Proceedings of the 2nd International Conference on Biological Engineering and Medical Science (ICBioMed 2022), Part I
DOI: 10.54254/2753-8818/3/20220489

Statistic Study of Relation between the Algal Blooms and Chemicals in Water Body

Qiu Yu

Department of Statistics, School of Mathematics and Statistics, Shandong University, Weihai, 264209, China

202000810084@mail.sdu.edu.cn

Abstract. The global trend of the algal blooms is rising fast without any questions. This paper is to investigate whether we can predict the number of the algae in water body by the content of some common chemicals to avoid the algal blooms and, at the same time, use less time and costs. Step regression will be used to filter the independent variables, then use principal component analysis to get rid of multilinearity in the model and use ordinary least squares regression to fit the linear relationship between the principal components chosen and the number of algae. The results show that there is some relationship between the variables studied but not very strong. The R-squared values calculated are 2.86 for the row data and 3.58 for the test data. And the major influential factors are mxPH, mnO2 and PO4. In conclusion, we can use the model to briefly predict the trend of the algal blooms but cannot get the number accurately. However, the results do reveal to us that maximum pH value, minimum value of oxygen and average of orthophosphate indeed have some significant impacts on this kind of algae.

Keyword: algal blooms, water body, chemicals, OLS, stepwise regression, principal component analysis, linear relationship, multilinearity

1. Introduction

1.1. Background

Nowadays, with the development of technology and industry, human's demand for energy is consistently increasing and the environmental quality deterioration and ecological imbalance which is caused by human activities are becoming more serious.

Among the top 10 environmental problems of the world, water pollution has always been a focus. Water body pollution means the number of the harmful substances that enter the water body is so huge that the condition is beyond the control of a water-based ecosystem. The entire ecosystem loses its ability to regulate and then starts with some positive feedback adjustments, which causes its value to decrease or even completely lose.

The pollutants might come from the nature world and industrial emissions or they might come from biological emission. If the water body contains a great amount of phosphorus or nitrogen, then the algae, such as blue-green algae and red algae, will bloom in a short time. The life process of algae growth and multiplication creates mare wastes, and that causes the water body to be further polluted and the algae to be more increased. This whole process is what we call 'Eutrophication'.

Without doubt, the climate changes have an essential influence on the bloom of algae. According to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/3/20220489

a report from EWG, the number of the toxic algae, along with the flood of it, has been rapidly increasing since 2010.The researchers from the University of Southern California in the United States also indicated that the global trend of the algal blooms is rising fast without any question. Meanwhile, the pollution resulted from the algae bloom will not only imbalance the whole ecosystem, but also have some negative effects on human health. Anyway, the problem of eutrophication of water body is imminent.

After reading some former research and studies, some conclusions about what factors are related to the algae bloom are drawn. There are some environment-related factors, like precipitation, wind speed, temperature, water flow rate, and the degree of turbidity, biology-related factors, such as the number of cyanobacteria and the content of organism like chlorophyll, and also some chemistry-related factors——the content of dissolved oxygen and nitrate.

Surely, there are some mature techniques that are precisely aimed at monitoring and predicting all kinds of factors mentioned above now, such as some bioanalysis techniques. However, these techniques show high requests on the experiment device and experimenters. Additionally, these high requests lead to expensive experimental costs and longer time for waiting for the experiment results. Although what this kind of results perform is more accurate, sometimes we need to obtain the conclusion quickly and take corresponding measures immediately to prevent the algae from excessive growth. To get the results in a short time, we can start with the chemical composition that is common and easy to detect in water body. For instance, chlorine and oxygen can be taken into consideration. In that case, the requests for the experimenters are decreased largely, the experimental fees are much cheaper and, at the same time, the results can be obtained and put into use easier and earlier.

On the other hand, according to some former research and studies, most researchers choose to focus on one particular factor or a combination of factors from different aspects. Seldom do they use homogeneous factors, such as factors all related to chemistry, so it is meaningful to do some research about it. Besides, it is easier to find the inner connection of homogeneous things and the character of the dimension is basically the same, so there is no need to make complex transformations between them. Therefore, the next step is to get some statistics about some common chemical compositions in water body.

Some regression methods are going to be used in this paper to discover the linear connection between the content of chemical compositions and the number of algae in water body samples and to solve the question of whether this can provide some good prediction on the basis of saving experimental costs and reducing detecting time to a large extent.

1.2. Literature review

Donald Scavia, Joseph V. Depinto and Isabella Bertani conducted research regarding a multi-model approach to evaluating target phosphorus (P) loads for Lake Erie. This article details the process of assessing existing targets and phosphorus (P) load targets for Lake Erie, provides the information needed to update these targets, summarizes some of the requirements for avoiding cyanobacterial blooms in the severe Western Basin, reducing hypoxia in the Central Basin, and meeting the Cladophora threshold, and gives some recommendations [1].

Amin Daghighi conducted research regarding harmful algae bloom (HAB) prediction model for western Lake Erie. The model uses stepwise multiple regression (SMS) and genetic programming (GP). In the article, all possible predictors and predictand variables collected from various data source are taken into consideration and then eight predictors and based on a spearman rank correlation coefficient between predictors and predictand variables, one predictand are selected [2]. The process is associated with SMR and GP. The results show that both SMR and GP can be used to simulate the past occurrence of HAB events and predict future HAB intensity.

Zheng-Xi Zhou et al. conducted research about the complex relationship between environmental factors and harmful algal blooms in water body close to the estuary of the Yangtze River. The research uses principal component analysis, multiple regression analysis and path analysis [3]. The results suggest different major factors, like Changjiang diluted water and Taiwan warm water, are contributed

DOI: 10.54254/2753-8818/3/20220489

to the dinoflagellate bloom and the diatom bloom.

Inchio Lou et al. conduct regarding combining support vector regression with particle swarm optimization for the model of freshwater algal blooms. This research proposes a hybrid intelligent model using the two methods mentioned above to yield best control of parameters that predict and forecast the phytoplankton dynamics and uses the model to reduce the algal bloom problems in a reservoir [4]. They used part of the data for training and the rest of it for testing and got some good results.

Hongwon Yun conducted research regarding the prediction model of algal blooms based on logistic regression and confusion matrix. The algal bloom dataset was analyzed by logistic regression analysis, and statistically tested and normalized to identify marine environmental factors that influence algal blooms [5]. And apply the actual values and the predicted values to the confusion matrix and improve the accuracy the prediction through improving the decision boundary of the existing logistic regression.

Sangmok Lee and Donghyun Lee conducted research regarding the improved prediction of harmful algal blooms in Korea. The research uses some deep learning models. The research conducted an algal bloom prediction model using regression analysis and some deep learning methods including multilayer perceptron, recurrent neural network and long short-term memory [6]. The model can predict chlorophyII- a, a recognized criteria for algal activity. The results show high prediction efficiency.

Mohamed Laabir et al. conducted research regarding the prediction of two kinds of algal blooms. The research uses some methods including decision trees and Zero Inflated Negative Binomial linear regression [7]. The research analyzed a 10-year dataset and used the methods mentioned above to reveal when will Dinophysis bloom events and shellfish contaminations most likely occur and to find out if salinity and temperature affect the occurrence and blooms of Alexandrium.

Jingjing Xia and Jin Zeng conducted research about the early warning of algal blooms in a famous bay in China. The research is based on the optimization support vector machine regression, traditional particle swarm optimization, Artificial Neural Network Model and grey wolf optimizer [8]. The research uses environmental factors to assist the prediction of Chl-a based on the method mentioned above, The final model achieved high accuracy.

Qiuwen Chen and Arthur E Mynett conducted research about predicting phaeocystis globsa bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression. The research uses these two methods to predict bloom timing and bloom intensity and also set up a multi-variable regression-model to predict bloom duration [9]. A variety of factors that affect the bloom have been found and tested. It turns out that the methods used are quite good techniques in modelling harmful algal blooms.

K. -W. Chang et al. conducted research about predicting algal bloom in the Techi reservoir using Landsat TM data. The research developed an empirical model with some independent variables matched the dinoflagellate algal cell densities to detect algal blooms phenomenon in the Techi reservoir using techniques like a forward selection method and showed some prediction results and the limitations of the constructed model [10].

2. Body of work

2.1. Data description

The data set used in this paper is downloaded from Kaggle, which consists the content of 8 different kinds of chemical compositions including maximum pH value, minimum value of oxygen and average value of chloride, nitrates, ammonium, orthophosphate, phosphate and chlorophyll and the quantity of 7 different kinds of algae in some samples from European rivers. For simplicity, the data set uses frequency numbers found in the respective water samples to represent the number of the algae. Most of the sample data for each indicator is concentrated in a rough range, which is considered a good start for the study.

2.2. Implementation flow

First, use 8 independent variables to construct an OLS model with each one of the 7 different kinds of algae's quantity data and choose the one that has the largest value of R-squared to continue.

DOI: 10.54254/2753-8818/3/20220489

Second, analyze the F-test and the t-test results. The results show that the primary model has some independent variable that is not significant to it. Based on this situation, using the idea of stepwise regression and, at the same time, taking AIC as an evaluation criterion to filter the independent variables are good choices.

After that, 5 variables are chosen to construct a new model. It is found that the problem that some variables are not significant is solved but the condition number of the rest part of the data matrix is still large and that indicates that there is still multilinearity existing in the model.

As a result, the next step is to use principal component analysis. After doing so, three principal components are chosen to construct the final model and do some prediction with 122 test data.

However, the outcome of the prediction is not so good. Logistic regression is taken into consideration to construct a new model and raise the accuracy of the prediction.

2.3. Methods

R-squared：

R-squared is a kind of statistic measure, which represents how much the independent variables chosen can explain the change of the dependent variable in a regression model. For example, if the value of the R-squared is 0.9, that means 90 percent of the variance of the dependent variable being studied can be explained by the variance of the independent variables. Its formula is:

$$R^2 = 1 - \frac{SSE}{SST} \tag{1}$$

SSE represents the error sum of squares. SST represents the total sum of squares.

$$SSE = \sum\left(Y_t - \widehat{Y_t}\right)^2 \tag{2}$$

where Y_t stands for the true value of the dependent variable and $\widehat{Y_t}$ stands for the predictive value of the dependent variable.

$$SST = \sum\left(Y_t - \underline{Y}\right)^2 \tag{3}$$

where $\underline{Y}$ stands for the mean of the dependent variable.

Adj. R-squared:

The formula is:

$$R^2_{adj.} = 1 - \left[\frac{\left(1-R^2\right)(n-1)}{n-k-1}\right] \tag{4}$$

n is the number of points in the data sample; k is the number of independent regressors.

The value of R-squared will rise, when the number of independent variables increases. This can cause an overfitting condition. The Adj. R-squared can be used to eliminate this situation and to have a better comparison between models which have the same dependent variable but different independent variables.

F-test:

After finishing the construction of a linear model, it is necessary to test the significance of it. Suppose the null hypothesis (H0) be:

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0 \tag{5}$$

where β_j stands for the coefficient before independent variable x_2.

Constructing the F statistic

$$F = \frac{\frac{SSR}{k-1}}{\frac{SSE}{n-k}} \tag{6}$$

n is the number of points in the data sample; k is the number of independent regressors.

DOI: 10.54254/2753-8818/3/20220489

$$SSR = \sum(\widehat{Y_t} - \underline{Y})^2 \quad (7)$$

where $\widehat{Y_t}$ stands for the predictive value of the dependent variable and $\underline{Y}$ stands for the mean of the dependent variable.

The F statistic follows a F distribution with (k-1) and (n-k) degrees of freedom. Then calculate the value of the F statistic, and decide whether to reject the null hypothesis depending on the theoretical critical value.

t-test:

Different from F-test, t-test is used to test whether every single independent variable in the linear model is significant for the dependent variable. Suppose the null hypothesis (H0) be:

$$H_0: \beta_j = 0 \quad (8)$$

Constructing the t statistic

$$t_{\widehat{\beta_J}} = \frac{\widehat{\beta_J}}{S_{\widehat{\beta_J}}} \quad (9)$$

where $\widehat{\beta_J}$ stands for the estimated value of the regression coefficient and $S_{\widehat{\beta_J}}$ stands for $\widehat{\beta_J}$'s standard deviation estimate.

The t statistic follows a t distribution. Then calculate the value of the t statistic, and decide whether to reject the null hypothesis depending on the theoretical critical value.

AIC:

Formula:

$$AIC = 2k - 2\ln(L) \quad (10)$$

where k stands for the number of parameters and L stands for the likelihood function.

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. AIC encourages the goodness of data fitting but also tries to avoid overfitting. The model with the smallest AIC value should be given priority.

BIC:

Formula:

$$BIC = kln(n) - 2\ln(L) \quad (11)$$

where k stands for the number of parameters, n stands for the number of samples and L stands for the likelihood function.

The Bayesian information criterion (BIC) functions in the same way as the AIC. BIC is used to avoid overfitting when the increase of the number of the parameters causes the increase in likelihood function. Compared with AIC, the penalty term in the BIC formula increases significantly as the number of parameters increases, which causes BIC to prefer a simpler model with fewer parameters included. So naturally, the model with the smallest BIC value also tends to be chosen first.

Ordinary least squares:

Ordinary least squares method is widely used to get the linear fit function of variables. Its main idea is to solve a conditional extremum equation to make the sum of squares error of the fit function the least.

Stepwise regression:

Stepwise regression is a technique that is widely used when selecting proper dependent variables to construct a model.

The basic ideas of this method are as follow:

First of all, add the variable into the model one by one and test the model (Any reasonable method can be used to test the model). The variable will be fixed in the model provided its influence on the model is significant by test. After fixing a new variable in the model, the old variables will be tested one by one to see if they are significant to the model. If not, they will be kicked off the model. Cycling

DOI: 10.54254/2753-8818/3/20220489

through the process above, until there is no variable to be added or kicked off.

By using this method, the multilinearity in the model will be reduced.

Principle component regression:

Principle component regression is also used to eliminate the multilinearity. The basic idea of this method is now that there is multilinearity between existing variables, we might as well use the linear combination of the variables as new independent variables to replace them.

The process is as follow:

First off, standardize the data

$$x_i^* = \frac{x_i - \mu_i}{\sigma_i} \tag{12}$$

where σ_i is the standard deviation of the variable x_i and μ_i is the mean of the variable x_i.

Second, solve the conditional extremum equation

$$con: \sum_{i=1}^{n} \quad c_i^2 = 1 \tag{13}$$

$$obj: \max[var(\sum_{i=1}^{n} \quad c_i x_i^*)] \tag{14}$$

What should be explained is because newly constructed variables are expected to have a big deviation so that it can contain enough information, the condition related to the combination coefficients above should be set to ensure the function can be solved.

The solution process uses knowledge of Lagrange multiplier and matrix derivation. The results show that each group of coefficients corresponds to an eigenvalue of the standardized data matrix. Moreover, the larger the value of the eigenvalue is, the larger the deviation of the corresponding new variable is, or in other words, the higher the contribution rate is. In that case, we can use the eigenvalues to define the contribution rate:

$$w_j = \frac{\lambda_j}{\sum_{i=1}^{n} \quad \lambda_i} \tag{15}$$

Where λ_j stands for corresponding eigenvalue.

After all of these, generally speaking, the contribution rate is arranged from large to small, and several principal components (new variables) are selected to make the cumulative contribution rate reach 85%. Then these chosen new variables can be used in further analysis and regression.

Logistic regression:

Logistic models (also known as logit models) are some kind of classification and predictive method which are commonly used in data analysis. Logistic regression estimates the probability that an event will occur based on the data. Since the outcomes are probabilities, the dependent variable's value range is between 0 and 1.

2.4. Results

After comparing 8 OLS model results, choose the one that has the largest value of R-squared.

DOI: 10.54254/2753-8818/3/20220489

OLS Regression Results

Dep. Variable:	y	R-squared:	0.387
Model:	OLS	Adj. R-squared:	0.359
Method:	Least Squares	F-statistic:	13.79
Date:	Fri, 26 Aug 2022	Prob (F-statistic):	1.93e-15
Time:	16:08:19	Log-Likelihood:	-469.78
No. Observations:	184	AIC:	957.6
Df Residuals:	175	BIC:	986.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	16.8336	4.888	3.444	0.001	7.187	26.480
x1	-1.6358	0.590	-2.774	0.006	-2.800	-0.472
x2	-0.2520	0.118	-2.135	0.034	-0.485	-0.019
x3	0.0046	0.006	0.777	0.438	-0.007	0.016
x4	-0.4490	0.099	-4.517	0.000	-0.645	-0.253
x5	0.0010	0.000	5.349	0.000	0.001	0.001
x6	-0.0053	0.007	-0.792	0.429	-0.019	0.008
x7	0.0137	0.005	2.637	0.009	0.003	0.024
x8	-0.0213	0.014	-1.478	0.141	-0.050	0.007

Omnibus:	170.415	Durbin-Watson:	1.629
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6181.353
Skew:	3.145	Prob(JB):	0.00
Kurtosis:	30.689	Cond. No.	4.39e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
[2] The condition number is large, 4.39e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 1. The results of preliminary OLS regression.

Based on figure 1, the R-squared value is 0.387 and that means 38.7 percent of the variance of the dependent variable being studied can be explained by the variance of the independent variables.

The P value of the F-test is very small, so it is safe to say at least one indicator is significant to the model.

Then let's turn to the t-test results. There are three indicator's P values of the t-test are larger than 0.05 so it is necessary to filter the indicators.

AIC is chosen as a criterion to operate a stepwise regression method. Here, a combination of forward and bidirectional regression methods will be used. (Forward regression and bidirectional regression are kind of the same as stepwise regression. Forward regression only add variable into the model while bidirectional regression kick variable off after add variable)

Here are the results:

DOI: 10.54254/2753-8818/3/20220489

```
(['mxPH', 'mnO2', 'NO3', 'NH4', 'PO4'],
 <class 'statsmodels.iolib.summary.Summary'>
 """
                            OLS Regression Results
==============================================================================
Dep. Variable:                     a4   R-squared:                       0.376
Model:                            OLS   Adj. R-squared:                  0.358
Method:                 Least Squares   F-statistic:                     21.43
Date:                Thu, 25 Aug 2022   Prob (F-statistic):           9.28e-17
Time:                        18:59:14   Log-Likelihood:                -471.40
No. Observations:                 184   AIC:                             954.8
Df Residuals:                     178   BIC:                             974.1
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         19.8606      4.347      4.568      0.000      11.282      28.440
mxPH          -1.9901      0.513     -3.879      0.000      -3.002      -0.978
mnO2          -0.2704      0.116     -2.335      0.021      -0.499      -0.042
NO3           -0.4397      0.093     -4.743      0.000      -0.623      -0.257
NH4            0.0009      0.000      5.294      0.000       0.001       0.001
PO4            0.0102      0.002      4.771      0.000       0.006       0.014
==============================================================================
Omnibus:                      170.450   Durbin-Watson:                   1.656
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5822.359
Skew:                           3.173   Prob(JB):                         0.00
Kurtosis:                      29.817   Cond. No.                     3.90e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
[2] The condition number is large, 3.9e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 2. OLS regression results after filtering the indicators.

According to figure 2, we use these five indicators to construct a new model. The Adj. R-squared value is nearly the same as the former one. That suggests the decrease of the R-squared value is only caused by the reduction of the indicators and each indicator is significant to the model. However, the notes remind us that the condition number is large and that means there is still multilinearity existing in the model.

As a result, let's do principal component analysis on these five indicators. Each components' contribution rate and coefficient matrix are as follow:

[0.37118467 0.29942638 0.17713968 0.10375083 0.04849845]
[[-0.19717818 -0.11791695 0.64128483 0.66207061 0.31246221]
[-0.3629598 0.67351246 0.2180702 0.08382575 -0.60004982]
[0.9047645 0.31595106 0.22815159 0.14684548 -0.08921538]
[0.10359844 -0.60684794 -0.05532186 0.31604653 -0.71976178]
[-0.00703657 0.25374608 -0.69719832 0.65816923 0.12763689]]

Then, we use the first three components to construct an OLS model again.

DOI: 10.54254/2753-8818/3/20220489

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.286
Model:	OLS	**Adj. R-squared:**	0.274
Method:	Least Squares	**F-statistic:**	23.98
Date:	Sat, 27 Aug 2022	**Prob (F-statistic):**	4.18e-13
Time:	14:49:27	**Log-Likelihood:**	-230.15
No. Observations:	184	**AIC:**	468.3
Df Residuals:	180	**BIC:**	481.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.776e-17	0.063	-4.41e-16	1.000	-0.124	0.124
x1	0.2055	0.046	4.443	0.000	0.114	0.297
x2	-0.2768	0.051	-5.376	0.000	-0.378	-0.175
x3	-0.3232	0.067	-4.828	0.000	-0.455	-0.191

Omnibus:	242.550	**Durbin-Watson:**	1.765
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19847.344
Skew:	5.321	**Prob(JB):**	0.00
Kurtosis:	52.755	**Cond. No.**	1.45

Notes:

[1] Standard Errors assume that the covariance matrix of the errors

Figure 3. Principal component regression results.

According to figure 3, the notes before don't appear, but correspondingly, the R-squared value is decreased. Simplify the model, we get these linear function:

$$y = -1.96165604\,x_1 - 0.51719040x_2 - 0.00237955x_3 + 0.00012812x_4 + 0.00798917x_5 + 21.12493410 \quad (16)$$

Use test data to predict and calculate the R-squared value.

$$R^2 = 0.35848435812353496 \quad (17)$$

3. Conclusion

After getting rid of the unrelated variables and the multilinearity in the model we do find that there is some linear relationship between the algal blooms and chemicals in water body but the relationship is

DOI: 10.54254/2753-8818/3/20220489

not so strong. Even so, we still can use the model constructed to estimate the number of algal in the water body.

According to the biological common sense, while other conditions are appropriate, In the case of other conditions, a certain condition will only achieve the best overall state within a certain range, which is obviously not in line with the linear relationship, so the method of changing the power of independent variables can be considered. Fit the dependent variable and the squares of the original independent variable to do the least squares linear regression and the R-squared value of the model constructed is 0.489 which is much better than before.

Sometimes you can discard some information to improve the accuracy of the model. The number of algae can be divided into two parts, using a Logistic regression model. With proper demarcation, it is also possible to predict the occurrence of red tides before the algae population reaches K/2. The preliminary model constructed predicts well when the number of algae is less than 5.

Also, maybe there are different linear relationships at different intervals. Therefore, piecewise linear regression and regression tree can be used.

References

[1] Scavia, Donald, Joseph V. DePinto, and Isabella Bertani. "A multi-model approach to evaluating target phosphorus loads for Lake Erie." *Journal of Great Lakes Research* **42**.6 (2016): 1139-1150.

[2] Daghighi, Amin. "Harmful algae bloom prediction model for western lake erie using stepwise multiple regression and genetic programming." (2017).

[3] Zhou, Zheng-Xi, Ren-Cheng Yu, and Ming-Jiang Zhou. "Resolving the complex relationship between harmful algal blooms and environmental factors in the coastal waters adjacent to the Changjiang River estuary." *Harmful Algae* **62** (2017): 60-72.

[4] Lou, Inchio, et al. "Integrating support vector regression with particle swarm optimization for numerical modeling for algal blooms of freshwater." Advances in Monitoring and Modelling Algal Blooms in Freshwater Reservoirs. *Springer, Dordrecht*, 2017. 125-141.

[5] Yun, Hongwon. "Prediction model of algal blooms using logistic regression and confusion matrix." *International Journal of Electrical and Computer Engineering* **11**.3 (2021): 2407.

[6] Lee, Sangmok, and Donghyun Lee. "Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models." *International Journal of Environmental Research and Public Health* **15**.7 (2018): 1322.

[7] Bouquet, Aurélien, et al. "Prediction of Alexandrium and Dinophysis algal blooms and shellfish contamination in French Mediterranean Lagoons using decision trees and linear regression: a result of 10 years of sanitary monitoring." **Harmful Algae 115** (2022): 102234.

[8] Xia, Jingjing, and Jin Zeng. "Early warning of algal blooms based on the optimization support vector machine regression in a typical tributary bay of the Three Gorges Reservoir, China." *Environmental Geochemistry and Health* (2022): 1-15.

[9] Chen, Qiuwen, and Arthur E. Mynett. "Predicting Phaeocystis globosa bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression." *Ecological Modelling* **176**.3-4 (2004): 277-290.

[10] Chang, K-W., Y. Shen, and P-C. Chen. "Predicting algal bloom in the Techi reservoir using Landsat TM data." *International Journal of Remote Sensing* **25**.17 (2004): 3411-3422.