0822 | 0.5000 | 0.6170 |
| **Reservoir** | **15.9994** | 0.1774 | 0.4057 | 0.4370 | 0.6620 |
| Significance code : (\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001) | | | | | |

According to the results, 17 measured variables were selected in the PRM, and 15 measured variables, excluding Chla and inflow, were selected in the NBRM. However, the two models had a considerable difference in the *P*-value-based significance. In other words, at the 0.05 significance level, all measured variables were significant in the PRM, whereas only TP, temperature, and transparency were significant in the NBRM. Consequently, there is no guarantee that all explanatory variables selected by the stepwise regression method are not always significant. In addition, assuming to what extent multicollinearity problems are reduced based on VIF values, two models had VIF values ≥ 10 only for the maximum of two measured variables, implying that the problem was somewhat resolved. Table 7 and Table 8 present the fitting results of ZIM, specifically speaking, the ZIPM and the ZINBM. However, as the identification of an accurate count process is the main purpose, the estimation of VIFs for the zero-inflation process is meaningless; therefore, we did not calculate them in the R program (Zhang et al. (2020)).

Table 7. Results of Zero-Inflated Poisson Model (ZIPM)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Selected variable** | **VIF** | **Estimate** | **Standard**  **Error** | **Z value** | **p-value** |
| **Poisson Count Process** | **(Intercept)** | X | 5.1169 | 0.0126 | 407.5020 | **0.0000\*\*\*** |
| **BOD** | 3.1347 | -0.1208 | 0.0137 | -8.8200 | **0.0000\*\*\*** |
| **COD** | **12.2424** | 0.3215 | 0.0171 | 18.8100 | **0.0000\*\*\*** |
| **TN** | 5.4373 | 0.2371 | 0.0146 | 16.1930 | **0.0000\*\*\*** |
| **TP** | **19.9139** | -1.1640 | 0.0252 | -46.2650 | **0.0000\*\*\*** |
| **TOC** | 6.8542 | 0.4067 | 0.0109 | 37.3720 | **0.0000\*\*\*** |
| **SS** | 3.1233 | 0.0250 | 0.0117 | 2.1340 | **0.0328\*** |
| **EC** | **11.5964** | 1.0511 | 0.0190 | 55.2510 | **0.0000\*\*\*** |
| **pH** | 1.5855 | 0.0190 | 0.0038 | 4.9520 | **0.0000\*\*\*** |
| **DO** | 8.0638 | 0.2018 | 0.0107 | 18.8590 | **0.0000\*\*\*** |
| **Temperature** | **10.1544** | 0.5045 | 0.0098 | 51.3550 | **0.0000\*\*\*** |
| **Turbidity** | 1.8453 | -0.0285 | 0.0034 | -8.4480 | **0.0000\*\*\*** |
| **Transparency** | 2.5787 | -0.3120 | 0.0064 | -49.1100 | **0.0000\*\*\*** |
| **Chla** | 1.7856 | -0.0768 | 0.0044 | -17.5620 | **0.0000\*\*\*** |
| **LowWaterLevel** | **29.7598** | -0.8615 | 0.0210 | -40.9570 | **0.0000\*\*\*** |
| **Inflow** | 1.5276 | -0.0114 | 0.0027 | -4.1560 | **0.0000\*\*\*** |
| **Discharge** | 1.3452 | 0.0160 | 0.0034 | 4.7180 | **0.0000\*\*\*** |
| **Reservoir** | **17.5851** | 0.3283 | 0.0156 | 21.0170 | **0.0000\*\*\*** |
| **Zero-Inflation Process** | **(Intercept)** | X | 2.8037 | 0.2142 | 13.0870 | **0.0000\*\*\*** |
| **COD** | X | -1.0133 | 0.2774 | -3.6530 | **0.0002\*\*\*** |
| **TOC** | X | 0.7012 | 0.2169 | 3.2320 | **0.0012\*\*** |
| **SS** | X | 0.5196 | 0.2130 | 2.4400 | **0.0147\*** |
| **DO** | X | 0.8204 | 0.2076 | 3.9520 | **0.0000\*\*\*** |
| **Temperature** | X | -2.4409 | 0.2442 | -9.9950 | **0.0000\*\*\*** |
| **Transparency** | X | 0.3274 | 0.1113 | 2.9420 | **0.0033\*\*** |
| Significance code : (\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001) | | | | | | |

Table 8. Results of Zero-Inflated Negative Binomial Model (ZINBM)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Selected variable** | **VIF** | **Estimate** | **Standard**  **Error** | **Z value** | **p-value** |
| **Negative Binomial Count Process** | **(Intercept)** | X | 5.2227 | 0.1361 | 38.3780 | **0.0000\*\*\*** |
| **TP** | 5.7435 | -0.6704 | 0.1997 | -3.3580 | **0.0007\*\*\*** |
| **TOC** | 2.5064 | 0.4740 | 0.0976 | 4.8570 | **0.0000\*\*\*** |
| **EC** | 3.6626 | 0.5630 | 0.1643 | 3.4260 | **0.0006\*\*\*** |
| **DO** | 5.8814 | 0.3243 | 0.1314 | 2.4680 | **0.0136\*** |
| **Temperature** | 7.3732 | 0.4458 | 0.1373 | 3.2480 | **0.0012\*\*** |
| **Transparency** | 1.2868 | -0.3003 | 0.0818 | -3.6720 | **0.0002\*\*\*** |
| **LowWaterLevel** | 2.4821 | -0.4054 | 0.0954 | -4.2480 | **0.0000\*\*\*** |
| **Log(theta)** | X | 0.0960 | 0.0841 | 1.1410 | 0.2539 |
| **Zero-Inflation Process** | **(Intercept)** | X | 2.7984 | 0.2146 | 13.0380 | **0.0000\*\*\*** |
| **COD** | X | -1.0166 | 0.2785 | -3.6500 | **0.0003\*\*\*** |
| **TOC** | X | 0.7062 | 0.2178 | 3.2420 | **0.0012\*\*** |
| **SS** | X | 0.5166 | 0.2136 | 2.4190 | **0.0156\*** |
| **DO** | X | 0.8242 | 0.2085 | 3.9540 | **0.0000\*\*\*** |
| **Temperature** | X | -2.4412 | 0.2452 | -9.9550 | **0.0000\*\*\*** |
| **Transparency** | X | 0.3269 | 0.1118 | 2.9250 | **0.0034\*\*** |
| Significance code: (\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001) | | | | | | |

The ZIM consists of a count process in which the cell count of blue-green algae is modeled based on Poisson distribution or negative binomial distribution, and the zero-inflated process in which the probability of having 0 is modeled. Accordingly, there are two fitted model equations. As a result, in the count process, all 17 measured variables were selected in the ZIMP, whereas seven measured variables (i.e., TP, TOC, EC, DO, Temperature, Transparency, and LowWaterLevel) were selected in the ZINBM. The measured variables that were selected in both models were significant at *P* < 0.05, and for the zero-inflation process, COD, TOC, SS, DO, temperature, and transparency were commonly selected. However, in the case of the ZIPM, there are six measured variables with VIF values ≥ 10, implying that multicollinearity problems were greatly reduced. Table 9 and Table 10 present the fitting results of the GAM, specifically the GAPM and the GANBM.

Table 9. Results of Generalized Additive Poisson Modelling (GAPM)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Selected variable** | **VIF** | **Effects** | | | |
| **Parametric** | | **Non-Parametric** | |
| **F value** | **p-value** | **Npar Chisq value** | **p-value** |
| **f(BOD)** | 2.2974 | 12.0186 | **0.0006\*\*\*** | 998.0000 | **0.0000\*\*\*** |
| **f(COD)** | **11.3675** | 0.0257 | 0.8727 | 609.4000 | **0.0000\*\*\*** |
| **f(TN)** | 2.9250 | 19.9309 | **0.0000\*\*\*** | 3059.2000 | **0.0000\*\*\*** |
| **f(TP)** | **10.6528** | 30.0516 | **0.0000\*\*\*** | 2876.0000 | **0.0000\*\*\*** |
| **f(TOC)** | 6.0887 | 24.2585 | **0.0000\*\*\*** | 1799.5000 | **0.0000\*\*\*** |
| **f(SS)** | 3.3468 | 11.8185 | **0.0006\*\*\*** | 4067.1000 | **0.0000\*\*\*** |
| **f(EC)** | 6.5145 | 1.8360 | 0.1757 | 781.6000 | **0.0000\*\*\*** |
| **f(pH)** | 1.4098 | 0.0448 | 0.8325 | 1312.3000 | **0.0000\*\*\*** |
| **f(DO)** | 1.9553 | 61.4436 | **0.0000\*\*\*** | 2011.4000 | **0.0000\*\*\*** |
| **f(Temperature)** | 1.6738 | 132.9822 | **0.0000\*\*\*** | 1744.4000 | **0.0000\*\*\*** |
| **f(Turbidity)** | 1.5512 | 3.3417 | 0.0679 | 3996.5000 | **0.0000\*\*\*** |
| **f(Transparency)** | 2.0297 | 24.0152 | **0.0000\*\*\*** | 1412.8000 | **0.0000\*\*\*** |
| **f(Chla)** | 1.5247 | 0.0010 | 0.9750 | 1224.2000 | **0.0000\*\*\*** |
| **f(LowWaterLevel)** | **26.4300** | 8.3021 | **0.0040\*\*** | 3740.4000 | **0.0000\*\*\*** |
| **f(Inflow)** | 1.2980 | 3.2911 | 0.0700 | 3241.2000 | **0.0000\*\*\*** |
| **f(Discharge)** | 1.2942 | 0.0018 | 0.9660 | 522.0000 | **0.0000\*\*\*** |
| **f(Reservoir)** | **16.6264** | 0.0360 | 0.8497 | 2974.9000 | **0.0000\*\*\*** |
| Significance code: (\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001) | | | | | |

Table 10. Result of Generalized Additive Negative Binomial Modelling (GANBM)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Selected variable** | **VIF** | **Effects** | | | |
| **Parametric** | | **Non-Parametric** | |
| **F value** | **p-value** | **Npar Chisq value** | **p-value** |
| **f(BOD)** | 2.9925 | 273.5440 | **0.0000\*\*\*** | 13.9050 | **0.0000\*\*\*** |
| **f(COD)** | 7.7668 | 1262.9950 | **0.0000\*\*\*** | 12.0730 | **0.0000\*\*\*** |
| **f(TN)** | 4.0189 | 228.8160 | **0.0000\*\*\*** | 11.6630 | **0.0000\*\*\*** |
| **f(TP)** | 8.6495 | 744.5320 | **0.0000\*\*\*** | 6.6480 | **0.0001\*\*\*** |
| **f(TOC)** | 5.3316 | 197.3340 | **0.0000\*\*\*** | 3.6570 | **0.0122\*** |
| **f(SS)** | 2.1566 | 217.7150 | **0.0000\*\*\*** | 29.6820 | **0.0000\*\*\*** |
| **f(EC)** | 5.0054 | 434.4890 | **0.0000\*\*\*** | 13.7040 | **0.0000\*\*\*** |
| **f(pH)** | 1.3557 | 252.1140 | **0.0000\*\*\*** | 47.8290 | **0.0000\*\*\*** |
| **f(DO)** | 4.9959 | 2157.1660 | **0.0000\*\*\*** | 2.7080 | **0.0441\*** |
| **f(Temperature)** | 4.6201 | 2840.3080 | **0.0000\*\*\*** | 37.3680 | **0.0000\*\*\*** |
| **f(Turbidity)** | 1.3728 | 4638.3240 | **0.0000\*\*\*** | 31.7080 | **0.0000\*\*\*** |
| **f(Transparency)** | 1.4800 | 2901.1650 | **0.0000\*\*\*** | 23.4130 | **0.0000\*\*\*** |
| **f(Chla)** | 1.7287 | 4312.8160 | **0.0000\*\*\*** | 164.0440 | **0.0000\*\*\*** |
| **f(LowWaterLevel)** | **18.5349** | 1373.9320 | **0.0000\*\*\*** | 5.3300 | **0.0012\*\*** |
| **f(Inflow)** | 1.4098 | 12297.9220 | **0.0000\*\*\*** | 251.8660 | **0.0000\*\*\*** |
| **f(Discharge)** | 1.3513 | 1994.2580 | **0.0000\*\*\*** | 8.5040 | **0.0000\*\*\*** |
| **f(Reservoir)** | **11.7430** | 87.8560 | **0.0000\*\*\*** | 8.8920 | **0.0000\*\*\*** |
| Significance code: (\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001) | | | | | |

As the GAM has an additive form of a non-linear function, and estimates the function for each measured variable, a semi-parametric estimation method was utilized. Consequently, significance tests for the parametric effect and non-parametric effect were performed, respectively. As a result, all 17 measured variables were selected in the GAPM and GANBM. However, as for the significance based on the *P* < 0.05, all selection variables were significant in the GANBM; however, BOD, TN, TP, TOC, SS, DO, temperature, transparency, and LowWaterLevel were significant at parametric effect in the GAPM. In addition, as there were VIFs ≥ 10 for a maximum of four variables in both models, multicollinearity problems were resolved to some extent.

**3.4.2. Results of statistical machine learning for predicting cell count of Cyanophytes**

In the present study, we used the RF to first classify the cell count of blue-green algae as 0 or not, used XGBoost for the surveyed points where cell count was classified as non-zero, and applied the XGBRF method to predict the cell count of blue-green algae. As a result of training data-based RF for the initial classification, as shown in Table 11 and Fig. 10, variable importance was observed, which is one of the results based on explainable artificial intelligence (Angelov et al., 2021).

Table 11. Variable Importance from Random Forest

|  |  |
| --- | --- |
| **Explanatory Variables** | **Variable Importance-based Mean Decrease Gini** |
| **BOD** | 6.687975 |
| **COD** | 9.914455 |
| **TN** | 15.680119 |
| **TP** | 11.415743 |
| **TOC** | 8.702956 |
| **SS** | 15.944473 |
| **EC** | 15.563142 |
| **pH** | 10.174798 |
| **DO** | **60.984920** |
| **Temperature** | **109.916273** |
| **Turbidity** | 23.336505 |
| **Transparency** | 10.917926 |
| **Chla** | 12.806143 |
| **LowWaterLevel** | 13.674825 |
| **Inflow** | 11.867803 |
| **Discharge** | 25.094874 |
| **Reservoir** | 15.867041 |

|  |
| --- |
|  |

Figure 10. Variable Importance from Random Forest

Consequently, DO and Temperature were observed to considerably influence environments where blue-green algae occur in Juam Lake and Tamjin Lake. Figure 11 presents the error graph for the results of the RF. According to the graph, a greater number of a tree indicates more convergence of error values to specific values; as the cell count of blue-green algae is a zero-inflated variable, it is possible to see that the error for the case in which the cell count is 0 (YES) had smaller values than the error for the case in which it is not 0 (NO). In addition, as for the legend in the graph, OOB (Out Of Bag) error refers to the calculated error by setting the remaining data, which were not included in the sampling, with replacement allowing repetition in training data, as validation data (Ramosaj et al., 2019).

|  |
| --- |
|  |

Figure 11. Error graph of Random Forest

Whether the cell count was 0 or not was classified based on the RF process, and the cell count of blue-green algae was predicted through XGBRF for the survey period when the cell count was not classified as 0.

**3.4.3. Checking prediction accuracy of seven Models**

To compare the predictive power of seven fitted models for blue-green algae cell count, as shown in Sections 3.4.1. and 3.4.2, the measured data in 2022 were set as test data, and the actual cell count of Cyanophytes were compared with the cell counts predicted using the six models. Although there are diverse indicators for evaluating the predictive power of prediction models, the present study used Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Root Mean Squared Log Error (RMSLE), which are frequently used indicators, to evaluate the predictive power of models, and the formulas are presented in Equations (14) ~ (17).

(14)

(15)

(16)

(17)

Table 12 presents the calculation results of the four criteria for evaluating predictive power. The model with the smallest values across the evaluation criteria is considered the model with the best predictive power.

Table 12. Criteria for checking prediction accuracy

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **PRM** | **NBRM** | **ZIPM** | **ZINBM** | **GAPM** | **GANBM** | **XGBRF** |
| **MSE** | 41017.6600 | 43409.0100 | 146372.0000 | 34793.3000 | 15399.5000 |  | **6771.5780** |
| **MAE** | 115.1636 | 119.6647 | 178.6169 | 108.0713 | 63.0644 |  | **33.6925** |
| **RMSE** | 202.5282 | 208.3483 | 382.586 | 186.5296 | 124.0947 |  | **82.2896** |
| **RMSLE** | 2.8510 | 2.9016 | 3.1611 | 2.9004 | 2.5518 | 15.8213 | **1.6291** |

According to the results, the XGBRF model was observed to be the model with the most accurate prediction for Cyanophyte cell count, based on the data measured from 2017 to 2022, at Juam Lake and Tamjin Lake. Based on the results, the actual cell count of Cyanophytes, and the log value of the cell count of Cyanophytes, which was predicted by each model, were expressed in a line graph based on each survey points (see Fig. 12). The GANB model, particularly, when compared with the other six models, was less accurate in predicting Cyanophyte cell count. However, the result was based on the analysis of monitoring data, which were measured at Juam Lake and Tamjin Lake from 2017 to 2022. If the analysis is performed based on data measured at different survey points or periods, there could be different outcomes. Therefore, although the XGBRF had the most accurate prediction for Cyanophyte cell count in the present study, the model may not always the best option.

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

Figure 12. Observed vs predicted Cyanophyte cell counts

**4. Conclusion**

In the present study, the distributional characteristics of algae found at Juam Lake and Tamjin Lake, which are representative water supply sources in the Yeongsan·Seomjin River in Korea, were investigated, from 2017 to 2022. Water quality measurement network data, algae alert system data, and hydraulics and hydrological data, which were measured during the corresponding survey period, were compiled. Based on the data, d explanatory data analysis and correlation analysis were performed to determine the distributional features of measured variables, and associations between variables. In addition, RDA was conducted to investigate the factors influencing algae presence during the survey. Afterward, the performance of seven prediction models in Cyanophyte cell count prediction was evaluated, and the model that accurately predicted Cyanophyte cell count was identified. The results are summarized as follows.

1. In the explanatory data analysis and correlation analysis based on monitoring data measured at Juam Lake and Tamjinho Lake from 2017 to 2021, transparency showed significantly negative correlations with most of the other measured variables. In addition, there were significant positive correlations among water quality characteristics such as BOD, COD, TN, and TP, and among hydraulics and hydrological variables such as LowWaterLevel, Inflow, and Discharge. In addition, there was a very significant and negative correlation between DO and temperature.

2. According to the RDA results, variables such as DO, temperature, Chla, and turbidity had high correlations with algal species distribution and seasonal environment. In particular, DO, Chla, and pH in spring and winter, and temperature and turbidity in summer and autumn exhibited high correlations with seasonal environments. In addition, in summer and autumn, when the water temperature is relatively high, Cyanophytes were dominant, whereas diatoms were primarily observed in spring and winter, when the water temperature is relatively low.

3. Seven prediction models were selected for prediction of Cyanophyte cell count, considering the distributional features of count variables like cell count. When the prediction performances of the models were compared based on the training data measured from 2017 to 2021, and test data measured in 2022, the XGBRF had the most accurate prediction performance.

The methods presented in this study could facilitate the efficient management of water supply sources, which are critical for the life activities of residents. Furthermore, the methods could be applied to water supply sources in other water systems to identify algal characteristics in corresponding points, and to predict the cell counts of Cyanophytes, which can be toxic to humans, animals, and aquatic organisms.

**Acknowledgements**

This work was supported by a grant from the National Institute of Environment Research (NIER), funded by the Ministry of Environment (MOE) of the Republic of Korea (NIER-2023-01-01-043).

**References**

Kim, S. G. (2017). Green algae and algae warning system. *Water for future, 50*(7), 22-26.

Falconer, I. R., & Humpage, A. R. (2005). Health risk assessment of cyanobacterial (blue-green algal) toxins in drinking water. *International journal of environmental research and public health, 2*(1), 43-50.

Fleming, L. E., Rivero, C., Burns, J., Williams, C., Bean, J. A., Shea, K. A., & Stinn, J. (2002). Blue green algal (cyanobacterial) toxins, surface drinking water, and liver cancer in Florida. *Harmful algae, 1*(2), 157-168.

Lee, K. L., Choi, J. S., Lee, J. H., Jung, K. Y., & Kim, H. S. (2021). Response of epilithic diatom assemblages to weir construction on the Nakdong River, Republic of Korea. *Ecological Indicators, 126,* 107711.

Seo, M. J., Lee, H. J., & Kim, Y. S. (2019). Relationship between coliform bacteria and water quality factors at weir stations in the Nakdong River, South Korea. *Water, 11*(6), 1171.

Choi, D. H., Jung, J. W., Lee, K. S., Choi, Y. J., Yoon, K. S., Cho, S. H., Park, H. N., Lim, B. J. & Chang, N. I. (2012). Estimation of pollutant load delivery ratio for flow duration using LQ equation from the Oenam-cheon watershed in Juam Lake. *Journal of Environmental Science International, 21*(1), 31-39.

Shim, Y. J., Cha, J. Y., Park, Y. S., Lee, D. J., Seo, Y. H., Hong, J. P., & Cho, D. G. (2014). A Study on the Land Purchase Priority Measurement of the Riparian Areas in Yeongsan and Seomjin River Basin-Focusing on the Riparian Areas of the Juam Lake. *Journal of the Korean Society of Environmental Restoration Technology, 17*(1), 173-184.

Lim, J. C., & Kim, T. S. (2018). Analysis of Biodiversity and Ecological Characteristics on Tamjin-river Estuarine Ecosytem. *Journal of Wetlands Research, 20*(2), 181-189.

Martinez, W. L., Martinez, A. R., & Solka, J. (2017). *Exploratory data analysis with MATLAB.* Chapman and Hall/CRC.

Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics, 5*(1), 5-12.

Thadewald, T., & Büning, H. (2007). Jarque–Bera test and its competitors for testing normality–a power comparison. *Journal of applied statistics, 34*(1), 87-105.

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences.* Psychology press.

Capblancq, T., & Forester, B. R. (2021). Redundancy analysis: A Swiss Army Knife for landscape genomics. *Methods in Ecology and Evolution, 12*(12), 2298-2309.

ter Braak, C. J., & Šmilauer, P. (2018). *Canoco reference manual and user's guide: Software for ordination (version 5.10).* Biometris, Wageningen University & Research.

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics, 9*(3), 499-513.

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics, 11*(3), e1460.

Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models.* Chapman and Hall/CRC.

Casella, G., & Berger, R. L. (2021). *Statistical inference.* Cengage Learning.

Mukhopadhyay, N. (2020). *Probability and statistical inference.* CRC Press.

Seo, K. A., Jung, S. J., Park, J. H., Hwang, K. S., & Lim, B. J. (2013). Relationships Between the Characteristics of Algae Occurrence and Environmental Factors in Lake Juam, Korea. *Journal of Korean Society on Water Environment, 29*(3), 317-328.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A. & Smith, G. M. (2009). Zero-truncated and zero-inflated models for count data. *Mixed effects models and extensions in ecology with R,* 261-293.

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero‐inflated count data: Beyond zero‐inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology, 65*(1), 163-180.

Yusuf, O. B., Bello, T., & Gureje, O. (2017). Zero inflated poisson and zero inflated negative binomial models with application to number of falls in the elderly. *Biostatistics and Biometrics Open Access Journal, 1*(4), 69-75.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249-307). Routledge.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine, 47*(1), 31-39.

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers, 12*(1), 469-477.

Kaur, P., Stoltzfus, J., & Yellapu, V. (2018). Descriptive statistics. *International Journal of Academic Medicine, 4*(1), 60.

Đorić, D., Nikolić-Đorić, E., Jevremović, V., & Mališić, J. (2009). On measuring skewness and kurtosis. *Quality and Quantity, 43,* 481-493.

Rebekić, A., Lončarić, Z., Petrović, S., & Marić, S. (2015). Pearson's or Spearman's correlation coefficient-which one to use?. *Poljoprivreda, 21*(2), 47-54.

Górecki, T., Hörmann, S., Horváth, L., & Kokoszka, P. (2018). Testing normality of functional time series. *Journal of time series analysis, 39*(4), 471-487.

He, J., Hou, X. L., & Wang, W. C. (2022). Study of Water Quality Pollution Index, Land-Use and Socio-Economic Factors in Yingkou Irrigation District of China Based on Redundancy Analysis. *Nature Environment & Pollution Technology, 21*(1).

Sakizadeh, M., & Zhang, C. (2021). Source identification and contribution of land uses to the observed values of heavy metals in soil samples of the border between the Northern Ireland and Republic of Ireland by receptor models and redundancy analysis. *Geoderma, 404,* 115313.

Yu, J. J., Lee, H. J., Lee, K. L., Lyu, H. S., Whang, J. W., Shin, L. Y., & Chen, S. U. (2014). Relationship between distribution of the dominant phytoplankton species and water temperature in the Nakdong River, Korea. *Korean Journal of Ecology and Environment, 47*(4), 247-257.

Zhang, X., & Yi, N. (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics, 36*(8), 2345-2351.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(5), e1424.

Ramosaj, B., & Pauly, M. (2019). Consistent estimation of residual variance with random forest Out-Of-Bag errors. *Statistics & Probability Letters, 151*, 49-57.

1. Corresponding author: Kang-Young Jung, National Institute of Environmental Human Resources Development, Education Planning Division, 42, Hwangyeong-ro, Seo-gu, Incheon, Republic of Korea (Tel: +82-32-560-7795, E-mail : happy3313@korea.kr) [↑](#footnote-ref-1)