Decision Support Systems 84 (2016) 89–103

Contents lists available at ScienceDirect

Decision Support Systems

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

A framework based on hidden Markov model with adaptive weighting for microcystin forecasting and early-warning

P. Jiang [a], X. Liu [a,b,*], J. Zhang [c], X. Yuan [d]

[a] Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai 200240, PR China
[b] Department of Industrial & System Engineering, National University of Singapore, Singapore 119260, Singapore
[c] Department of Civil and Environmental Engineering, National University of Singapore, Singapore 117576, Singapore
[d] Department of Cell Biology, University of Alberta, AB T6G 2H7, Canada

ARTICLE INFO

Article history:
Received 9 July 2015
Received in revised form 9 February 2016
Accepted 9 February 2016
Available online 16 February 2016

Keywords:
Decision support systems
Framework
Hidden Markov model
Adaptive exponential weighting
Microcystin forecasting
Early warning of risk

ABSTRACT

Harmful algal blooms during the eutrophication process produce toxins, such as microcystins (MCs), which endanger the ecosystems and human health. Accurate forecasting and early-warning of MCs can provide theoretical guidance for quick identification of risk in water management systems. The variation of MC concentration is affected by not only the status quo of numerous manifest biotic and abiotic factors, but also a hidden variable that represents the uncertainty of variations of these factors. Traditional approaches focus on fitting data precisely but less consider such a hidden variable, which would experience formidable barriers when encountering fluctuations in time-serial data. In this study, to address the forecasting problem with a hidden state variable and the problem of early-warning-of-risk, we build a novel integrated framework which is consist of three parts: 1) a forecasting model based on a Principal Component Analysis (PCA) and an improved Continuous Hidden Markov Model (CHMM) with adaptive exponential weighting (AEW), where the AEW-CHMM is proposed to forecast both the single-step-ahead concentration for general points and fluctuating points, and the three-step-ahead concentration existing immediately after the fluctuating point; 2) Bayesian hierarchical modeling for a ratio estimation; and 3) revised guidelines for the risk-level grading. The case study tests a real dataset of one shallow lake with the proposed approaches and other supervised machine learning methods. Computational results demonstrate that the proposed approaches are effective to offer an intelligent decision support tool for MC forecasting and early warning of risk by risk-level grading.

1. Introduction

Eutrophication, which is an aging process of water bodies caused by nutrient enrichment, constitutes one of the most serious environmental issues regarding the degradation of water quality in rivers, reservoirs, lakes, and oceans. Harmful algal blooms (e.g., microcystis blooms) formed during the eutrophication process can release toxins, which belong to compounds of emerging organic contaminants (EOCs). The most common of these are the microcystis toxins (microcystins or MCs) that include over 80 different congeners [46], of which the microcystin-LR (MC-LR) is the most toxic and most frequently occurring [2]. MCs threaten safe drinking water and the recreational use of beaches and lakes, which lead to illness or death for animals and humans, as well as many environmental and recreational related problems [40,56]. In recent years, as emergency cases of acute liver failure and gastrointestinal syndromes have been increasing, growing attentions have been paid globally to water bodies contaminated by MCs [46,49]. Unfortunately, residents cannot directly observe the degree of MC contamination. For this reason, they cannot make informed choices regarding which aquatic recreational activities are safe to participate in and which are not. They also have difficulties judging whether or not tap-water is drinkable at a certain time because MCs cannot be removed by conventional processes in a drinking water treatment plant. These serious health issues caused by these "invisible killers" indicate that it is critical to identify effective early-warning-of-risk systems to detect dangerous levels of microcystis blooms in reservoirs, lakes, and oceans.

Actually, the variation of MC concentration is affected by: 1) the status quo of numerous manifest biotic and abiotic factors, such as predators, nutrients (e.g., nitrogen and phosphorus), other phytoplankton species, dissolved oxygen, dissolved organics, salinity, pH, transparency, sunlight, temperature, and wind speed [20], and 2) the uncertainty of variations of these factors. The variation uncertainty is a hidden variable, which can be synthetically reflected by the variation of the water body eutrophication level. Traditional approaches focus on fitting or training data precisely and implementing forecasting according to the

* Corresponding author at: Department of Industrial Engineering & Management, Shanghai Jiao Tong University, 800 Dongchuan Road, Min-Hang District, Shanghai 200240, PR China. Tel.: +65 6601 4089; fax: +65 6601 4089.
E-mail address: x_liu@sjtu.edu.cn (X. Liu).

http://dx.doi.org/10.1016/j.dss.2016.02.003
0167-9236/

fit or training principle but less consider such a hidden variable, which would present enormous difficulties when encountering fluctuations in time-serial data.

In most cases, historical data regarding MC concentrations are not available as chemical extraction and determination of MCs are much more difficult than those of biomass or chlorophyll-a (Chl-a). Without possessing the original available data, accurate forecasting of MCs is quite difficult. However, a large amount of field data have shown that a highly correlated linear relation exists between MCs and Chl-a [19,21,47]. Thus, the Chl-a concentration could serve as a helpful indicator of the MC concentration [39,47]. Nevertheless, how to estimate the relation between them without ignoring sampling uncertainties of water samples poses another challenge.

To the best of our knowledge, there is not currently an integrated framework offering an intelligent decision support tool for MC forecasting and early warning of risk. The absence of such a framework presents great inconvenience for reservoir, lake, or ocean management systems. In fact, few researchers and organizations have implemented early warning of risk for MCs due to the absence of original data regarding MCs, and that the World Health Organization (WHO) guidelines for relative risk of exposure to MCs are not detailed. Thus, residents often possess no direct impressions of the potential risk.

Motivated by the aforementioned three challenges, our solutions and contributions of this study are listed as below: 1) In order to address the Chl-a or MC forecasting problem with a hidden state variable in a more robust manner, we propose an improved CHMM with AEW schemes as an extension of forecasting modeling via the CHMM [24,26], by discovering similar patterns and bestowing exponential weighting to them adaptively; 2) To cope with the issue of historical data absence regarding MCs, and to perform uncertainty analysis, a Bayesian hierarchical model is proposed to estimate the ratio of MCs/Chl-a, and transform MC forecasting into Chl-a forecasting, which helps to reduce work complexity; and 3) To effectively implement early warning of risk for MCs, a novel integrated framework is built, which is consist of the PCA-based AEW-CHMM forecasting model, Bayesian hierarchical modeling for the estimation of the above ratio, and revised guidelines for the risk-level grading.

The remainder of this paper is structured as follows: Section 2 reviews the related work. Section 3 introduces forecast targets and the problem characteristics. The framework is presented in Section 4. Section 5 demonstrates a case study to test the effectiveness of the proposed approaches. Finally, in Section 6, we close this paper with some conclusions and future work.

2. Related work

This section reviews the major forecasting approaches and analyzes their advantages and weaknesses, and interprets the deficiencies of existing early-warning-of-risk systems.

2.1. Forecasting approaches

In the literature, approaches that have been newly established to forecast Chl-a, MCs, or algal blooms can be divided into six streams, which include: 1) traditional regression-based models; 2) supervised machine learning; 3) genetic-based models; 4) system simulation models; 5) Markov chain models; and 6) Bayesian-based models. We summarize their advantages and weaknesses in Table 1.

According to Table 1, it is difficult to determine which one is the best approach if no concrete application scenario exists. Nonlinear forecasting, limitation of data volume, and algorithm speed do not constitute main challenges for MC forecasting because most approaches possess their own nonlinear processing abilities, the consciousness of data collection are raising in the big data era, and there are sufficient time and computing power to perform MC forecasting as concentration does not tend to change radically in real time, respectively. However, certain substantial deficiencies do exist in applications of current approaches, including: 1) lacking coping capacity for fluctuation and uncertainty (e.g., MLR and ARIMA); 2) "black-box" system properties (e.g., ANN and SVM); 3) expert experience dependency (e.g., FL and BN); and 4) only the median forecasting (e.g., MC and BN).

To the best of our knowledge, no study has yet forecast MCs via a CHMM which has been successively applied to the fields of genetic analysis, customer relationship forecasting, stock market forecasting, failure prognosis, etc. In this study, to cope with the above deficiencies, the AEW-CHMM is constructed to elucidate the complex relationship between observations and the hidden parameter of the water body eutrophication level, and forecast possible changes of MCs driven by related factors.

2.2. Early-warning-of-risk systems

Some literature has presented early warning approaches or early-warning-of-risk systems for algal blooms [11,29,41]. However, they did not directly measure adverse health effects, which prevents residents from possessing direct impressions of the risk, or permissible range of aquatic activities.

To measure adverse health effects, major institutions, such as the WHO and various national government agencies, released guidelines with varying degrees of elaboration. For example, in 1999 and 2003, the WHO successively promulgated standards on the MC-LR in which the upper limit for drinking water is 1 μg/L and for recreational water is 20 μg/L [12,53]. In Australian government guidelines, recreational water is divided into water of whole-body contact (primary contact), water of incidental contact (secondary contact), and water of no contact (esthetic uses) [37]. The Great Lakes Environmental Research Laboratory (GLERL) indicated that the Australian guideline for recreational water of whole-body contact (e.g., swimming, surfing, and bathing) is 20 μg/L, and for water of incidental contact (e.g., fishing, boating, and walking on the beach) is approximately 100 μg/L [18].

In this study, revised guidelines regarding WHO and Australian guidelines are utilized to rank the risk level of exposure to MCs which are forecast by the proposed approaches.

3. Forecast targets and problem characteristics

This section first defines forecast targets and then introduces the characteristics of MC concentration forecast problem.

3.1. Forecast targets

To validate the proposed approaches in a detailed manner, the targets to be forecast are divided into two types: single-step-ahead forecasting (including general points and fluctuating points) and three-step-ahead forecasting immediately after the fluctuating point. Generally, there exist two kinds of points in conventional time-series, i.e., fluctuating points and general points. They are defined as follows:

Definition 1. Let the scatter plots in Fig. 1 be a scatter function $y = f(x), x\epsilon\{1,2,\ldots,t,\ldots\}$. If $f(t-1)\le f(t)\ge f(t+1)$ or $f(t-1)\ge f(t)\le f(t+1)$, t is an extreme value point in the time series.

Definition 2. Let the first point following the extreme value point be the fluctuating point. All points except for the fluctuating points are called general points in this paper.

For example, in Fig. 1, E_1, E_2, E_3, and E_4 are extreme value points; f_1, f_2, f_3, and f_4 are fluctuating points; and T_{i1}, T_{i2}, and T_{i3} are continuous three-day general points immediately after the fluctuating point f_i $(i = 1, 2, 3, 4)$.