Science of the Total Environment 407 (2009) 2536–2545

Contents lists available at ScienceDirect

Science of the Total Environment

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

Determination of the optimal parameters in regression models for the prediction of *chlorophyll-a*: A case study of the Yeongsan Reservoir, Korea

Kyung Hwa Cho [a], Joo-Hyon Kang [a], Seo Jin Ki [a], Yongeun Park [a], Sung Min Cha [a], Joon Ha Kim [a,b,*]

[a] *Department of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, South Korea*
[b] *Sustainable Water Resource Technology Center, Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, South Korea*

ARTICLE INFO

Article history:
Received 8 October 2008
Received in revised form 28 December 2008
Accepted 8 January 2009
Available online 10 February 2009

Keywords:
Principal Component Regression (PCR)
Multiple Linear Regression (MLR)
F-overall statistics
Optimal parameters
Chlorophyll-a Prediction
Uncertainty

ABSTRACT

Statistical regression models involve linear equations, which often lead to significant prediction errors due to poor statistical stability and accuracy. This concern arises from multicollinearity in the models, which may drastically affect model performance in terms of a trade-off scenario for effective water resource management logistics. In this paper, we propose a new methodology for improving the statistical stability and accuracy of regression models, and then show how to cope with pitfalls in the models and determine optimal parameters with a decreased number of predictive variables. Here, a comparison of the predictive performance was made using four types of multiple linear regression (MLR) and principal component regression (PCR) models in the prediction of *chlorophyll-a* (*chl-a*) concentration in the Yeongsan (YS) Reservoir, Korea, an estuarine reservoir that historically suffers from high levels of nutrient input. During a 3-year water quality monitoring period, results showed that PCRs could be a compact solution for improving the accuracy of the models, as in each case MLR could not accurately produce reliable predictions due to a persistent collinearity problem. Furthermore, based on R^2 (goodness of fit) and *F*-overall number (confidence of regression), and the number of explanatory variables (*R-F-N*) curve, it was revealed that PCR-F(7) was the best model among the four regression models in predicting *chl-a*, having the fewest explanatory variables (seven) and the lowest uncertainty. Seven PCs were identified as significant variables, related to eight water quality parameters: pH, 5-day biochemical oxygen demand, total coliform, fecal indicator bacteria, chemical oxygen demand, ammonia–nitrogen, total nitrogen, and dissolved oxygen. Overall, the results not only demonstrated that the models employed successfully simulated *chl-a* in a reservoir in both the test and validation periods, but also suggested that the optimal parameters should cautiously be considered in the design of regression models.

1. Introduction

1.1. Environmental risks and challenges, eutrophication

Worldwide, effective water resource management has become one of the most challenging issues in recent years, where special attention has been given to the eutrophication of reservoirs and lakes due to increased public health threats, significant ecological damage, and economic loss (Wu, 1999). Over 28–54% of reservoirs on all continents (UNEP, 2002) or, more recently, 415 sites globally (Selman et al., 2008) have been found to suffer from moderate to high levels of eutrophication (hypoxic conditions and areas of concern), and this trend is expected to increase. Eutrophication is a degradation process that is accelerated by high levels of nutrient input as a result of human activities such as point sources (sewage and storm overflows) and diffuse sources (runoff of commercial fertilizer, herbicides, and animal waste), which typically cause excessive algal growth (enhanced phytoplankton biomass) and depletion of dissolved oxygen (DO) (hypoxia conditions, DO ≤ 5 mg /l) in waterbodies (Glasgow and Burkholder, 2000; Parr and Mason, 2004). This deteriorative process, in turn, stimulates severe water quality impairment with toxic side-effects such as the intensification of health-related problems (Pitois et al., 2001), alteration of the biodiversity and species distribution of ecosystems (Kagalou et al., 2008), restrictions on various water uses, and even increases in the total cost of water treatment for drinking water sources (Kuo et al., 2008).

1.2. Are existing regression models sufficient for predicting levels of chlorophyll-a?

Eutrophication covers a multidisciplinary domain with regards to this subject. However, in terms of a qualitative (statistical) model, most research to date has been devoted to identifying a significant relationship between exploratory variables and *chlorophyll-a* (*chl-a*; Giovanardi and Tromellini, 1992; Innamorati and Giovanardi, 1992;

* Corresponding author. Department of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, South Korea. Tel.: +82 62 970 3277; fax: +82 62 970 2434.
E-mail address: joonkim@gist.ac.kr (J.H. Kim).

0048-9697/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2009.01.017