Ecological Engineering 53 (2013) 68–78

Contents lists available at SciVerse ScienceDirect

Ecological Engineering

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

Forecasting the cyanotoxins presence in fresh waters: A new model based on genetic algorithms combined with the MARS technique

J.R. Alonso Fernández [a,*], C. Díaz Muñiz [a], P.J. Garcia Nieto [b], F.J. de Cos Juez [c], F. Sánchez Lasheras [d], M.N. Roqueñí [c]

[a] Cantabrian Basin Authority, Ministry of Agriculture, Food and Environment, 33071 Oviedo, Spain
[b] Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain
[c] Mining Exploitation and Prospecting Department, University of Oviedo, 33004 Oviedo, Spain
[d] Department of Construction and Manufacturing Engineering, University of Oviedo, 33204 Gijón, Spain

ARTICLE INFO

Article history:
Received 7 August 2012
Received in revised form 17 October 2012
Accepted 1 December 2012
Available online 25 December 2012

Keywords:
Statistical machine learning techniques
Cyanobacteria
Cyanotoxins
Genetic algorithms (GAs)
Multivariate adaptive regression splines (MARS)

ABSTRACT

Cyanobacteria are one of the major concerns to public health since some of them produce a range of potent toxins (cyanotoxins). This group of microorganism can be present in drinking and recreation waters representing a health risk for animals and human being. For this reason, as prevention, it is important to bring forward their presence. In this study, using physical–chemical and biological parameters, a hybrid approach based on genetic algorithms (GAs) combined with the multivariative adaptative regression splines (MARS) technique, was developed and applied for forecasting the presence of cyanobacteria in a water reservoir (Trasona reservoir, Northern Spain) and in consequence, the cyanotoxin risk. The significance of each biological and physical–chemical variables used for its determination was assessed and a predictive model useful for preventing the presence of cyanobacteria, and consequently of cyanotoxins, was defined.

1. Introduction

Cyanobacteria are now classified as bacteria although the term *blue-green algae* is still used frequently in practice. Cyanobacteria can be found in almost every conceivable environments: oceans, lakes and rivers as well as on land. Even they flourish in Artic and Antarctic lakes (Quesada et al., 2006; Reynolds, 2006), hotsprings and wastewater treatments plants (Smith et al., 2008; Huisman et al., 2010).

Cyanobacteria can produce a range of potent toxins called *cyanotoxins* (Spoof et al., 2006; Reynolds, 2006), and in freshwater ecosystems are the most common cause of eutrophication. The blooms are not always green (Smith et al., 2008; Huisman et al., 2010).

Cyanotoxins are commonly referred as hepatotoxins and neurotoxins (Reynolds, 2006). Several cases of poisoning wild and domestic animals, and also humans, involving cyanotoxins have been reported worldwide (Hillebrand et al., 1999; Reynolds, 2006). Specifically, cyanotoxins are an important environmental problem in reservoirs (Quesada et al., 2004). The association of toxicity with cyanobacterial blooms has frequently led to the closure of recreational waters when blooms are observed. Generally these blooms are harm-less. Otherwise, they are called harmful algal blooms (HABs) (Quesada et al., 2004).

Therefore, cyanotoxins are an important environmental problem in reservoirs (Vasconcelos, 2006; Stewart et al., 2006). Water is never perfectly clean and polluted water is also a continuing threat to human health and welfare (Dasí et al., 1998; de Hoyos et al., 2004). The cyanotoxins include neurotoxins, hepatotoxins, cytotoxins, and endotoxins (Dixit et al., 2005; Willame et al., 2005; Seckbach, 2007; David et al., 2009; Peschek et al., 2011). Most reported incidents of poisoning by microalgal toxins have occurred in freshwater environments, and they are becoming more common and widespread (Negro et al., 2000).

Cyanotoxins are often implicated in what are commonly called red tides or *harmful algal blooms* (HABs) (Fogg et al., 1973). Lakes and oceans contain many single-celled organisms called phytoplankton. Under certain conditions, particularly when nutrient concentrations are high, these organisms reproduce exponentially. The resulting dense swarm of phytoplankton is called an algal bloom. These can cover hundreds of square kilometers and can be easily seen in satellite images. Individual phytoplankton rarely live more than a few days, but blooms can last weeks (de Hoyos et al., 2004).

* Corresponding author. Tel.: +34 658 061580; fax: +34 985 103354.
E-mail address: jrafernandez@chcantabrico.es (J.R. Alonso Fernández).

0925-8574/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecoleng.2012.12.015

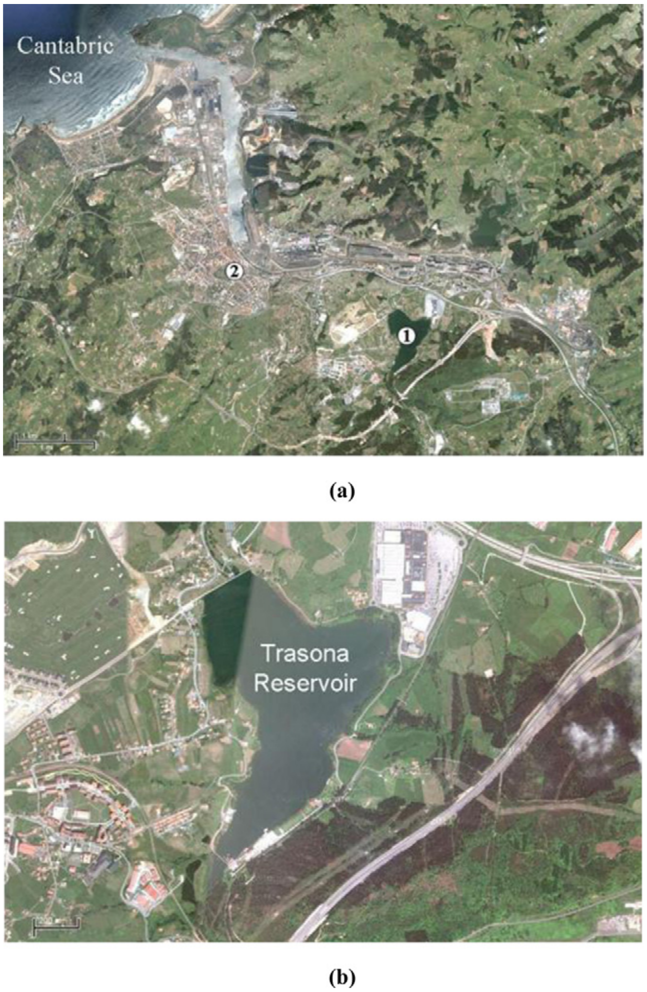

Fig. 1. (a) Aerial photograph of the city of Avilés (Northern Spain) (2) and Trasona reservoir (1); and (b) an aerial photograph of Trasona reservoir in great detail.

Cell counts per se are inadequate as a measure of relative algal biomass. A standard biomass estimate is essential for comparing the relative contribution of different microalgae. As a result, depending on the equivalent geometric shapes for each microalgae, different sets of equations are used in order to estimate their biovolume (Hillebrand et al., 1999).

The purpose of this research is to perform a more complete regression model based on the combination of a genetic algorithm (GA) (Goldberg, 1989) to reduce the initial total number of predicting variables and then the application of the MARS technique (Friedman, 1991; Friedman and Roosen, 1995; Vapnik, 1998; Chou et al., 2004; de Cos Juez et al., 2009; García Nieto et al., 2012) to identify cyanotoxins in the Trasona reservoir (Asturias, Northern Spain) (see Fig. 1) as a function of the main predicting variables selected previously using a suitable genetic algorithm. This reservoir, initially destined to an industrial supply, is now complemented with a recreational use as a high performance training center of canoeing. It is a eutrophic ecosystem, which has been characterized for cyanobacteria outcrops in certain periods, which sometimes has produced variable concentrations of cyanotoxins, mainly mycrocistins.

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution (Goldberg, 1989; Davis, 1991; Sivanandam and Deepa, 2010). In this study, this heuristic is used to carry out a dimensional reduction by identifying patterns in the experimental data set. This technique permits the selection of six main variables from a total number of twenty-four predicting variables in this complex problem, with minimal loss of information. GAs belong to the larger class of evolutionary algorithms (EA) (Haupt and Haupt, 2004; Engelbrecht, 2007), using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

The aim of this research work is to construct a hybrid GA–MARS model to identify spatial cyanotoxins in waterways in the Trasona reservoir (Principality of Asturias, Northern Spain). It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models non-linearities and interactions as those analyzed in this innovative research work successfully.

This innovative research work is structured as follows. In the first place, the necessary materials and methods are described to carry our this study. Next the obtained results are shown and discussed. Finally, the main conclusions drawn from the results are exposed.

2. Materials and methods

2.1. Experimental data set

The data set used for the hybrid GA–MARS model developed here were collected over 6 years (2006–2011) from lots of samples in the Trasona reservoir. The total number of data processed was about 151 values (see Appendix A). The time distribution of this data set depends on the cyanobacterial blooms occurrence. In this way, four samples per month were collected in July, August, September and October of 2006, 2007 and 2008 (months with cyanobacterial blooms occurrence). However, in November of 2006, 2007 and 2008, the sampling was carried out three times per month. For the remaining months, the reservoir was sampling twice per month, except in 2011 where only two samples were collected: in July and September. The information of the biological parameters is expressed in biovolume (cubic millimeters per liter) of phytoplankton species. Specifically, this reservoir was sampled several times a month from January 1, 2006 to December 31, 2011, following the sampling protocols for lakes and reservoirs of the Spanish Ministry of Agriculture, Food and Environment, which are consistent with the guidelines established by the European Union and international agencies dealing with these issues (Quesada et al., 2004). In practice, a single point of sampling is taken into account in the place of greater depth of the reservoir. A Niskin hydrographic bottle was used to collect the water samples. The Niskin bottle is a plastic cylinder with stoppers at each edge, connected by an elastic cord (see Fig. 2(a)). The stoppers are held open by plastic cords attached to a release mechanism. Two side clamps of the cylinder are used to attach the bottle to a hydrographic line so that it can be lowered to a specific depth in the water. When a small weight encircling the hydrographic line is released underneath, the release mechanism is started, so that the water is collected from that depth. The samples were taken at different depths in the euphotic zone (Dasí et al., 1998), that is to say, the layer closer to the surface that receives enough light so that photosynthesis to take place. Its depth, determined by the transparency of the water, is measured using a Secchi disk (see Fig. 2(b)). In this sense, the euphotic zone depth is equal to 2.5 times the depth at which the pattern on that disk is no longer visible when it is lowered down in the water. The values of phytoplankton and concentrations of cyanotoxins, chlorophyll and other physicochemical parameters were determined from a sample composed of five homogeneous subsamples obtained with the