Ecological Indicators 58 (2015) 225–234

Contents lists available at ScienceDirect

Ecological Indicators

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

Nonparametric approaches for estimating regional lake nutrient thresholds

Shouliang Huo, Chunzi Ma, Beidou Xi*, Zhuoshi He, Jing Su, Fengchang Wu

State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Science, Beijing 100012, China

ARTICLE INFO

Article history:
Received 5 December 2014
Received in revised form 5 March 2015
Accepted 25 May 2015

Keywords:
Nutrient threshold
Classification and regression tree
Nonparametric changepoint analysis
Bayesian hierarchical modeling
Lake ecoregion

ABSTRACT

Nonparametric approaches including a classification and regression tree (CART), a nonparametric change-point analysis (nCPA) and a Bayesian hierarchical modeling (BHM) method were developed to determine ecoregional nutrient response thresholds. A CART analysis revealed that hierarchical structure was important for predicting Chl *a* concentrations from total nitrogen (TN) and total phosphorus (TP). The nCPA and BHM methods confirmed the CART results for each node in the tree, and the 90% confidence interval for each threshold was calculated to quantify uncertainty. The CART, nCPA, and BHM methods suggested that the nutrient criteria differed significantly within certain nutrient ecoregions and that numerical nutrient criteria of 0.0150–0.222 mg/L TP and 0.300–1.766 mg/L TN may control Chl *a* concentrations in the various lake ecoregions. The results of this analysis suggest that the integration of CART, nCPA and BHM might be useful for determining nutrient thresholds.

1. Introduction

Nutrient criteria representing the enrichment status of surface waters that are minimally impacted by human development may be defined as the threshold value that supports a particular beneficial designated use (US EPA, 1998, 2000). In many developed countries, the establishment of nutrient criteria has recently been assumed to be important for protecting waterbodies from the deleterious effects of nitrogen and phosphorus pollution (Dodds, 2003). These criteria provide a basis for assessing the attainment of designated uses and measuring progress toward achieving water quality goals (US EPA, 1998, 2000, 2008).

Since 2008, China has implemented the Regional Nutrient Criteria Research Plan to create nutrient ecoregions and develop a set of regional lake nutrient criteria. Many studies have examined the methodology of nutrient criteria development in China (Huo et al., 2013, 2014a; Zhang et al., 2014). Three types of scientifically based approaches recommended by the US Environmental Protection Agency (EPA) have contributed to the development of numerical criteria for Chinese lake ecoregions: the reference condition approach, mechanistic modeling, and stressor-response analysis (US EPA, 2000, 2010). These methods also have been applied in Europe and Canada (Cardoso et al., 2007; Carvalho et al., 2008; Ramin et al., 2011). The reference condition approach, which includes the reference lake, lake population distribution and trisection methods, is preferred for ecoregions with available reference lakes (Dodds and Oakes, 2004; Dodds et al., 2006). The mechanistic modeling approach requires sufficient data to identify the appropriate equations for characterizing a waterbody or group of waterbodies and to regulate the parameters in these equations (US EPA, 2010). Stressor-response analysis is used to derive numerical criteria if minimally affected sites cannot be identified and paleoecological or historic data are not available (Bowman and Somers, 2005; Stoddard et al., 2006; Huo et al., 2013). The stressor-response model is more appropriate for establishing nutrient criteria in China because of the widespread contamination of aquatic ecosystems by industrialization, urbanization, and agriculture across various lake ecoregions (Huo et al., 2013).

Many statistical models have been developed to explore the stressor-response relationship between nutrient variables and response variables that are directly or indirectly related to a designated water use (Lamon and Qian, 2008; Ramin et al., 2011; Huo et al., 2014a; Zhang et al., 2014). For example, Lamon and Qian (2008) employed a Bayesian multilevel modeling approach to estimate a linear model for the prediction of chlorophyll a (Chl *a*) from total nitrogen (TN) and total phosphorus (TP). However, the biological response to nutrient gradients might be subtle and would most likely be difficult to detect with a linear regression analysis (Brian et al., 2013). Moreover, ecological responses to environmental gradients are often nonlinear, non-normal, and

* Corresponding author. Tel.: +86 01084937970.
E-mail addresses: huoshouliang@126.com, xibeidou@263.net (B. Xi).

http://dx.doi.org/10.1016/j.ecolind.2015.05.065
1470-160X/

heterogeneous (Legendre and Legendre, 1998). Hence, in this study, nonparametric approaches including a classification and regression tree (CART), a nonparametric changepoint analysis (nCPA) and a Bayesian hierarchical modeling (BHM) method are developed to determine nutrient thresholds at which Chl *a* variation changes across nutrient concentrations.

CART can be used to explore complex nonlinear relationships between a response variable and a set of independent continuous or categorical variables and to quantify uncertainty in model predictions (Breiman et al., 1984; Zheng et al., 2009). CART model derives simple prediction rules by dividing the predictor variable space into subspaces and each subspace is assigned a single response variable value (Qian, 2009). Using a CART model as an exploratory data analysis tool, we are able to explore the structure of data while imposing few prior assumptions (Qian and Anderson, 1999). As changepoint analysis methods, nCPA and BHM are used to evaluate the positions of thresholds or changepoints in binary relationships and provide natural candidates for nutrient criteria in certain cases (US EPA, 2010). These approaches are based on the idea that a structural change in an ecosystem may lead to changes in both the mean and the variance of the ecological response variable (Qian et al., 2003). The preliminary survey of scatter plots or ecological knowledge manifests that a threshold exists prior to applying nCPA and BHM because these two methods will identify a changepoint no matter one truly exists or not (US EPA, 2010). These two methods may be useful in the development of criteria because they (1) estimate discrete, numerical values of the predictor variable that lead to ecological changes; (2) provide an estimation of uncertainty by developing confidence intervals; and (3) make very few assumptions about the data (Qian et al., 2003).

In this study, the CART, nCPA and BHM methods are used to develop region-specific nutrient criteria to address the problem of eutrophication. The specific objectives of the study are (1) to identify variables that significantly contribute to the variability in the response of Chl *a* using the CART method, (2) to verify the CART results for each node in the tree by applying the nCPA and BHM methods, (3) to interpret the uncertainty of each response threshold by calculating the 90% confidence interval, and (4) to explore ecoregional differences and seasonal variations in nutrient criteria.

2. Materials and methods

2.1. Data sources and data analysis

The lakes of China have been divided into eight lake nutrient ecoregions based on a spatial cluster analysis that considered climate (precipitation and temperature), physiography (elevation and geomorphology), and moisture index (Fig. 1) (Huo et al., 2014b). Nutrient data for seven ecoregion lakes were obtained from environmental agencies and scientific institutes (the Qingzang Plateau Lake Ecoregion was excluded because of its high water quality and a lack of monitoring data), including the ambient monitoring network supported by the Department of Provincial Environmental Protection of China. The obtained data consisted of measurements for stressor variables such as TN and TP and response variables such as Chl *a*. The final dataset was heterogeneous because certain lakes were sampled every month for 20 years, whereas others were sampled only once or three times per year. In this study, we only used those lakes sampled in at least three surveys (in three water periods) in separate years over this time interval and averaged these data for every year for a given lake. The final dataset contained a total of more than 170 lakes, which were primarily sampled between 1991 and 2010. Observations of TP, TN, and Chl *a* from April to September were used as the dataset for the spring and summer to represent seasonal effects. The compiled data included those monitoring stations for which at least one Chl *a* measurement was available. The nutrient indexes and Chl *a* were analyzed using standard testing procedures as recommended by the Ministry of Environmental Protection of China (PRC EPA, 2002).

The detection limits for TP and TN were 0.01 mg/L and 0.1 mg/L, respectively. Observations in the database below the detection limits were assigned values equal to one-half the detection limits because these observations were encountered infrequently (less than 15% of the total dataset). This method of addressing the limits of detection has been reported to be sufficiently accurate for determining descriptive statistics such as the mean and standard deviation (Dodds et al., 2006; Suplee et al., 2007; US EPA, 2006).

3. Methods

3.1. Classification and regression tree analysis (CART)

CART is attractive for exploratory environmental and ecological studies due to its ability to address both continuous and discrete variables, the inherent capability of the predictors of the model to interact, and its hierarchical structure (Qian and Anderson, 1999; Qian, 2009). In the CART method, as a binary recursive partitioning method, the distribution/variation of a target variable is constructed and divided into two mutually exclusive branches based on the explanatory variable showing the largest reduction in variation of a target variable after the partition (Choi et al., 2013). Each branch is then split into two sub-branches by the other variables such that a dataset is successively split into increasingly homogeneous subsets until it is infeasible to continue (Clark and Pregibon, 1992). The optimal CART model is selected based on a cross-validation procedure to minimize a model's prediction error (Breiman et al., 1984).

CART analysis is a non-parametric method and makes no assumptions about the underlying distribution of values of the predictor variables. Thus, CART can accommodate numerical data that are highly skewed or multi-modal as well as categorical predictors with either an ordinal or non-ordinal structure. These qualities facilitate analysis by reducing the time spent estimating whether variables are normally distributed and transforming non-normally distributed data (Qian and Anderson, 1999). In this study, CART was used to identify the most important variables affecting Chl *a* concentrations.

3.2. Nonparametric changepoint analysis (nCPA)

Suppose that $y_1, \ldots, y_n$ is the sequence of values of the response variable observed along the ordered stressor gradient $x_1, \ldots, x_n$. The solution to a changepoint problem is a value $r (1 \le r \le n)$ that divides the response variable into two categories, $y_1, \ldots, y_r$ and $y_{r+1}, \ldots, y_n$, where each category has distinct characteristics such as the mean or the variance. The corresponding value of the environmental variable x_r is the threshold value (Qian et al., 2003).

Nonparametric changepoint analysis (nCPA) is an approach for calculating the location of thresholds or changepoints in bivariate relationships; these changepoints can provide natural candidates for nutrient criteria (US EPA, 2010). If observations from multiple sites are ordered along the gradient, a threshold or sudden change in the statistical attributes of the dependent variable will occur in the relationship between a stressor and a response. A changepoint analysis can therefore be used to determine the point at which the change occurs (Breiman et al., 1984; Qian et al., 2003).

Diverse methods are available to identify changepoints. The choice of a particular method depends primarily on the statistical attribute estimated. Deviance reduction methods were used to develop the nCPA method for environmental threshold