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-2)

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

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

**1. Introduction**

Water 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 cyanophytes. There are several phases of the prevention activity, which are based on the cell count of harmful cyanophytes, 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 the cell count of harmful cyanophytes, 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 sources, since four types of harmful cyanophytes (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 in South Korea collected samples from several points across water sources every week to monitor water source quality, investigate diverse water quality characteristics, including presence of algae, and 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 cyanophytes and green algae, and appropriate distribution characteristics of TDI and DAIpo were not considered while 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 total coliform count and fecal coliform count. However, they did not consider the distributional characteristics 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 sources in the Yeongsan and Seomjin Rivers, and the data was 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 the algal species emergence were examined. When fitting the prediction model for the cell count of cyanophytes 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 explained how to use a model fitting while considering distribution characteristics of response variables, and suggest how to predict the cyanophyte cell count by combining Random Forest (RF) and extreme gradient boosting statistical machine learning techniques. Furthermore, we suggested strategies for addressing the limitations of the these studies.

**2. Materials and Methods**

**2.1 Study Area**

This study investigated the representative water sources in Yeongsan and Seomjin River basins, Juam Lake, and Tamjin Lake. The Yeongsan River Environment Research Center of the National Institute of Environmental Research attempted to respond to the algae alert system, and to monitor the quality of water sources that provides local residents with domestic water, and to collect as well as 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) respectively. Figure 1 shows the points where samples were collected. Juam Lake is the largest freshwater lake in Jeollanam-do which is also an artificial lake that was established during the construction of the Juam Dam. It was completed after eight years of construction (1984 to 1992) based on the government's long-term multi-purpose dam development plan. The total water storage capacity is 457 × 10⁶ ton, with the basin area being 1,029.41 km², and the reservoir area being 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 of 457 × 10⁶ ton 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. Its total basin area is 193 km² and the total water storage is 191 × 10⁶ ton. Overall, 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 used water quality monitoring network, algae alert system, and hydraulic and hydrological data, which were collected every seven days from January to December 2022 at the corresponding target points from the Environment Information System of the National Institute of Environmental Research. After comprehensively compiling this data, it was combined with data from four target points, and refined into analyzable formats. The variables included in the final collected 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** |

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 monitoring network data, were obtained (Table 1). In addition, data such pH, dissolved oxygen (DO), temperature, turbidity, transparency, chlorophyll-a (Chla), and cell count were collected from algae alert system data. The remaining variables, such as low water level (Low Water Level), 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. The algal species observed at 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. Flowchart of the methodology followed

**2.3. Explanatory Data Analysis**

Explanatory data analysis is a process of performing a broad analysis of the variables in the data to determine the general framework of the data, before analyzing the given data (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 performed a correlation analysis to examine the correlations between water quality items, and hydraulic and hydrological variables. In addition, an RDA was performed to examine the relationships between algal species and algal habitats observed at survey sites, and the environmental variables.

**2.3.1. Correlation Analysis**

Correlation analysis analyzes the relationship between continuous variables included in the target data. In general, Pearson correlation coefficients were calculated, as shown in Equation (1), and corresponding significance tests were 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, these two normality test methods have limitations, as they can be applied only when the variables have the 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 was determined to be appropriate for such variables. The normality of this method was 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 the theoretical values of normal distribution (Thadewald & Büning, 2007). In particular, as the water quality items, or hydraulic and hydrological variables included in the data of this study mostly do not satisfy the normality, analyses based on the Pearson correlation coefficients have lower validity (Cohen et al., 2014). Therefore, we used the Spearman correlation coefficient, which is a correlation coefficient defined by rank, and is irrelevant to distribution, as shown in Equation (2):

(2)

**2.3.2. Redundancy Analysis**

Redundancy analysis (RDA) is frequently used for analyzing ecological data. Researchers can use this analysis to study diverse environmental variables, establish a multivariate regression model for a particular species, conduct 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 the factors with 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 were selected to eliminate multi-collinearity among 17 explanatory variables (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** | **Chloroph**  **-ytes** | **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* |

Based the data measured from 2017 to 2020, five algal species, which were frequently dominant, were selected (Table 2). An RDA was performed on a total of 20 species. The selected species and Code are displayed in Table 3.

**2.4. Modeling for cyanophyte cell count**

The model proposed in the present study aims to predict the cyanophyte cell count. The cyanophyte cell count 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 used, the validity of the analysis would inevitably decrease (Ajiferuke et al., 2015). The probability distribution, which can reflect distribution characteristics of count variables, can be generally classified into Poisson distribution and negative binomial distribution. Therefore, a model for predicting count variables should be based on the two types of distributions. To reduce the multi-collinearity issue prior to model fitting, the standardization process of subtracting the mean of every measured variable used as explanatory variable and dividing the resulting value by the standard deviation was performed, as shown in Equation (3):

(3)

Furthermore, when selecting measured variables with great impacts on prediction of cyanophytes 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 generalized linear model (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). This also includes the multiple regression model. Using this, we explained models to which count variables, such as the cyanophyte cell count, can be applied.

**2.4.1.1. Poisson Regression Model**

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

(4)

This distribution is characterized by the same mean and variance. In other words, .PRM based on the distribution has the representative form [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 as there is a limitation in applying it in diverse situations.

(5)

Unlike the multiple regression model using the least square estimation, the estimator of the regression coefficient for this model should be 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 [Equation (6)]. It is a 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 [Equation (7)]. Therefore, this 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 cyanophyte cell count 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; thus, 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 used: 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 cases where non-zero observed values follow the Poisson distribution, and an excess of observed values are 0 (Loeys et al., 2012). 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 and 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 cases where non-zero observed values follow the negative binomial distribution, and excessive observed values are 0 (Yusuf et al., 2017). The model is based on the same distribution as Equation (10):

(10)

where is the mean of the negative binomial distribution, and is the over-dispersion parameter. This is a gamma function that is defined as for , which is a real number greater than 0. The mean of this distribution is and the variance is . Based on this, and are modeled separately in the ZINBM as follows:

(11)

**2.4.3. Generalized Additive Model**

The GLM assumes that there is a linear relationship between the log value of the cyanophyte cell count, and the measured variables. However, there are some 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, allow non-linear functions for each explanatory variable, and provide an extended form of a GLM (Hastie, 2017). The specific form can be expressed as the part indicating the linear relationship in the GLM being converted to an additive form of a non-linear function. 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 used.

**2.5. Prediction of Cyanophyte Cell Count using Statistical Machine Learning Technique**

This section, explains how to combine and apply two types of statistical machine learning techniques in the prediction of the cyanophyte cell count. These include the random forest, which exhibits excellent performance in classification problems, and extreme gradient boosting, which is appropriate for predicting time series variables. We attempted to first classify whether the cell count is 0 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 (RF)**

The process of random forest (RF) is as follows: First, number of samples () for analysis are extracted via sampling with replacement. Based on that, numerous decision tree models () are established, and prediction results obtained through the process are averaged (). In contrast, based on classification results, bagging is performed to derive final outcomes through multi-voting (). Bagging may be have problems due to correlations between samples, This method compensates for the disadvantages of bagging by randomly and differently selecting all types of explanatory variables for each extracted sample. Through this 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 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 constructing a model using gradients, and repeating the process of remodeling with resultant residuals. The slow performance time and overfitting risks are compensated via improvement through the additional support of parallel learning (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 and predicted values) is modeled, as shown in Equation (14).

(14)

After clarifying the function through this process, the updated new models 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 when overfitting would emerge in advance. 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 the new training data, and setting the remaining 1 data as the 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 cyanophyte cell count by simultaneously using RF, which is frequently used for classification problems, and XGBoost, which exhibits good performance in problems of predicting time series variables. It was possible to classify whether cell count is 0 based on RF and predict cell count at surveyed points based on XGBoost, thereby 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 cyanophyte cell count 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 be 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 measured from 2017 to 2021. Furthermore, we calculated the p-value for the J-B test to identify the normality of the measured time series variables. Figure 6 presents a boxplot with a density plot to visually confirm the distribution of each measurement variable. Based on the results, it is possible to confirm the overall distribution characteristics 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, the central part of the distribution appears sharp if the value is > 0, whereas the central part appears smooth if it is < 0 (Đorić et al., 2009). Regarding descriptive statistics as a whole, skewness had positive values for all variables, except for temperature and low water level, and kurtosis had positive values for all variables, except for DO, temperature, low water level, and reservoir. Therefore, 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. Likewise, since all measured variables violated the normality assumption, a correlation analysis, using spearman correlation coefficients rather than Pearson correlation coefficients, needed to be performed (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)** | **Chl a**  **(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 |

|  |
| --- |
|  |

Figure 6. Boxplot with density plot of measured variables

**3.2. Correlation analysis**

The analysis results in Section3.1 indicate that correlation analysis of the measured variables in the data measured from 2017 to 2021, should be performed via Spearman correlation coefficients, which is a non-parametric method that uses a rank irrelevant to the distribution. Figure 7 displays the results of the correlation analysis that identifies associations between measured variables within the analysis data. 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 the same. 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 among hydraulic and hydrological variables, such as low water level, inflow, and discharge. Additionally, a significant and strong negative correlation was observed between DO and temperature *(P <* 0.05); thus, indicating that DO decreases as the temperature of water 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 a correlation analysis is performed based on this, there is no major challenge in interpreting corresponding results. However, researchers should be careful while analyzing and interpreting the results because there is a clear limitation in the theoretical part (Górecki et al., 2018).

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

Figure 7. Correlation analysis

**3.3. Redundancy analysis**

In this section, we discuss 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, June to August, September to November, and December to February as spring, summer, autumn, and winter, respectively. 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 irrespective of the season, both in Juam Lake and Tamjin Lake.

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

Figure 8. Seasonal variation of dominant algal species of the study lake

The RDA results are illustrated in Figure 9. The figure on the left presents the distribution of algal species found during the survey, and the figure on the right indicates the distribution of the seasonal environment. These figures are a visual representation of the correlations between each measured variable, and distributions of algal species and seasonal environments to some extent (Sakizadeh et al., 2021; He et al., 2022). Figure 9 shows that measured variables, such as DO, temperature, Chl a, and turbidity, had high associations with the distributions of algal species and seasonal environments. In particular, DO, Chl a, and pH showed high associations with seasonal environments in spring and winter, whereas temperature and turbidity exhibited high correlations with the seasonal environment in summer and autumn. Notably, in the right figure of Figure 9, the arrow for DO indicates the direction of spring and winter, and the arrow for temperature indicates the direction of summer and autumn. However, the length of the arrow was relatively longer than that of other measured variables. This is consistent with the strong negative correlation between DO and temperature, which was observed through correlation analysis. Therefore, as the water temperature is relatively high in summer and autumn, DO decreases during these periods. Since the water temperature is relatively low in spring and winter, DO increases during these periods. Furthermore, the left figure of Figure 9 indicates the presence of diatoms, particularly *Asterionella, and Fragilaria*, around the arrow for DO. There are cyanophytes, particularly *Oscillatoria, and Microcystis*, around the arrow for temperature. These 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 as it facilitates the accurate classification of algae found in water 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. They reported that 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 seasonal samples in Juam Lake and Tamjin Lake. Circles: spring samples, Squares: summer samples, Diamonds: autumn samples, Up triangles: winter samples.

**3.4. Modeling for cyanophyte cell count**

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

**3.4.1. Modeling results**

We performed 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 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\*\*\*** |
| **Chl a** | 1.5972 | –0.0447 | 0.0039 | –11.5210 | **0.0000\*\*\*** |
| **Low Water Level** | **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\*\*** |
| **Low Water Level** | **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. 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 multi-collinearity issues are reduced based on VIF values, two models had VIF values ≥ 10 only for a maximum of two measured variables, implying that the issue was somewhat resolved. Table 7 and 8 present the fitting results of ZIM, specifically speaking, the ZIPM and the ZINBM respectively. However, as the identification of an accurate count process was the main purpose, the estimation of VIFs for the zero-inflation process was meaningless. Therefore, we did not calculate them using the R program (Zhang et al. (2020)).

Table 7. Results of the 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\*\*\*** |
| **Chl a** | 1.7856 | –0.0768 | 0.0044 | –17.5620 | **0.0000\*\*\*** |
| **Low Water Level** | **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 the 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\*\*\*** |
| **Low Water Level** | 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 comprises a count process in which the cyanophyte cell count is modeled based on Poisson distribution or negative binomial distribution, and the zero-inflated process in which the zero probability is modeled. Accordingly, there are two fitted model equations. Therefore, 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 Low Water Level) were selected in the ZINBM. The measured variables that were selected in both models were significant at *P* < 0.05, and COD, TOC, SS, DO, temperature, and transparency were selected for the zero-inflation process. However, in the case of the ZIPM, there were six measured variables with VIF values ≥ 10, implying that multi-collinearity issues were greatly reduced. Table 9 and Table 10 present the fitting results of the GAM, specifically the GAPM and GANBM, respectively.

Table 9. Results of the generalized additive Poisson model (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(Chl a)** | 1.5247 | 0.0010 | 0.9750 | 1224.2000 | **0.0000\*\*\*** |
| **f(Low Water Level)** | **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 the generalized additive negative binomial model (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(Chl a)** | 1.7287 | 4312.8160 | **0.0000\*\*\*** | 164.0440 | **0.0000\*\*\*** |
| **f(Low Water Level)** | **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) | | | | | |

A semi-parametric estimation method was used because the GAM has an additive form of a non-linear function and estimates the function for each measurement variable. Consequently, significance tests for the parametric and non-parametric effect were performed. All the 17 measured variables were selected in 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 low water level were significant at parametric effect in the GAPM. In addition, there were VIFs ≥ 10 for a maximum of four variables in both models, indicating that multi-collinearity issues were resolved to some extent.

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

In the present study, we used RF to first classify the cyanophyte cell count as 0 or non-zero, then used XGBoost for the surveyed points where the cell count was classified as non-zero, and finally applied the XGBRF method to predict the cyanophyte cell count. As a result of training data-based RF for the initial classification (Table 11 and Figure 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 |
| **Chl a** | 12.806143 |
| **Low Water Level** | 13.674825 |
| **Inflow** | 11.867803 |
| **Discharge** | 25.094874 |
| **Reservoir** | 15.867041 |

|  |
| --- |
|  |

Figure 10. Variable Importance from Random Forest

Consequently, DO and temperature considerably influenced environments where cyanophytes occurred 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 trees indicates more convergence of error values to specific values. Since the applied the XGBRF method to predict the cyanophyte cell count is a zero-inflated variable, it can be observed 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 the cell count is not 0 (NO). In addition, as for the legend in the graph, OOB (Out Of Bag) error refers to the error calculated 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 cyanophyte cell count was predicted through XGBRF for the survey period when the cell count was not classified as 0.

**3.4.3. Verifying the prediction accuracy of the seven Models**

To compare the predictive power of the seven fitted models for cyanophytes cell count, as shown in Sections 3.4.1. and 3.4.2, the measured data in 2022 were set as the test data, and the actual cyanophyte cell count was 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. The formulae are presented in Equations (14) ~ (17).

(14)

(15)

(16)

(17)

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

Table 12. Criteria for verifying 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 have the most accurate prediction for cyanophyte cell count, based on the data measured from 2017 to 2022 at the Juam Lake and Tamjin Lake. The actual cyanophyte cell count and the log value of the cyanophyte cell count, which was predicted by each model, were expressed in a line graph based on each survey points (Figure 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. An analysis performed based on data measured at different survey points or periods could result in different outcomes. Therefore, although the XGBRF had the most accurate prediction for cyanophyte cell count in the present study, the model may not always be 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 sources in the Yeongsan and Seomjin Rivers in Korea, were investigated from 2017 to 2022. Water quality monitoring network, algae alert system, and hydraulic and hydrological data, measured during the corresponding survey period, were compiled. Based on the data, explanatory data analysis and correlation analysis were performed to determine the distribution characteristics of measured variables, and associations between variables. In addition, RDA was conducted to investigate the factors influencing algal presence during the survey. Subsequently, 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 Tamjin 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 hydraulic and hydrological variables, such as Low Water Level, Inflow, and Discharge. Furthermore, there was a highly significant and negative correlation between DO and temperature.

2. According to the RDA results, the variables DO, temperature, Chla, and turbidity exhibited 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, cyanophytes were dominant in summer and autumn when the water temperature is relatively high, whereas diatoms were primarily observed in spring and winter when the water temperature is relatively low.

3. Seven prediction models were selected for the prediction of cyanophyte cell count, considering the distribution characteristics of count variables, such as 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 method proposed in this study can facilitate efficient management of water sources, which are critical for the life activities of residents. Furthermore, the methods can be applied to water sources in other water systems to identify algal characteristics in corresponding points, and to predict the cyanophyte cell count, 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-2)