ABSTRACT
Due to many environmental and biology factors influenced the phytoplankton biomass, there are multiple predictor variables often themselves related in cases of severe multicolinearity. The paper introduced a new iterative dimensionality reduction data mining method based on partial least-squares regression to study the control mechanism of phytoplankton biomass. Results showed that control mechanism changed from bottom-up in winter to top-down in summer. There were always negative relations between phytoplankton and phosphate, and phytoplankton was negative correlated with NH$_4^+$ and NO$_3^-$ but positive correlated with NO$_2^-$ in two seasons.

Keywords: phytoplankton biomass; partial least squares regression; control mechanism

1. INTRODUCTION
Phytoplanktons are the autotrophic component of the plankton community and responsible for much of the oxygen present in the Earth's atmosphere. Marine phytoplankton plays an important role in the recycling of carbon and nutrients in the pelagic zone. So, it's one of the central issues of aquatic microbial ecology to understand the factors that regulate phytoplankton productivity and biomass [1].

Phytoplankton biomass can be controlled by supply of resources (light or nutrients) or from physical factors such as temperature of the environment (bottom-up). For example, in many coastal-zone areas, marine phytoplankton populations can increase a lot suddenly (blooms) in the spring under some suitable conditions. The bloom ceases and populations decline when they have used up all the nutrients in the water [2-3]. On the other hand, phytoplankton biomass can be controlled by grazing effect from high trophic levels (top-down). So, planktonic populations are regulated by a combination of both resource limitation from the bottom up and predation from the top down [4-5]. The influences of grazing top-down control and nutrient bottom-up controls have been deeply investigated in the classic food web by traditional statistic method [6]. However, due to many environmental and biology factors influenced the phytoplankton biomass, there are often large number of explanatory variables in comparison with the number of observations, and the multiple predictor variables often themselves related in cases of severe multicolinearity [7].

Partial least squares (PLS) regression is a recent technique that generalizes and combines features from principal component analysis and multiple regression [8]. Partial least-squares regression was first developed by Wold in the late 1960s [9] and then be used to analyze data from chemical applications [10]. Now, PLS has been found used to study the effect of natural or human disturbance to the different ecosystems, such as community ecology [11-12], large scale influence of climate [13] and environmental effects on biodiversity [14-15]. Luis M (2008) provided that PLS was more reliable than other techniques such as multiple regression and a combination of principal component analysis and multiple regression when identifying relevant variables and their magnitudes of influence, especially in cases of small sample size and low tolerance [16]. Research of Zhang-yu et al. (2007) also showed that PLS was more effective than stepwise multiple regression or regression analyses with principal component analysis in relating hyperspectral leaf reflectance in rice Oryza sativa crops to the disease severity of the fungus Bipolaris oryzae [17].

The present study focuses on the control mechanism of phytoplankton biomass under many environmental and biology factors with a new data mining method of partial least squares.
(PLS) regression.

2. MATERIALS AND METHODS

2.1. Study area

Bohai Bay is a semi-enclosed inner sea in north China with an area of about 80,000 km² and an average depth of about 18 m. This area is highly developed, and about 1 billion tons of wastewater were discharged into the bay from Beijing, Tianjin, and Hebei province every year.

In March and June 2006, surficial seawater samples were collected from 24 and 30 stations in coastal areas of Tianjin Bohai Bay respectively. The survey region covers 117° 37′ E—117° 50′ E, 38° 49′ N—38° 54′ N (Fig. 1).

2.2. Sampling and analysis

The measurement parameters include field pH value (pH), NH₄⁺, NO₃⁻, PO₄³⁻, dissolved oxygen (DO), NO₂⁻, chemical oxygen demand (COD), chl-a, phytoplankton and zooplankton biomass. Sampling, preservation and analytical protocols were conducted by standard methods [18].

Surficial seawater samples were taken at each of the stations and then filtered through a membrane filter (0.45 μ m). Phytoplankton and zooplankton were obtained from vertical tows (bottom to surface) respectively. The phytoplankton samples mainly consisted of algae of the taxonomic groups, and the samples of zooplankton mainly consisted of small copepods which are primary herbivores.

2.3. PLS regression method

The goal of PLS regression is to predict Y from X and to describe their common structure. If there were serious multicollinearity in the arguments, as well as multicollinearity between the dependent variables, the ordinary multiple regression methods could not be used. But partial least-squares regression analysis method could be used to solve the problem perfectly [19-23]. PLS regression finds components from X that are also relevant for Y. Specifically, PLS regression searches for a set of components that performs a simultaneous decomposition of X and Y with the constraint that these components explain as much as possible of the covariance between X and Y.

The algorithm of PLS regression can be summarized as follows. First there are two matrices: E = X and F = Y. These matrices are then column centered and normalized. The vector u is initialized with random values before starting the iteration process.

- Step 1. w ∝ ETu (estimate X weights).
- Step 2. t ∝ Ew (estimate X factor scores).
- Step 3. c ∝ FTt (estimate Y weights).
- Step 4. u ∝ Fc (estimate Y scores).

If t has converged, then compute the value of b which is used to predict Y from t as b = tTu, and compute the factor loadings for X as p = Eᵀt. If t has not converged, then go to Step 1. Now subtract t from both E and F. If E is a null matrix, then the whole set of latent vectors has been found, otherwise the procedure can be re-iterated from Step 1 on.

The algorithm would terminate if the regression equation has reached a satisfactory accuracy. Otherwise, the residual explanation information would be used to conduct the second round of the principal component extraction. At the end, the dependent variables are predicted using the multivariate regression formula as Y = TBCᵀ.

3. RESULTS

3.1. Control mechanism in March 2006

The t[1]/t[2] scatter plot of the former two principal components in favorites window was displayed in Fig. 2. The main role of plot was to identify specific points.

![Fig. 2 The t[1]/t[2] Scatter Plot and T² ellipse (March 2006)](image-url)
The coefficients plot of predictors to response variable phytoplankton biomass in March 2006 was showed in Fig. 3 based on PLS model standardized variables.

Fig. 3 The coefficients plot of predictors to response variable phytoplankton biomass (March 2006)

The final regression equation was:

\[
\text{Phytoplankton} = 0.79 + 0.23pH + 0.14DO + 0.15\text{NO}_3 + 0.03\text{PO}_4 \\
-0.016\text{NH}_4 - 0.18\text{P} + 0.04\text{Chla} + 0.51\text{Zooplankton}
\]

3.2. Control mechanism in June 2006

To identify specific points, we also use the t[1]/t[2] scatter plot of the former two principal components in favorites window (Fig. 4).

Fig. 4 The t[1]/t[2] Scatter Plot and T² ellipse (June 2006)

Only one sample points of 30 samples located outside the T² ellipse. It showed that the existence of one specific points. So we delete this samples and re-fitting model.

The coefficients plot of predictors to response variable phytoplankton biomass in June 2006 was showed in Fig. 5 based on PLS model standardized variables.

Fig. 5 The coefficients plot of predictors to response variable phytoplankton biomass (June 2006)

Last, we get the final regression equation:

\[
\text{Phytoplankton} = 0.73 - 0.26pH + 0.04\text{CO}_2 + 0.13\text{DO} + 0.28\text{NO}_3 - 0.08\text{NO}_2 \\
-0.05\text{NH}_4 - 0.15\text{P} + 0.32\text{Chla} - 0.32\text{Zooplankton}
\]

4. CONCLUSIONS

Using the PLS regression method, we analysed the control mechanism of phytoplankton in two different season in 2006. In March 2006, the phytoplankton biomass had a positive relation with the zooplankton biomass. But in June 2006, there were a negative relation between phytoplankton and zooplankton. The reason for this difference may be the control mechanism changed from bottom-up to top-down.

For the nutrient-phytoplankton relations, there were always negative relations between phytoplankton and phosphate. To the nitrogen, phytoplankton was negative correlated with \(\text{NH}_4\) and \(\text{NO}_3\) but positive correlated with \(\text{NO}_2\).

ACKNOWLEDGEMENT

This research was supported by technology development and research funds provided by Ministry of Science and Technology of China (2008EG124218).

REFERENCES


