Environmental Pollution 223 (2017) 676–684

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Comparison of models for predicting the changes in phytoplankton community composition in the receiving water system of an inter-basin water transfer project☆

Qinghui Zeng [a, b], Yi Liu [c], Hongtao Zhao [a], Mingdong Sun [a], Xuyong Li [a, *]

[a] State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] The Center of Space Surveying and Mapping in China, Beijing 102102, China

ARTICLE INFO

Article history:
Received 29 September 2016
Received in revised form
3 January 2017
Accepted 1 February 2017
Available online 10 February 2017

Keywords:
Random forest
Support vector machine
Artificial neural network
Phytoplankton community
Water transfer

ABSTRACT

Inter-basin water transfer projects might cause complex hydro-chemical and biological variation in the receiving aquatic ecosystems. Whether machine learning models can be used to predict changes in phytoplankton community composition caused by water transfer projects have rarely been studied. In the present study, we used machine learning models to predict the total algal cell densities and changes in phytoplankton community composition in Miyun reservoir caused by the middle route of the South-to-North Water Transfer Project (SNWTP). The model performances of four machine learning models, including regression trees (RT), random forest (RF), support vector machine (SVM), and artificial neural network (ANN) were evaluated and the best model was selected for further prediction. The results showed that the predictive accuracies (Pearson's correlation coefficient) of the models were RF (0.974), ANN (0.951), SVM (0.860), and RT (0.817) in the training step and RF (0.806), ANN (0.734), SVM (0.730), and RT (0.692) in the testing step. Therefore, the RF model was the best method for estimating total algal cell densities. Furthermore, the predicted accuracies of the RF model for dominant phytoplankton phyla (Cyanophyta, Chlorophyta, and Bacillariophyta) in Miyun reservoir ranged from 0.824 to 0.869 in the testing step. The predicted proportions with water transfer of the different phytoplankton phyla ranged from −8.88% to 9.93%, and the predicted dominant phyla with water transfer in each season remained unchanged compared to the phytoplankton succession without water transfer. The results of the present study provide a useful tool for predicting the changes in phytoplankton community caused by water transfer. The method is transferrable to other locations via establishment of models with relevant data to a particular area. Our findings help better understanding the possible changes in aquatic ecosystems influenced by inter-basin water transfer.

1. Introduction

Changes in the phytoplankton community in large temperate freshwater lakes have been regarded as a good indicator of water quality, ecological health, and trophic status of the system (Jiang et al., 2014; Xu et al., 2001). Any inter-basin water transfer project causes complex hydro-chemical and biological variation to the receiving water system (Zeng et al., 2015). Nevertheless, changes in the phytoplankton community composition caused by inter-basin water transfer projects have rarely been studied. As such, monitoring and early prediction of phytoplankton community compositions are particularly important, especially in the receiving water system of an inter-basin water transfer project.

Machine learning models are a promising approach to examine complicated, non-linear ecological data. Artificial neural network (ANN) technologies have been widely used for the prediction of algal blooms and the succession of dominant species (Kuo et al., 2007; Lee et al., 2003; Recknagel, 1997; Wei et al., 2001; Yabunaka et al., 1997). However, ANN are based on empirical risk

☆ This paper has been recommended for acceptance by Dr. Harmon Sarah Michele.

* Corresponding author. Chinese Academy of Sciences, State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Shuangqing Road 18, Beijing 100085, China.
E-mail address: xyli@rcees.ac.cn (X. Li).

http://dx.doi.org/10.1016/j.envpol.2017.02.001
0269-7491/

minimization, which can cause the solution to be captured in a local minimum and the network overfitted (Yoon et al., 2011). Support vector machine (SVM) technology is based on the structure risk minimization, which is a method for overcoming dimensionality and over-learning problems (Wang et al., 2014). SVM has been successfully applied to a large number of environmental applications, such as water quality management, lake water level prediction, forecasting the presence of cyanotoxins, and water quality assessment (Buyukyildiz et al., 2014; Khan and Coulibaly, 2006; Liao et al., 2012; Singh et al., 2011; Vilán et al., 2013). Regression trees (RT) are ideally suited for the analysis of complex ecological data (De'ath and Fabricius, 2000), with RT increasingly prevalent in environmental sciences as a useful tool for modeling species-environment relationships (De'Ath, 2002; Park and Vlek, 2002; Sass et al., 2008; Sorrell et al., 2013). Random forest (RF) technology is a non-parametric method where a large number of decision trees are generated and each tree is grown with a randomized subset of predictors (Breiman, 2001; Cortes et al., 2013). Responses are predicted by aggregating the predictions of all trees (i.e., majority votes for classification and average for regression) (Bergström et al., 2013). By growing large numbers of tree, the generalization error is limited and the RF model is less prone to overfit the data than the single tree regression is (Prasad et al., 2006). The RF model has been well established in many disciplines (Albert et al., 2008; Chen and Liu, 2005; Ellis et al., 2014; Pal, 2005; Shi et al., 2005) and is gaining popularity with ecologists to model and understand environmental systems (Catherine et al., 2010; Kehoe et al., 2012, 2015; Prasad et al., 2006). To the best of our knowledge, despite the growing applications and success of machine learning models in surface water problems, which model is more effective in predicting total algal densities and densities of different phytoplankton groups with easily measured hydro-chemical variables is little known.

China launched the South-to-North Water Transfer Project (SNWTP) in 2002 to alleviate water shortage problems in northern China (Liu and Zheng, 2002). The middle route of SNWTP only started trials of transferring water into Miyun reservoir from October 2015. There is a difference between the hydro-chemical and biological characteristics of the sending (Danjingkou reservoir) and the receiving (Miyun reservoir) drinking water systems (Table S1, Supporting materials). The water in the Danjiangkou reservoir has relatively low pH and low total hardness (Li et al., 2009). Therefore, it is important to consider not only the concentrations of phosphorus and nitrogen but also the effect other hydro-chemical factors might have on the receiving waters when predicting changes in the phytoplankton community caused by the SNWTP. The objectives of the present study were to (1) apply machine learning models using easily measured hydro-chemical variables to predict total algal densities and densities of different phytoplankton phyla and (2) investigate the simulated changes in phytoplankton community composition caused by water transfer using a machine learning model.

2. Material and methods

2.1. Study area

Miyun reservoir (latitude 40°30′N, longitude 116°56′E) is located in the northeast section of Beijing City, and has been the main source of drinking water for Beijing since 1981. Miyun reservoir has a total watershed area of 15,788 km^2, a water surface area of 188 km^2, and the maximum water depth is approximately 40 m, with shallower water in the north and deeper water in the west and south. It is the largest reservoir in northern China, with a total storage capacity of 4375 million m^3. The primary sources of water into Miyun reservoir are the Chao River and the Bai River, while the two main outflows are the Bai River Dam and the Chao River Dam (Fig. 1).

2.2. Data collection

Water samples were collected monthly from 0.5 m below the water surface from 2009 to 2014. There were only four water quality monitoring sites (1, 2, 6 and 7) in 2009 and three new monitoring sites (3, 4 and 5) were added since 2010 (Fig. 1). Because the surface water of Miyun reservoir freezes in the winter, water sampling time is from April/May to November every year. In addition, for some locations, water samples were failed to be collected sometimes due to big stormy waves. Finally, the collected data contained 280 groups of observations. The hydro-chemical parameters measured include water temperature (T), Secchi disk depth (SD), pH, calcium (Ca), magnesium (Mg), dissolved oxygen (DO), dissolved inorganic nitrogen (DIN), total dissolved phosphorus (TDP), chloride, and sulfate. Water samples of 1.5 L were collected for phytoplankton quantitative analysis. Phytoplankton samples were fixed with Lugol's iodine solution (2% final concentration) and 1.5 L water samples were settled for 48 h. Then, the supernatant was siphoned away with a thin hose and samples were concentrated to 30 ml. Phytoplankton phyla were identified according to (Hu and Wei, 2006) and cell densities were measured using a Sedgwick-Rafter counting chamber under microscopic magnification of 200–400 × (Olympus BX51 microscope). For each sample, triplicate of 1 mL each concentrated solution (from the same bottle) was collected and counted separately. At least 400 algal cells were counted per sample to provide a counting error of <10% (Lund et al., 1958; Stanković et al., 2012; Su et al., 2014). Solar radiation data were downloaded from the China Meteorological Data Sharing Service System.

In the present study, we imputed missing values in our data with function “rfImpute” in the RF model. A total of 280 records of observations (each record consisted of eleven variables) remained to calibrate and validate the models. To eliminate the influence of the unit of each variable and the discrepancy of numerical magnitude, the data set were normalized to the range of [0, 1] (Liao et al., 2012).

2.3. Model development

2.3.1. Identification of factors affecting algal growth

Machine learning modeling may undergo over-parameterization if large numbers of input variables are included without rigorous selection (Catherine et al., 2010). The influences of environmental factors on phytoplankton community composition were identified with canonical correspondence analysis (CCA) using CANOCO 5. The relationship between phytoplankton phyla and environmental factors in the present study is shown in Fig. S1 (Supporting materials). Forward selection indicated that nine of the eleven environmental variables (except for chloride and sulfate) made independent and significant contributions to the variance in phytoplankton data ($P < 0.05$, Monte Carlo test). Results from CCA ordination indicate that the most discriminant variable was T, which explained 32.6% of the total variance followed by solar radiation (24.9%), SD (17.7%), DO (9.1%), Mg (4.3%), TDP (3.2%), pH (2.0%), Ca (2.0%), and DIN (1.6%). Based on previous modeling studies of eutrophication prediction (Çamdevýren et al., 2005; Karul et al., 2000; Kuo et al., 2007; Lee et al., 2003; Wei et al., 2001; Xu et al., 2015), T was used to simulate the effect of heat. Solar radiation provides the essential light and heat energy for primary production. SD (a measure of water transparency) is a good indicator for eutrophication status because there is a strong