Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2019.24195

157 (2019) 195–208
July

Extracting the factors influencing chlorophyll-a concentrations in the Nakdong River using a decision tree algorithm

Yeongdae Cho[a], Yejin Kim[b,*]

[a]*Department of Environmental Engineering, Pusan National University, Busan 46241, Korea, email: nvirocho@gmail.com*
[b]*Department of Environmental Engineering, Catholic University of Pusan, Busan 46252, Korea, Tel. +82-51-510-0621, email: yjkim@cup.ac.kr*

Received 24 October 2018; Accepted 29 March 2019

ABSTRACT

Phytoplankton blooms have a simple generation mechanism but myriad influencing factors. The compounding effects of water quality, weather conditions, and geological factors are site-specific to single river basin and can exhibit distinctive characteristics. To investigate the site-specific patterns of Chlorophyll-a outbreaks for each sub-part of the river, we statistically analyzed the results of water quality monitoring for the Nakdong River to classify the locations of the monitoring stations into upstream, midstream, and downstream sites. Then, based on the combination of the influencing factors, the rules explaining Chlorophyll-a concentration levels were extracted and interpreted using a decision tree algorithm, CHAID. The results revealed that for upstream, weather conditions were the primary factor influencing phytoplankton blooms because of the relative absence of pollutants compared with the midstream and downstream. For midstream, weather conditions and nutritional factors were influential in the generation of phytoplankton blooms. For downstream, the notable amount of pollutants originating from upstream not only reflected a high nutrient level for the phytoplankton but also caused the water quality factor to be the primary cause of phytoplankton blooms. The deduced tree, a tool for data-driven modeling, demonstrated its usefulness by extracting practical influencing factors and patterns of interest from the given data.

Keywords: CHAID; Chlorophyll-a; Data-driven modeling; Feature extraction; Decision tree

1. Introduction

Since the mid-twentieth century, seasonal algae growth in downstream areas has become an important problem for water quality management [1–4]. To investigate the core factor of algal overgrowth, the nutritional factor, research was conducted to identify the level and source of nutrients, including nonpoint sources [5–7]. As a part of this intense interest, research was consistently carried out to identify the biological mechanisms of algal blooming [8,9]. Due to the numerous site-specific environmental factors that exist in algae overgrowth sites, constant efforts have been made to develop a better system modeling technique for accurate forecasting [9,10].

OECD (1982) modeled a linear regression by comparing the relation between Chl-a and total nitrogen (T-N) and total phosphorus (T-P) [11], which are the most significant factors influencing algae overgrowth [12–14]. Furthermore, many researchers [9,15–18] have attempted to model a linear function to describe the relationship between Chl-a and the influencing factors; however, researchers have yet to derive a relationship that can be widely accepted. Such shortcomings are evident, as exhibited in Borics et al. [19], where the coefficients of various linear models differ in value.

Huszar et al. [16] noted that the difficulty in describing the above relation with a linear function demonstrates that phytoplankton blooms are caused by compounding effects and that they cannot be exclusively delimited by nutritional

* Corresponding author.

factors to describe the Chl-a concentration levels. The factors influencing algae overgrowth must include geomorphological, geological, and sociological factors surrounding the river basin. Because the environmental factors described above are considerably site-specific, the regression model previously developed by researchers is useful in a scholarly context but is inappropriate for wide applications.

To address the complexity of the target, data-driven modeling has been applied with good estimation accuracy, such as artificial neural networks [20,21], fuzzy logic [22], and multivariate linear regression [9,23], showing its powerful computational capability and successful performance for nonlinear targets [24]. However, whether the purpose of the study is to create a predictive model or to understand the structure of the problem should be identified. In the case of the preceding examples, the development of predictive models is meaningful as a tool to predict the concentration of Chl-a. However, it is hard to say that they have studied the cause-effect relationships in Chl-a outbreaks.

Data exploration for extracting some meaningful information from a given data set, which is mainly based on statistical data analysis, can be successfully applied to solve this problem by mapping the patterns and deriving influencing factors. The site-specific characteristics and cause-effect relationship related to the topic might be hidden in the values in the database. To extract the relationship between the target and independent variables, the decision tree algorithm has been regarded as a powerful tool. Decision trees, called classification trees for categorical target variables and regression trees for continuous variables, are data mining methods that have been widely used in a number of fields [25], including environmental sciences, for example, the dynamics of zooplankton [26]. Although the extracted decision tree was too simple to explain the complex dynamics, it was meaningful for building the rules to explain natural dynamics from the database. Model trees, having linear regression functions at the leaf nodes, have been suggested as an alternative to neural networks for modeling rainfall-runoff relationships [27] and for modeling water level-discharge relationships [28]. As a generalized concept of a decision tree, M5 model trees showed that tree-type regression can have powerful estimation performance [29,30]. For the cases actively exploring the relationships, statistical methods were applied, such as Bayesian networks [31,32]. As a recent case using the regression tree as a tool to express cause-effect relationships, Park et al. [33] developed stressor-response models for Chl-a. To develop stressor-response trees with acceptable performance, trees were limited to simple structures for each detailed specific condition such as the month. This study attempts to extract various tree routes under as many influencing variables as possible given the aim of extracting the relationship between the influencing factors and Chl-a. Therefore, the use of decision trees in this study was aimed at classification rather than prediction. Thus, the CHAID algorithm was used to perform statistical splitting, and the deduced trees were called decision trees.

In this research, South Korea's Nakdong River is classified into upstream, midstream, and downstream subparts, and we determine whether the segmented subparts share the same characteristics in terms of statistical significance deduced from analysis of variance (ANOVA). Then, the algae growth patterns that reflect the special characteristics of the respective subparts were derived. Finally, the decision tree algorithm was used to determine the factors influencing algae growth in the respective subparts. The decision tree, although its original purposes are in classification or regression, was used for the extraction and interpretation of influencing factors.

2. Materials and methods

2.1. Nakdong river basin

The river basin pertinent to this research is the Nakdong River basin, which is located in the southeastern region of South Korea (127°~129° E, 35°~37° N). The Nakdong River, measuring 506.17 km, is the longest river in South Korea and spans 23,384.21 km^2 to cover 24% of South Korea's territory (Fig. 1). The river basin comprises approximately 780 tributaries and seven metropolitan dams [34]. Approximately 6.7 million people reside within the river basin area, including two metropolitan cities and numerous industrial and agricultural complexes. Sewage water and other wastewater from cities and industrial areas create an inflow of pollutants into the Nakdong River. The annual precipitation rate of the river basin is approximately 1,200 mm, 60% of which occurs in summer from June to September [35]. The precipitation pattern is caused by monsoon conditions and frequent typhoon occurrences in summer. Many multipurpose dams have been constructed along the river, such as the Imha, Andong, Hapcheon, and Namgang dams. The junction between the downstream region and the sea is blocked by estuary dikes. Furthermore, the Four Major Rivers Restoration Project prompted the construction of eight barrages in 2011 to strictly control the flow of the Nakdong River. After the project, hydraulic conditions and other environmental-ecological situations have been evaluated as going through a substantial change. Using all data before and after the project was considered to increase uncertainty of results due to the hydrodynamic effects of the barrages operation [34]. Therefore, the data used were limited to those collected prior to the project.

2.2. Data collection

2.2.1. Water quality data

The water quality data used in this research were collected from the water quality measurements of the Nakdong River, which is managed by the nearby Nakdong River Environment Research Center of the Ministry of Environment. There are 538 water quality monitoring stations scattered throughout the Nakdong River basin operated by Korea Ministry of Environment, ranging from downstream locations, lakes, ponds, agricultural water sources, and others (city and industrial complex sources). Among these, this research used the monitoring results collected from eight water quality measuring locations of the Nakdong River basin, which were regarded as representative point of the mainstream. The water quality data were obtained by four times sampling a month for 19 different categories, which are pH, dissolved oxygen (DO), biological

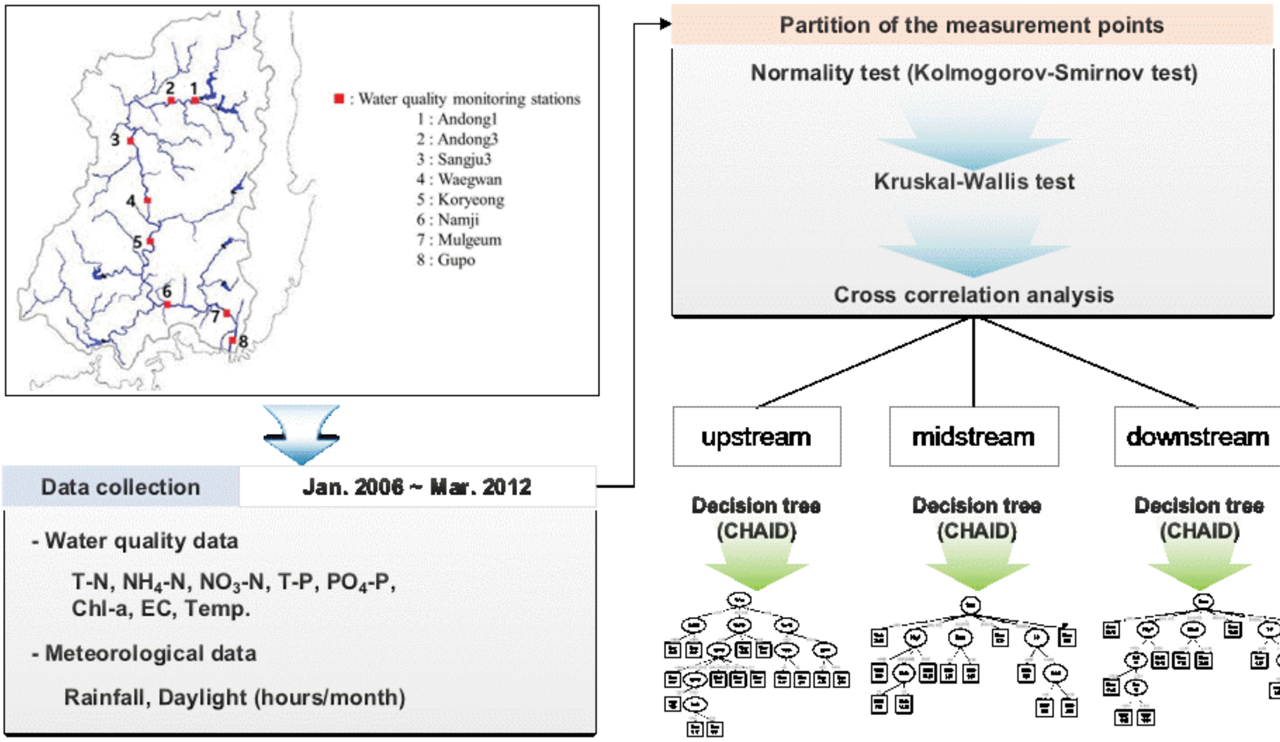

Fig. 1. Nakdong River basin and data treatment process.

oxygen demand (BOD), chemical oxygen demand (COD), total organic carbon (TOC), suspended solids (SS), T-N, NH_4–N, NO_3–N, T-P, Temperature, phenols, electrical conductivity (EC), fecal coliform, total coliform, dissolved total nitrogen (DTN), dissolved total phosphorus (DTP), PO_4–P, and Chl-a. For each measurement, samples were obtained by automatic pumping through sampling pipeline installed to reach proper point for sampling. The standard methods for the examination of water pollution was applied to the measurement of each water quality category except measurable categories by sensor such as pH, DO, and temperature. The water quality monitoring project has been started from 2002 and the data collected and used in this research covered from January 2006 to March 2012, a total of six years and two months considering the effect of the Four Rivers Restoration Project.

For application of the decision tree, the selection of independent variables was required. With respect to the water quality, independent variables were selected among the water quality monitoring station data that exhibited an absence of collinearity. DTN and DTP were not included among the independent variables because they exhibited collinearity with the nitrogen groups (T-N, NH_4–N, and NO_3–N). The high collinearity between T-N and DTN was shown as correlation factor as 0.923 ($p < 0.01$) for Mulgeum and Gupo, 0.971 ($p < 0.01$) for Sangju 3 and Waegwan, 0.971 ($p < 0.01$) for Angdong 1 and Andong 3. The factor values as 0.852 and 0.859 were derived for Mulgeum and Gupo, Andong 1 and Andong 3, respectively. Carbon groups (COD and TOC) and BOD were also not included because algae are autotrophic microorganisms. In addition, SS was not included for exhibiting collinearity with Chl-a concentration levels. The nitrogen groups as T-N, NH_4–N, and NO_3–N and phosphorus groups as T-P and PO_4–P were included as independent variables which can imply the causes of Chl-a outbreaks.

Considering the growth factors of algal blooms, DO and pH levels were not included because they are dependent variables of algal respiration and photosynthesis, whereas water temperature and electrical conductivity were chosen as the independent variables related to water quality. Further, phenol and e-coli were not included because they are not related to algal growth directly. Finally, the water quality independent variables include electrical conductivity, T-N, NH_4–N, NO_3–N, T-P, PO_4–P, and water temperature.

2.2.2. Meteorological data

The meteorological data used in this research were collected by retrieving regional meteorological data (daily precipitation and daylight exposure) from the website of the Korea Meteorological Administration (http://www.kma.go.kr/index.jsp). The regional meteorological administration locations include the Andong, Sangju, and Gumi meteorological offices as well as the Busan Regional Meteorological Office. All regional data were measured in the respective regional locations listed above.

The independent variables for weather conditions include precipitation rate and daylight exposure. The precipitation rate was chosen to reflect the flushing and dilution effect of precipitation on algae. Daylight intervals were chosen to indicate the amount of photons required for algal photosynthesis [14]. However, only daylight intervals cannot reflect the amount of photons over an area due to the seasonal angular path changes of the sun [36]. Thus, the corresponding month was added as another independent variable. Finally, because the daily precipitation rate and daylight exposure can vary significantly, cumulative precipitation rates, ranging from 2 d precipitation to 60 d precipitation rates, were added as independent variables. Additionally, to reflect the effect of time intervals of rainless and the amount of photons, cumulative daylight exposure hours as same way was aggregated into independent variables. As examples, 9 d cumulative precipitation was called "Rain9," and 16 d cumulative daylight exposure hours was called "Sun16".

2.3. Statistical analysis of the characteristics of each region

In the case of Nakdong River, upstream and downstream water quality is different due to discharge of point pollution sources from large cities and industrial complexes from upstream to downstream. In other words, if a decision tree algorithm is applied to all data from upstream to downstream, due to too many different causes for the same Chl-a level, it is difficult to extract meaningful trees having clear patterns of Chl-a outbreaks. Therefore, in this study, statistical preprocessing was performed as correlation analysis, ANOVA, and cross-correlation analysis to confirm that the Nakdong River can be divided into upstream, midstream, and downstream subparts with the measured Chl-a data from 8 monitoring station. Prior to correlation analysis, the data were subjected to normality tests to ensure the implementation of a correlation analysis that best fits the data. The locations that exhibited high correlation coefficients in Chl-a concentration levels were grouped into upstream, midstream, and downstream sections, respectively. Next, a mean difference test was performed to determine whether there was a statistically significant difference in Chl-a concentrations among the three categories. In the mean difference test, if the data in each category passed the normality test, then the data were run through ANOVA method. If the data failed the normality test, then the Kruskal-Wallis test should be conducted as a nonparametric method. Furthermore, a cross-correlation test was conducted to understand the influence of upstream on downstream. The time difference interval for the cross-correlation analysis was set at 7 units before and after the sampling time (–7, 7). Because data measurements were taken once per week, the cross-correlation analysis effectively covered a duration of approximately 50 d. If the results of the cross-correlational test exceeded the confidence limit, then a cross-correlation relationship would exist. This means that despite the statistical significance of the mean difference in Chl-a concentration levels, the upstream influenced the concentration levels of the downstream in some manner. In other words, it would be difficult to ascertain whether the water quality of each subpart is statistically independent from other subparts before confirming using statistical analysis. Therefore, if the results of the cross-correlational test do not exceed the confidence limit, then it could be ascertained that each subpart is statistically independent from other subparts. Of course, the downstream water quality can be affected by the upstream, but this relation means that water quality with statistically independent characteristics of downstream can be produced by the watershed characteristics and pollutant input characteristics of each subpart. This research used a series of statistical analyses to classify a single river basin into three subparts. Then, algal bloom patterns were statistically inferred from the special characteristics of the respective subparts.

2.4. Decision tree

The decision tree algorithm is one of the data mining methods, and it can be used for classification and regression. When the trees were developed for numerical or continuous target variables, it is called the regression tree for prediction. There are popular algorithms for making decision trees, including CART (classification and regression trees), C4.5, CHAID, and QUEST (Quick, Unbiased, and Efficient Statistical Tree) [37,38]. The algorithm used in this research is called CHAID (Chi-squared automatic interaction detection). The CHAID algorithm, originally proposed by Kass [39] and further developed by Magidson [40], is a recursive partitioning method. Originally, it was developed for making classification rules for categorical target variables and later extended to regression [41].

CHAID does not use entropy or the Gini coefficient matrix and instead uses the Chi-square test or *F*-test to initiate a multiway split. In particular, the Chi-square test is used for categorical properties and a *F*-test for continuous variables [40]. When the target variable is continuous, it is transformed into ordinal type, and the algorithm is applied to split the target. In this research, CHAID was chosen to analyze the continuous variable of this research, which is the water quality data. Although Chl-a, a continuous variable, was set as the independent variable to be classified, the decision tree algorithm explored the relationship between the water quality and weather condition variables for various locations. In other words, the individual paths leading to terminal nodes that comprise the decision tree signify the combination of environmental factors that contribute to a certain Chl-a concentration level.

3. Results

3.1. Statistical analysis of the characteristics of each class

3.1.1. Normality test and correlation analysis

After conducting a normality test using the Kolmogorov–Smirnov test, none of the locations met the required statistically significant *p*-value of 0.05 to reject the null hypothesis. The normality test failed, and a nonparametric measure of statistical dependence test had to be used. Spearman's rank correlation analysis was used to classify the data into groups of upstream, midstream, and downstream. The results are summarized in Table 1.

The variables used to determine the classification of data into the three groups were correlation coefficient, location, proximity, and data richness. As a result of the correlation analysis, Andong 1 and Andong 3 (0.601, $p < 0.01$) were categorized as upstream, and Sangju 3 and Waegwan (0.834, $p < 0.01$) and Mulgeum and Gupo (0.893, $p < 0.01$) were categorized as midstream and downstream. Namji and Koryeong (0.744, $p < 0.01$) failed to be categorized because of insufficient information. Locations that exhibited high correlations were paired to display their Chl-a concentration levels in Fig. 2. Koryeong and Namji (0.744, $p < 0.01$) were

Table 1
Results of Spearman's rank correlation analysis for Chl-a from each monitoring station

	Correlation Coefficient							
Spearman's rho	Andong 1	Andong 3	Sangju 3	Waegwan	Koryeong	Namji	Mulgeum	Gupo
Andong 1	1.000	0.601**	0.434**	0.437**	0.422	0.119	–0.059	–0.020
Andong 3		1.000	0.559**	0.507**	0.905**	–0.175	–0.081	–0.035
Sangju 3			1.000	0.834**	0.696**	–0.140	–0.359**	–0.163
Waegwan				1.000	0.641**	0.129	–0.254*	–0.138
Koryeong					1.000	0.744**	0.139	0.161
Namji						1.000	0.417**	0.662**
Mulgeum							1.000	0.893**
Gupo								1.000

*Correlation *p*-value 0.05 (both sides).
**Correlation *p*-value 0.01 (both sides).

excluded because they are affected by the joining of big tributaries as Geumho River and Hwang River, respectively. This explains the relatively low correlation coefficient as 0.744 compared with the 0.834 of between Sangju 3 and Waegwan and to the 0.893 of between Mulgeum and Gupo. Therefore, in consideration of the abundance of data and unknown uncertainty by big tributaries, the Koryeong and Namji were not selected and Sangju 3 and Waegwan were assigned as midstream.

From the results of previous studies that explored the changes in water quality according to the flow of the Nakdong River as a mean value from 2003 to 2016 [42], BOD concentration was less than 1 mg/L until the Andong 3 but it was elevated to about 1.37 at Sangju 3 due to join of pollutant loadings from three tributaries, Naesung, Byeonseong, and Wicheon. BOD concentration, maintained as under 2 mg/L from Sangju 3 to Waegwan, was elevated again as over 2 mg/L at Koryeong due to the joining of Keumho River with high pollutant loadings originated from Daegu metropolitan City. Between Waegwan and Koryeong, there are three industrial complex, Daegu, Sungseo and Gumdan. After Koryeong to Namji, Dalsung 1 and Dalsung 2 industrial complexes are located. Taking these geographic locations together, it cannot be regarded as all the measurement points from Sangju 3 to Namji are bounded as one category. To be more precise, Sangju 3 and Waegwan can be bounded as first part of midstream and Koryeong and Namji as second part of it. As the characteristics of pollutant loading, Andong 1 and Andong 2 can be assigned as upstream, and Sangju 3 and Waegwan represent a part affected by the pollutant loads. The part after Waegwan to the end of the river can be described as a part which receives a lot of pollutant loads and also has dilution effect by joining of large tributaries.

In all three subparts, the patterns of Chl-a growth were similar between paired locations. To conduct a comparative analysis of the subparts, Andong 1, Waegwan, and Mulgeum were selected as the primary upstream, primary midstream, and primary downstream locations, respectively. Andong 1 and Waegwan were selected as the corresponding primary subparts because their large distance would better highlight the different characteristics of the two subparts. Mulgeum was selected as the primary downstream location instead of Gupo because Gupo's proximity to the coast could cause the data to reflect the characteristics of coastal areas rather than that of downstream areas. The water quality data of each subpart from 2006 to 2010 are summarized in Table 2.

3.1.2. Analysis of variance

An ANOVA was required to determine the statistical significance of the mean difference of Chl-a concentrations among the primary subpart locations. As demonstrated earlier, a common ANOVA method cannot be used because the Chl-a concentration data of each subpart failed the normality test. Instead, the ANOVA was conducted using a nonparametric variance method, the Kruskal-Wallis test. The test revealed that the p-value did not exceed 0.05, confirming that the mean difference in Chl-a concentrations among subparts was statistically significant. The three subparts were clearly differentiated with respect to their Chl-a concentration levels.

3.1.3. Cross-correlation

A cross-correlation analysis was conducted to understand the influence of the upstream subpart on the midstream subpart and the midstream subpart on the downstream subpart and to determine whether each subpart was independent from the other subparts. The time difference interval was set at 7 units before and after the sampling time (–7, 7) as covering 50 d before and after the measurement. The result of the cross-correlation analysis is presented in Fig. 3. The upstream and midstream subparts did not exhibit cross-correlation because no time intervals unit exceeded the confidence limit (Fig. 3a). On the other hand, the midstream and downstream subparts exhibited a negative cross-correlation as most time intervals exceeded the negative confidence limit (Fig. 3b). However, the increasing distributive characteristics of Chl-a concentration levels should reflect a positive cross-correlation. A negative cross-correlation is a nonsignificant result. In conclusion, the cross-correlation analysis demonstrated that the influence of the upstream subpart on the conditions of the downstream subpart is virtually negligible, suggesting that each subpart is statistically independent from other subparts.

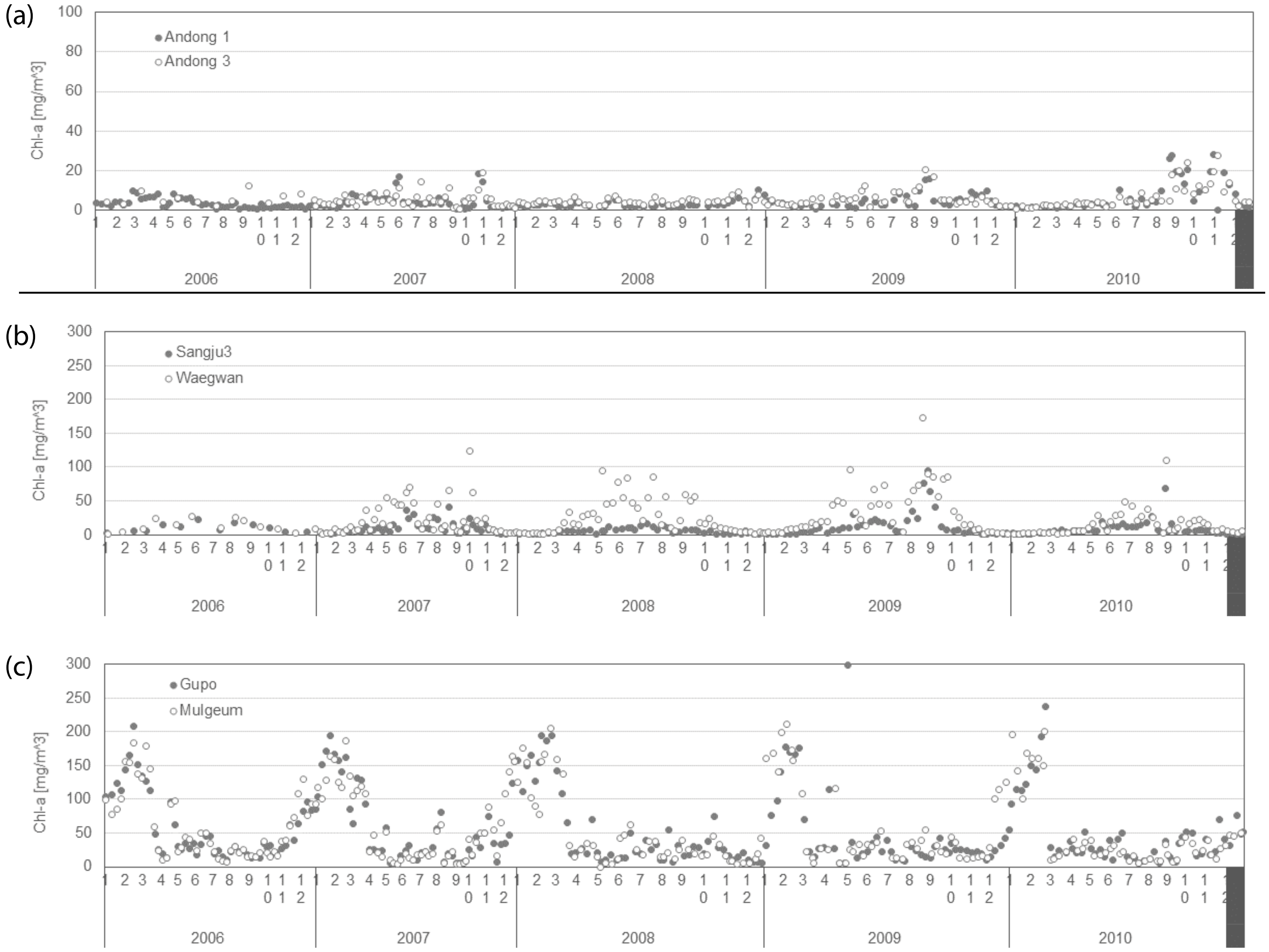

Fig. 2. Chl-a concentration levels (a) Upstream, (b) Midstream, and (c) Downstream.

Table 2
The water quality and grade of each subpart

Section	Mean concentrations of water quality index					
	pH	BOD (mg/L)	SS (mg/L)	DO (mg/L)	T-P (mg/L)	Total coliforms (coliforms/100 mL)
Andong 1	7.5	0.9	5.6	10.4	0.030	666
Waegwan	7.9	1.4	14.2	10.1	0.050	285
Mulgeum	7.8	2.5	18.1	10.6	0.140	472

3.2. Interpretation of the decision trees

3.2.1. Interpretation of the upstream decision tree

The upstream decision tree was statistically inferred from the Andong 1 and Andong 3 water quality data and regional weather data. The results are presented in Fig. 4.

For growing the decision tree, the max tree depth was set to 4, at least 25 cases of data for upper nodes and 15 cases of data for lower nodes were applied as restrictions of growing trees. The margin of error was set as 0.05 % in the algorithm and the independent variable range was 10. A tree depth of 4 and independent range of 10 imply that the number of internal nodes excluding the root node would be 4 with 10 or less independent variables. The minimum number of cases as 25 and 15 mean a lower limit on the amount of data to be included to each edge for an internal node to be divided for tree growth and a terminal node to stop the tree growing, respectively. The margin of error means the 0.05% significance limit of Chi-square test and *F*-test when executing a multiway split of independent variables.

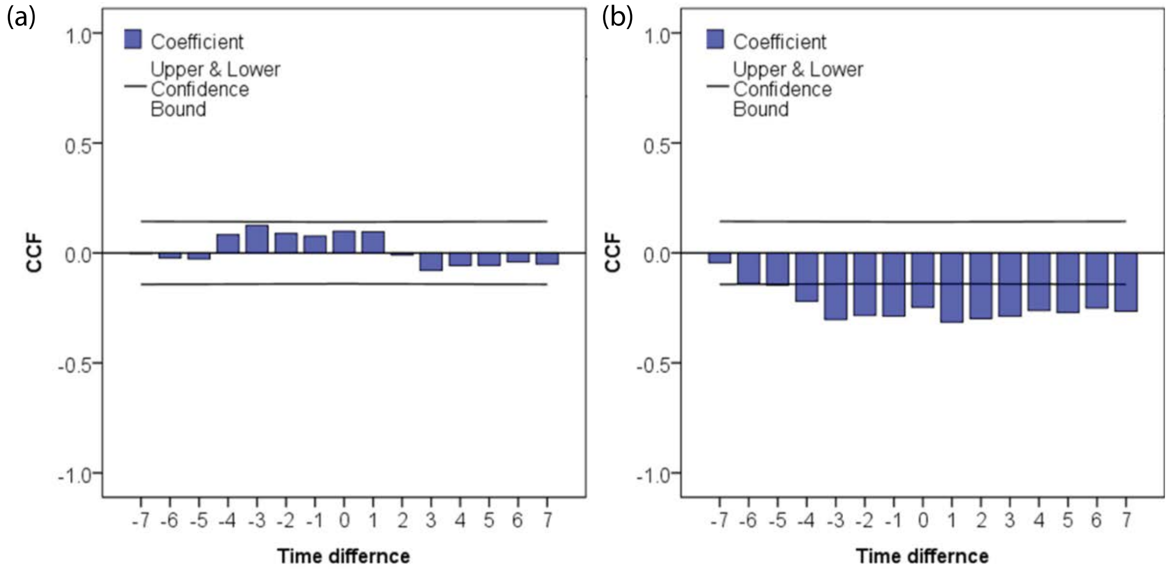

Fig. 3. Cross-correlations (a) Upstream-Midstream (Andong 1 including Waegwan) and (b) Midstream–downstream (Waegwan including Mulgeum).

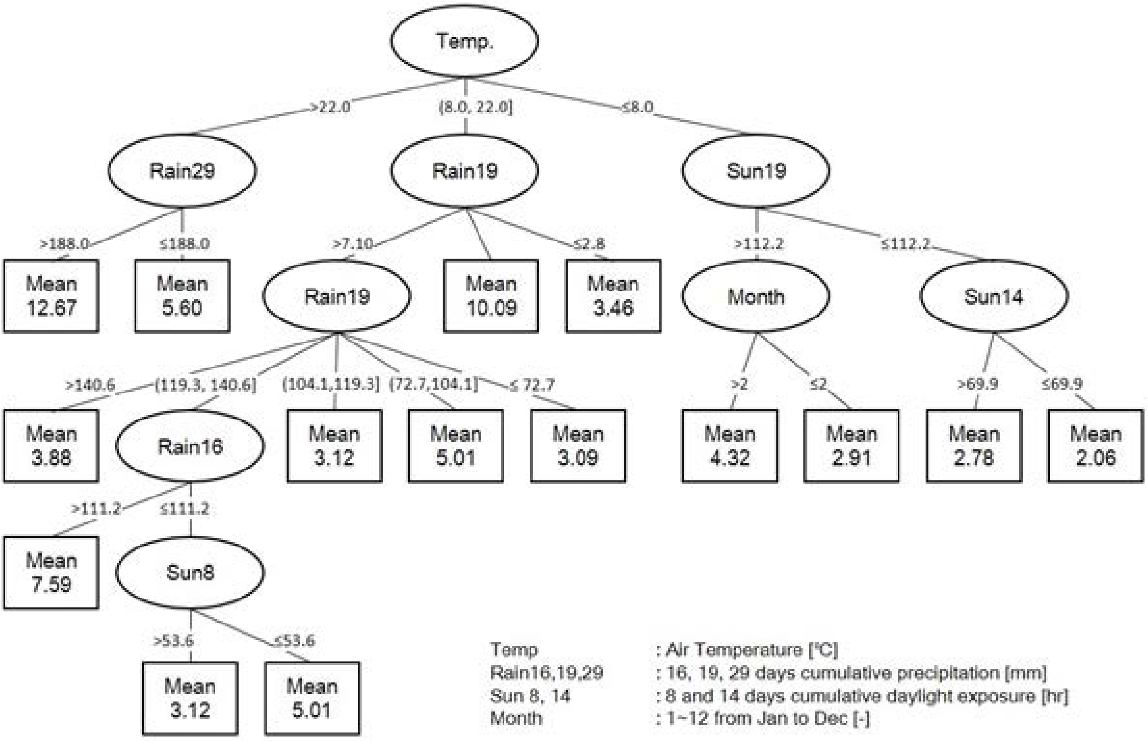

Fig. 4. The upstream decision tree.

These specific conditions for the tree growth were selected through trial and errors method. The large value of the tree depth and independent variable caused larger trees including meaningless patterns. The tree growth results in eight independent variables being highlighted among all the variables in the decision tree. The eight variables obtained from November 2011 to March 2012 were applied to the tree for the purposes of validation and results are presented in Fig. 5.

The distinctive characteristics of the upstream decision tree algorithm for classifying Chl-a growth were that the organizational structure of the tree was relatively simple and that all the influencing variables were related to weather. Because the water quality of the upstream is satisfactory, the weather condition was the determinant that affected the Chl-a concentration levels. This observation can be explained by the lack of industrial complexes and metropolitan cities

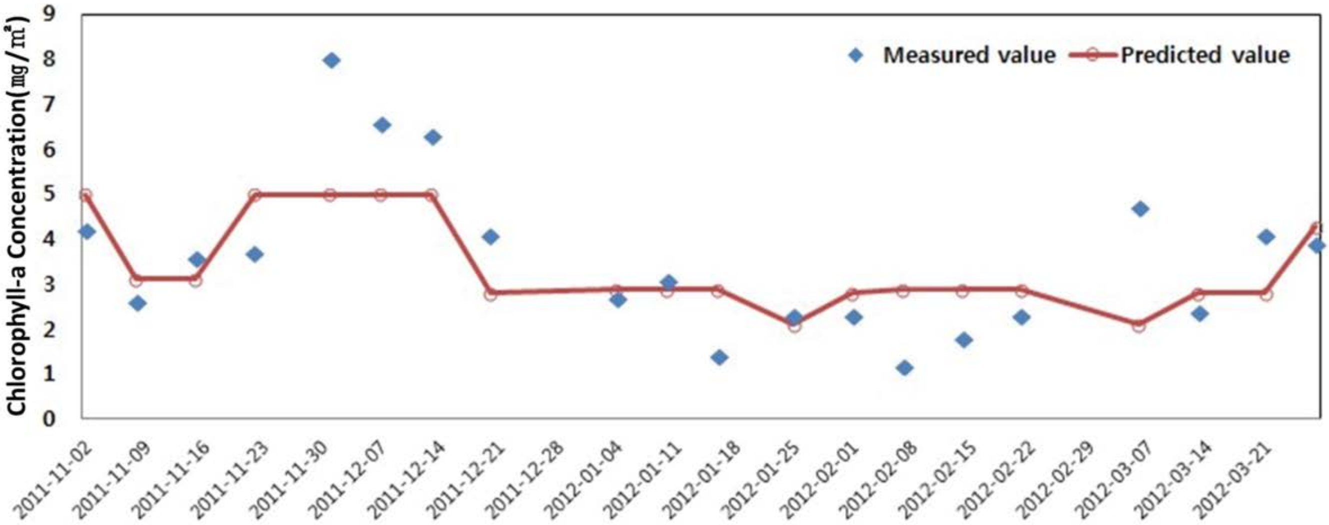

Fig. 5. Upstream tree validation (Andong 1).

near the upstream area, which decreases the inflow of pollutants. According to the upstream decision tree algorithm, the primary influencing factor of the Chl-a concentration was water temperature, which is a factor that is closely related to weather. Regarding secondary factors that contribute to the Chl-a concentration levels, the fact that the precipitation rate variable changes at regular intervals in tandem with the daylight exposure variable suggests that grouping the precipitation rate and daylight exposure as a single factor is appropriate. Through the validation graph designed to determine the predictive capacity of decision trees, it was observed that the tree's classifying rules correspond relatively accurately to the Chl-a concentration levels (RMSE:1.296). Such a correspondence indicates that the Chl-a blooming patterns were accurately inferred by the decision tree algorithm.

3.2.2. Interpretation of the midstream decision tree

To classify the Chl-a concentration levels in the midstream area, a decision tree algorithm was constructed using the water quality and weather data of Sangju 3 and Waegwan. Because the Chl-a concentration varies greatly by location for midstream areas, location was added as a special variable for the algorithm. Fig. 6 shows the decision tree algorithm statistically inferred from the data of midstream locations, both Sangju 3 (Fig. 6a) and Weagwan (Fig. 6b). It was derived as one decision tree including a variable called a measuring point Sangju 3 and Weagwan, but expressed separately in Figs. 6a and 6b, showing the same starting node as "Temp". Also, Figs. 6a and 6b have same patterns when the temperature under 8°C and higher than 26°C. For the temperature is in the range of 8°C ~ 12°C, Chl-a level of Sangju 3 was 5.2 mg/m^3, whereas 10.18 mg/m^3 for Waegwan. Temperature between 12°C ~ 19°C for Sangju 3, 4 d accumulated rainfall was selected as an effecting factor to Chl-a level. The relatively high temperature, between 19°C ~ 26°C, showed inverse relationship between phosphorus concentration and algal concentration for both cases of Sangju 3 and Waegwan. This tendency is also found in the literature on algal bloom dynamics, because in order to produce algal blooms above a certain concentration, the inorganic phosphorus present in the water body must be uptaken [43,44].

For growing the decision tree, the conditions were set as a maximum tree depth of 5, a minimum case number of 20 upper nodes and 10 lower nodes, 0.05% margin of error in the algorithm, and an independent variable range of 10. Among the input independent variables, eight variables were utilized except the location of measurement. To determine the functional capacity of the decision tree, the algorithm was additionally tested with the water quality data from November 2011 to March 2012, as shown in Fig. 7. Similar to the upstream subpart, the primary variable that affects Chl-a concentration levels is water temperature, indicating that weather conditions are the main influencing factor of Chl-a concentration levels. However, the midstream decision tree differs from the upstream decision tree because it exhibits an even distribution of water quality and weather condition variables. This increased influence of water quality variables in this decision tree compared with that of the upstream can be explained as reflecting the pollutant loading effect from three tributaries to the water quality of Sangju 3 and the pollutant loadings from the Gumi industrial complex located between Sangju 3 and Waegwan.

The inflow of wastewater consequently increased the significance of water quality variables in influencing the Chl-a concentration levels. The influence of water quality variables was especially evident during a drought, where a low precipitation rate limited the flushing and dilution effect that would normalize the water quality. Furthermore, in the case of NO_3–N and PO_4–P, because the decrease in the concentration of nutrients increased the Chl-a concentration levels, it can be concluded that the decrease in the concentration of nutrients is the remaining nutrient after the algal blooms had already utilized the available nutrients for consumption. The above results align with the previous findings that there is a decrease in the concentration level of inorganic compounds and nitrogen in locations with abundant algal blooms.

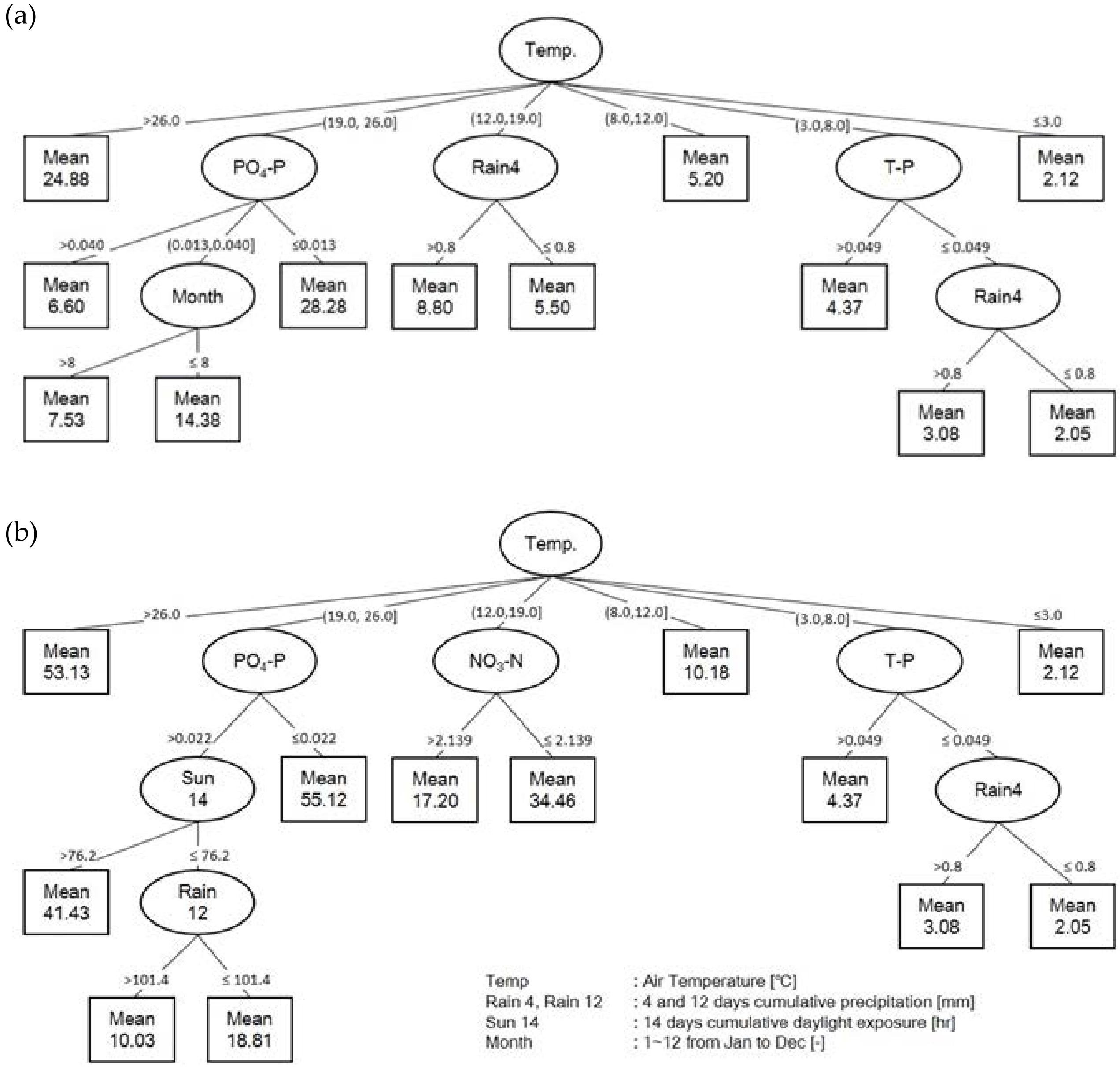

Fig. 6. The Midstream decision tree (a) Sangju 3 and (b) Waegwan.

Through the validation graph (Fig. 7) designed to determine the reliability of decision trees, it was shown that the tree's classifying rules correctly reflected low Chl-a concentrations but failed to accurately reflect high Chl-a concentrations. After performing a causal analysis with the March data of Chl-a/T-P/T-N/precipitation rate from 2005 to 2012, it was observed that the concentration of T-P had decreased from previous years despite a steady precipitation rate and a great increase in Chl-a concentration levels. It was presumed that this observation was the result of a barrage construction in the Nakdong River that disrupted the influence of environmental factors since 2011. To derive a more accurate model, the above human factor must be accounted for, and the consequent data must be reapplied to the decision tree algorithm.

3.2.3. Interpretation of the downstream decision tree

To classify the Chl-a concentration levels in the downstream area, a decision tree algorithm was constructed using water quality and cumulative 60-day weather condition data from Mulgeum and Gupo. The result is shown in Fig. 8. For growing the decision tree, the conditions were set as a minimum case number of 5 upper nodes and 5 lower nodes, a 0.05% margin of error in the algorithm, and an independent variable range of 10. Among the independent input variables, 33 variables were utilized. To determine the functionality of the decision tree, the algorithm was additionally tested with the water quality data from November 2011 to March 2012, as shown in Fig. 9. Similar to the midstream decision tree, weather condition variables and

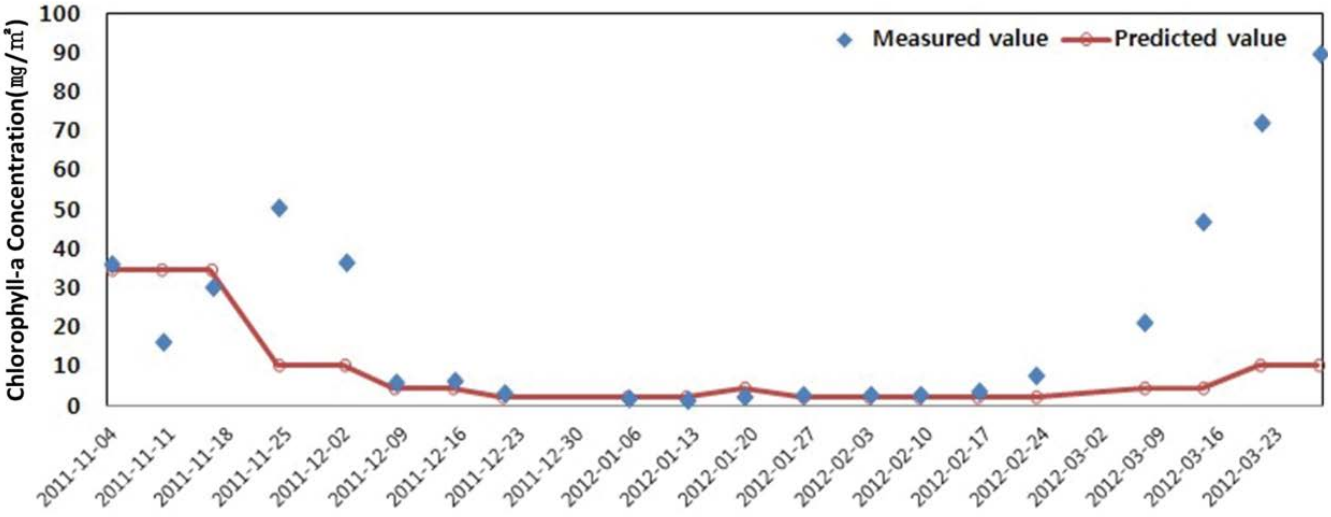

Fig. 7. Midstream tree validation (Waegwan).

water quality variables were evenly distributed, with water quality variable as the primary factor in the distribution. This result is due to the higher Chl-a concentrations in the downstream water quality variables compared with the midstream water quality variables. Similarly, water quality variables remain the variables of primary importance during a drought because the nutritional factor is the most important during a drought.

The factor that has the most significant influences on Chl-a concentration levels is the "corresponding month". The effect of monthly variations in mechanisms explaining Chl-a breaking out has also been mentioned by existing fundamental research [45,46]. To reflect the seasonality of Chl-a, the trees suggested in Fig. 8 were divided into three sections based on the "corresponding month". Fig. 8a shows the results from January to June, from the winter season to the dry summer. Figs. 8b and c show the results from the wet summer season and from the fall to the winter season, respectively.

This difference in Chl-a concentrations during different seasonal intervals demonstrates that the varying intensity of sunlight and rain play a critical role in determining the degree of a phytoplankton bloom. Because the Nakdong downstream is a river-reservoir hybrid system owing to its blockage by the estuary dikes, the flushing effect of precipitation diminishes, and the intensity of the sunlight effect increases.

The effects of electrical conductivity were confirmed in the months of December and September. In December, as the electrical conductivity level increased, the concentration of Chl-a decreased. In September, as the electrical conductivity decreased, the concentration of Chl-a increased. The December data correspond to the observations of an inflow of sea water due to drought, which increases electrical conductivity and decreases phytoplankton blooms. The September data are the result of a combination of factors including adequate water levels, water temperature, nutrition level, and stable water composition, all of which facilitate the growth of Chl-a concentrations.

Through the validation graph designed to determine the reliability of decision trees, it was shown that the tree's classifying rules correspond considerably accurately to the Chl-a concentration levels (RMSE:30.768). Because the downstream decision tree algorithm is influenced by the directly measured water quality variables as opposed to the indirectly measured weather condition variables, the decision conforms especially well to the Chl-a concentrations.

4. Discussion

This research attempted to classify the Chl-a breakout patterns by constructing decision tree algorithms based on the statistical analysis of water quality and weather condition data from eight core locations of the Nakdong River basin. The Nakdong River basin was divided into three subparts based on the Chl-a concentration levels, and the classification rules of the decision tree algorithms determined that the conditions created by the weather condition variables are at least as influential or more influential than the nutrition levels. As we progressed from upstream to downstream, the primary influencing factor progressively changed from weather conditions to water quality.

In the upstream region, the maximum concentration of Chl-a was lower than 50 mg/m^3. Thus, in the tree derived for the upstream, it has been shown that the factors influencing the variation of Chl-a in the low concentration range are meteorological factors such as rainfall and daylight hour rather than pollutants. This can be supported by the fact that the average T-P concentration at Andong 1 in Table 2 is 0.030 mg/L. The tree for midstream started with the temperature. It means that the first factor to divide the concentration range of Chl-a was the temperature. However, the following factors were water pollutants such as PO_4–P, T-P and NO_3–N. This is due to the input of pollutants from point and non-point sources as the river flows from the upper to the middle stream. One characteristic of the midstream tree is that the mean value of Chl-a of Sangju 3 was 5.2 mg/m^3 at temperatures between 8°C–12°C, while

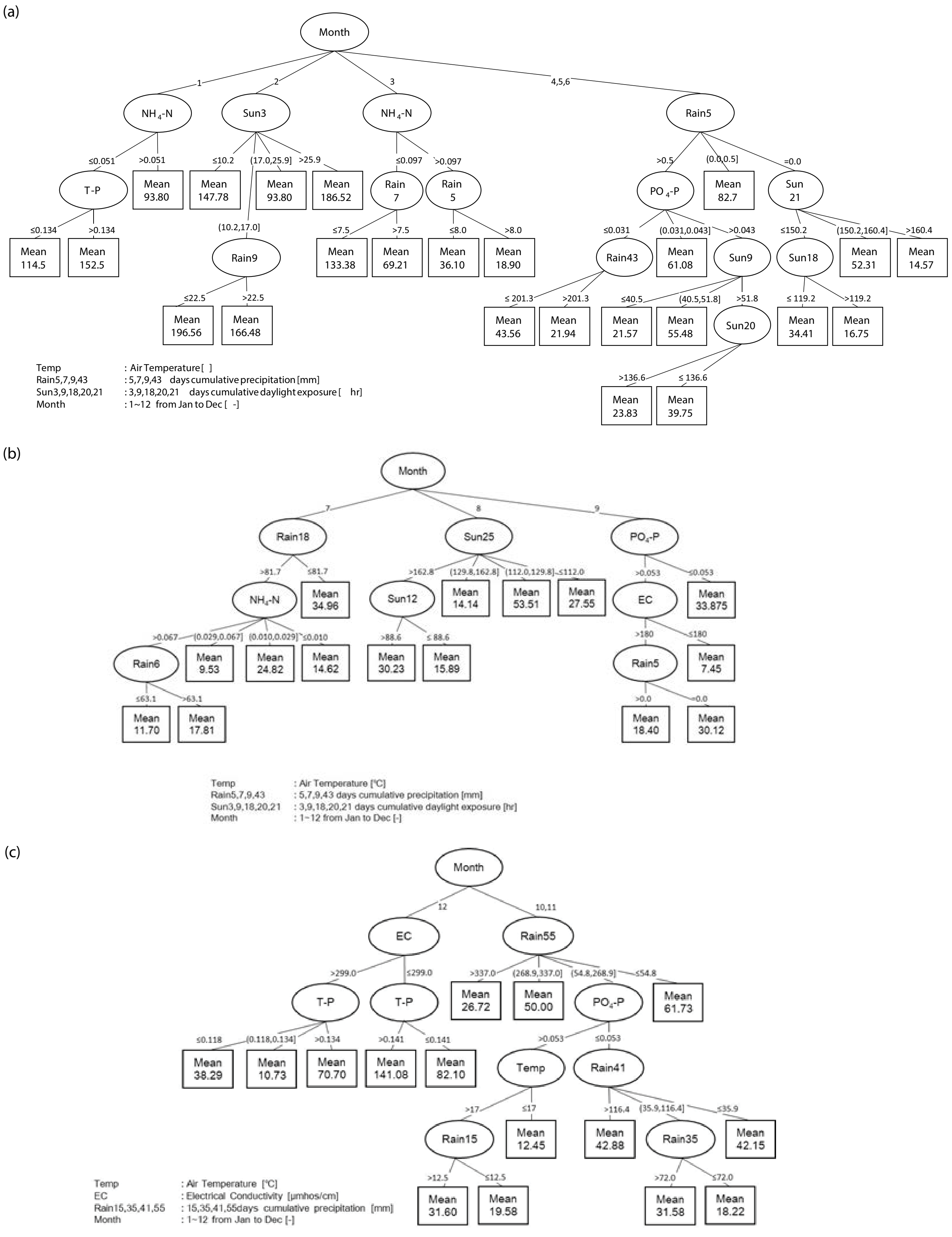

Fig. 8. The downstream decision tree (a) From January to June, (b) From July to September, and (c) From October to November.

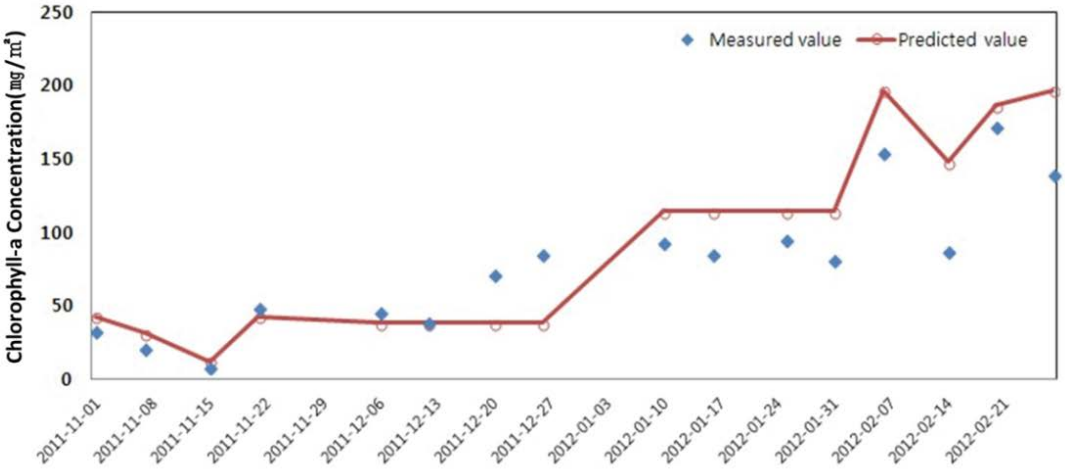

Fig. 9. Downstream tree validation results (Mulgeum).

it was 10.08 at Waegwan at the same conditions. This shows that the tree extracted the pattern of the concentration effect of Chl-a or the occurrence of additional Chl-a along the river flow direction. In addition, the middle tree could express the combined effect of pollutant and meteorological factors. For example, when the temperature was greater than 19°C and less than 26°C, PO_4–P greater than 0.022 mg/L, and the 14 d cumulative daylight exposure greater than 76.2 h, the concentration of Chl-a was as high as 41.43 mg/L. On the other hand, when the 12 d accumulated rainfall was greater than 101.4 mm under the same conditions, the mean of the concentration of Chl-a was 10.03 mg/m^3, indicating that the dilution and wash out by rainfall could be expressed. For the downstream, a complex tree with mixed effects of pollutants and meteorological factors was derived. This means the frequent algal blooms in downstream and it was possible to distinguish the factors mainly influential in monthly.

For the validation of each tree, it would have been more desirable if the data was prepared for longer period or for events of algal blooms. The data from 6 years contained only 6 times of summer, so it was not enough to be used for both of tree formation and validation. More amount of accumulated data will make possible to build more reasonable tree. However, it should be noticed that the Chl-a levels of validation data, under 40 mg/m^3 for Waegwan and under 150 mg/m^3 for Mulgeum, covers the range of 79% of and 89.5% for total Chl-a distribution, respectively. If limited to the period from November to February, the 12 validated data for Waegwan covered 22.2% of total and 16 data for Mulgeum was 19.0%.

Thus, various patterns of Chl-a occurrence could be extracted using decision tree. This shows that the effecting factors for the Chl-a concentration are very complex and site-specific, as Borics et al. [19] and Huszar et al. [16]. However, the decision tree algorithm also has its limitations, for example, that an overly large tree can be derived, or it can be interpreted as if causal relations exist by calling meaningless variables. This depends on how much of the raw data entered into the decision tree algorithm has information that is useful to the purpose of the study. This is why raw data is divided into upstream, downstream and downstream.

Also, the results indicate that one must consider the compounding effects of weather condition variables, hydraulic retention, and water quality to effectively control algal overgrowth. Furthermore, a closely connected data-sharing with the weather forecast is recommended.

An important issue in this paper is that this study does not aim at predicting or estimating algal concentrations. The results suggested are meaningful in that many of the factors already known to affect the algal blooms are site-specific and that they can be identified through data – driven modeling methodologies. Therefore, the results of how well the trees simulates actual values as presented in Figs. 5, 7, 9 do imply the reliability of the trees, not that the possibility as a tool of prediction.

5. Conclusions

This research confirmed the usefulness of statistical methods in analyzing the wide range of data characteristics of river basins. The implementation of statistical methods was especially useful considering the large amount of data and conditions that limit experimentation. Furthermore, if the decision tree algorithm constructed in this research is used as a data analytics tool, it should be noted that the decision tree has the advantages of fewer variables, faster computation, and minimal experimentation over conventional mathematical modeling. However, because the decision tree algorithm cannot process natural scientific mechanisms on its own, the reliability of the algorithm will improve if the algorithm usage is supported by interpretations of natural scientific mechanisms.

Funding

This study was supported by Korea Research Foundation (Project no. 2017R1A2B4007114) and Korea Ministry of Environment as ''Program for promoting commercialization of promising environmental technologies (Project no. 2017001930001).

Conflicts of interest

The authors declare that they have no conflict of interest.

References

[1] W.K. Dodds, W.W. Bouska, J.L. Eitzmann, T.J. Pilger, K.L. Pitts, A.J. Riley, J.T. Schloesser, D.J. Thornbrugh, Eutrophication of U.S. freshwaters: analysis of potential economic damages, Environ. Sci. Technol., 43 (2009) 12–19.

[2] M.B. Edlund, D.R. Engstrom, L.D. Triplett, B.M. Lafrancois, P.R. Leavitt, Twentieth century eutrophication of the St. Croix River (Minnesota–Wisconsin, USA) reconstructed from the sediments of its natural impoundment, J. Paleolimnol., 41 (2009) 641–657.

[3] H.-I. Eum, S.P. Simonovic, Integrated reservoir management system for adaptation to climate change: the Nakdong River Basin in Korea, Water Resour. Manage., 24 (2010) 3397–3417.

[4] G.J. Smith, V. Daniels, Algal blooms of the 18th and 19th centuries, Toxicon, 142 (2018) 42–44.

[5] A.F. Bouwman, L.J.M. Boumans, N.H. Batjes, Estimation of global NH_3 volatilization loss from synthetic fertilizers and animal manure applied to arable lands and grasslands, Global Biogeochem. Cycles, 16 (2002) 8-1–8-14.

[6] P.M. Glibert, J.M. Burkholder, The Complex Relationships Between Increases in Fertilization of the Earth, Coastal Eutrophication and Proliferation of Harmful Algal Blooms, Ecology of Harmful Algae, Springer, Berlin, Heidelberg, 2006, pp. 341–354.

[7] S. Ding, M. Chen, M. Gong, X. Fan, B. Qin, H. Xu, S. Gao, Z. Jin, D.C.W. Tsang, C. Zhang, Internal phosphorus loading from sediments causes seasonal nitrogen limitation for harmful algal blooms, Sci. Total Environ., 625 (2018) 872–884.

[8] K.J. Flynn, A. Mitra, Building the "perfect beast": modelling mixotrophic plankton, J. Plankton Res., 31 (2009) 965–992.

[9] P.M. Glibert, J.I. Allen, A.F. Bouwman, C.W. Brown, K.J. Flynn, A.J. Lewitus, C.J. Madden, Modeling of HABs and eutrophication: status, advances, challenges, J. Mar. Syst., 83 (2010) 262–275.

[10] S. Bae, D. Seo, Analysis and modeling of algal blooms in the Nakdong River, Korea, Ecol. Modell., 372 (2018) 53–63.

[11] Eutrophication of Waters: Monitoring, Assessment and Control, Organisation for Economic Co-operation and Development, OECD Publications and Information Center, Washington, 1982.

[12] M.V. Hoyer, J.R. Jones, Factors affecting the relation between phosphorus and chlorophyll a in Midwestern reservoirs, Can. J. Fish. Aquat. Sci., 40 (1983) 192–199.

[13] V.H. Smith, G.D. Tilman, J.C. Nekola, Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems, Environ. Pollut., 100 (1999) 179–196.

[14] W.R. Hill, S.E. Fanta, B.J. Roberts, Quantifying phosphorus and light effects in stream algae, Limnol. Oceanogr., 54 (2009) 368–380.

[15] D.E. Canfield Jr., Prediction of chlorophyll a concentrations in Florida lakes: the importance of phosphorus and nitrogen, J. Am. Water Resour. Assoc., 19 (1983) 255–262.

[16] V.L.M. Huszar, N.F. Caraco, F. Roland, J. Cole, Nutrient–chlorophyll relationships in tropical–subtropical lakes: do temperate models fit?, Biogeochemistry, 79 (2006) 239–250.

[17] G. Phillips, O.-P. Pietiläinen, L. Carvalho, A. Solimini, A. Lychees Solheim, A.C. Cardoso, Chlorophyll–nutrient relationships of different lake types using a large European dataset, Aquat. Ecol., 42 (2008) 213–226.

[18] Y.T. Prairie, C.M. Duarte, J. Kalff, Unifying nutrient–chlorophyll relationships in lakes, Can. J. Fish. Aquat. Sci., 46 (1989) 1176–1182.

[19] G. Borics, L. Nagy, S. Miron, I. Grigorszky, Z. László-Nagy, B.A. Lukács, L. G-Tóth, G. Várbíró, Which factors affect phytoplankton biomass in shallow eutrophic lakes?, Hydrobiologia, 714 (2013) 93–104.

[20] K.-S. Jeong, G.-J. Joo, H.-W. Kim, K. Ha, F. Recknagel, Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network, Ecol. Modell., 146 (2001) 115–129.

[21] D.F. Millie, G.R. Weckman, W.A. Young II, J.E. Ivey, H.J. Carrick, G.L. Fahnenstiel, Modeling microalgal abundance with artificial neural networks: demonstration of a heuristic 'Grey-Box' to deconvolve and quantify environmental influencesk, Environ. Model. Software, 38 (2012) 27–39.

[22] A.N. Blauw, P. Anderson, M. Estrada, M. Johansen, J. Laanemets, L. Paperzak, D. Purdie, R. Raine, E. Vahtera, The use of fuzzy logic for data analysis and modelling of European harmful algal blooms: results of the HABES project, Afr. J. Mar. Sci., 28 (2006) 365–369.

[23] B. Paudel, D. Velinsky, T. Belton, H. Pang, Spatial variability of estuarine environmental drivers and response by phytoplankton: a multivariate modeling approach, Ecol. Inf., 34 (2016) 1–12.

[24] S. Chen, S.A. Billings, Neural networks for nonlinear dynamic system modelling and identification, Int. J. Control, 56 (1992) 319–346.

[25] G. Gal, M. Skerjanec, N. Atanasova, Fluctuations in water level and the dynamics of zooplankton: a data-driven modelling approach, Freshwater Biol., 58 (2013) 800–816.

[26] M. Kuhn, K. Johnson, Applied Predictive Modeling, Vol. 26, Springer, New York, 2013.

[27] D.P. Solomatine, K.N. Dulal, Model trees as an alternative to neural networks in rainfall-runoff modeling, Hydrol. Sci. J., 48 (2003) 399–411.

[28] B. Bhattacharya, D.P. Solomatine, Neural networks and M5 model trees in modelling water level–discharge relationship, Neurocomputing, 63 (2005) 381–396.

[29] S. Schnier, X. Cai, Prediction of regional streamflow frequency using model tree ensembles, J. Hydrol., 517 (2014) 298–309.

[30] S. Heddam, O. Kisi, Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree, J. Hydrol., 229 (2018) 499–509.

[31] S.J. Moe, S. Haande, R.-M. Couture, Climate change, cyanobacteria blooms and ecological status of lakes: a Bayesian network approach, Ecol. Modell., 337 (2016) 330–347.

[32] C.M. Mutshinda, Z.V. Finkel, A.J. Irwin, Which environmental factors control phytoplankton populations? A Bayesian variable selection approach, Ecol. Modell., 269 (2013) 1–8.

[33] Y. Park, Y.A. Pachepsky, K.H. Cho, D.J. Jeon, J.H. Kim, Stressor–response modeling using the 2D water quality model and regression trees to predict *chlorophyll-a* in a reservoir system, J. Hydrol., 529 (2015) 805–815.

[34] S. Han, E. Kim, S. Kim, The water quality management in the Nakdong River watershed using multivariate statistical techniques, KSCE J. Civil Eng., 13 (2009) 97–105.

[35] H.-W. Kim, S.-J. Hwang, K.-H. Chang, M.-H. Jang, G.-J. Joo, N. Walz, Longitudinal difference in Zooplankton grazing on phyto- and bacterioplankton in the Nakdong River (Korea), Int. Rev. Hydrobiol., 87 (2002) 281–293.

[36] Z. Yang, P. Xu, D. Liu, J. Ma, D. Ji, Y. Cui, Hydrodynamic mechanisms underlying periodic algal blooms in the tributary bay of a subtropical reservoir, Ecol. Eng., 120 (2018) 6–13.

[37] M.J.A. Berry, G.S. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, Wiley, USA, 1997.

[38] T.S. Lim, W.Y. Loh, Y.S. Shin, An Empirical Comparison of Decision Trees and Other Classification Methods, Technical Report 979, Department of Statistics, UW Madison, 1997.

[39] G.V. Kass, An exploratory technique for investigating large quantities of categorical data, Appl. Stat., 29 (1980) 119–127.

[40] J. Magidson, The use of the new ordinal algorithm in CHAID to target profitable segments, J. Database Mark., 1 (1993) 29–48.

[41] W.Y. Loh, Fifty years of classification and regression trees, Int. Stat. Rev., 82 (2014) 329–348.

[42] S. Kim, S. Kim, Spatial water quality analysis of main stream of Nakdong River considering the inflow of tributaries, J. Korean Soc. Water Environ., 33 (2017) 640–649.

[43] J. Köhler, Origin and succession of phytoplankton in a river-lake system (Spree, Germany), Hydrobiologia, 289 (1994) 73–83.

[44] D.J. McQueen, D.R.S. Lean, Influence of water temperature and nitrogen to phosphorus ratios on the dominance of blue-green algae in Lake St. George, Ontario, Can. J. Fish. Aquat. Syst., 44 (1987) 598–604.

[45] R.E. Hecky, P. Kilham, Nutrient limitation of phytoplankton in freshwater and marine environments: a review of recent evidence on the effects of enrichment, Limnol. Oceanogr., 33 (1988) 796–822.

[46] D.J. Conley, H.W. Paerl, R.W. Howarth, D.F. Boesch, S.P. Seitzinger, K.E. Havens, C. Lancelot, G.E. Likens, Policy Forum Ecology/Controlling eutrophication: nitrogen and phosphorus, Science, 323 (2009) 1014–1015.