Science of the Total Environment 494–495 (2014) 202–210

Contents lists available at ScienceDirect

Science of the Total Environment

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

Assessment of arsenic concentration in stream water using neuro fuzzy networks with factor analysis

Fi-John Chang [a,*], Chang-Han Chung [a], Pin-An Chen [a], Chen-Wuing Liu [a], Alexandra Coynel [b], Georges Vachaud [c]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei 10617, Taiwan, ROC
[b] Laboratoire d'Environnements et Paléoenvironnements Océaniques et Continentaux, University Bordeaux 1, UMR EPOC, France
[c] Laboratoire Transferts en Hydrologie et Environnement, LTHE, UMR 5564 CNRS-IRD-UJF, Grenoble, France

HIGHLIGHTS

- A novel hybrid model (ANFIS-Gamma test) is used to estimate arsenic concentration.
- Data scarcity is overcome by key factor selection and cross-validation technique.
- Gamma test identifies 3 key input factors by evaluating factor occurrence frequency.
- Impacts of key factors on arsenic variation are drawn by ANFIS membership degree.
- The proposed method gives a quick and reliable way of estimating arsenic concentration.

ARTICLE INFO

Article history:
Received 20 February 2014
Received in revised form 27 June 2014
Accepted 29 June 2014
Available online xxxx

Editor: F.M. Tack

Keywords:
Arsenic
Water quality
Neuro-fuzzy network
Gamma test
Adaptive network-based fuzzy inference system (ANFIS)

ABSTRACT

We propose a systematical approach to assessing arsenic concentration in a river through: important factor extraction by a nonlinear factor analysis; arsenic concentration estimation by the neuro-fuzzy network; and impact assessment of important factors on arsenic concentration by the membership degrees of the constructed neuro-fuzzy network. The arsenic-contaminated Huang Gang Creek in northern Taiwan is used as a study case. Results indicate that rainfall, nitrite nitrogen and temperature are important factors and the proposed estimation model (ANFIS(GT)) is superior to the two comparative models, in which 50% and 52% improvements in RMSE are made over ANFIS(CC) and ANFIS(all), respectively. Results reveal that arsenic concentration reaches the highest in an environment of lower temperature, higher nitrite nitrogen concentration and larger one-month antecedent rainfall; while it reaches the lowest in an environment of higher temperature, lower nitrite nitrogen concentration and smaller one-month antecedent rainfall. It is noted that these three selected factors are easy-to-collect. We demonstrate that the proposed methodology is a useful and effective methodology, which can be adapted to other similar settings to reliably model water quality based on parameters of interest and/or study areas of interest for universal usage. The proposed methodology gives a quick and reliable way to estimate arsenic concentration, which makes good contribution to water environment management.

1. Introduction

The potential effects of arsenic (As) on human health and ecosystems have raised serious concerns. It still remains a great challenge to effectively model the temporal variation of As concentration in a river due to the inherent rigorous complex hydro-geochemical relation of As with limited water samples. As is a ubiquitous metalloid naturally present in rocks, soils and water. However, large amounts of As could be introduced into freshwater by anthropogenic activities such as agricultural fertilizers, combustion of fossil fuels, and/or domestic waste incineration. As has become a major environmental and human-health preoccupation due to its high bioavailability and toxicity. High As concentration in natural water has turned into a global problem in the USA, China, Bangladesh, Taiwan, Mexico, Argentina, Czech, Canada, Japan and India (Mandal and Suzuki, 2002; Mohan and Pittman, 2007; Armienta et al., 2008; Kulp et al., 2008; Kuo and Chang, 2009; Miyashita et al., 2009; Concha et al., 2010; Novak et al., 2010). Millions of people in the South and Southeast Asia routinely consume groundwater containing unsafe levels of As, in which As concentration is higher than the permissible limit of 10 μg/l (Polizzotto et al., 2008). In the Yun-Lin Country of Taiwan, the Blackfoot disease (BFD) is known to be caused through the direct drinking of As-contaminated groundwater (Chang et al., 2010). Previous studies indicated that the rapid infiltration of surface water would increase

* Corresponding author at: No. 1, Sec. 4, Roosevelt Road, Da-An District, Taipei 10617, Taiwan, ROC. Tel.: +886 2 33663452; fax: +886 2 23635854.
E-mail address: changfj@ntu.edu.tw (F.-J. Chang).

http://dx.doi.org/10.1016/j.scitotenv.2014.06.133
0048-9697/

groundwater elevation in wet seasons and produce reducing conditions, inducing a reductive dissolution of As-bearing Fe (hydr)oxides (Wang et al., 2011). High As concentration is found in shallow groundwater (<60 m) of southern Choushui River alluvial fan in the Yun-Lin County, and reducing conditions created by infiltrated rainfall water in shallow groundwater yield the release of As ions via the reductive dissolution of As-rich Fe oxy-hydroxides (Costa Goncalves et al., 2007; Wang et al., 2011). In turn, nearby biota is exposed to As and certain As compounds tend to accumulate in animal tissues (Peshut et al., 2008). The release of As into water, soil or biologic media results from both geologic (e.g., As is a major element in different types of ore deposits) and anthropogenic sources (e.g., As derives from the percolation of fertilizer residues) (Amini et al., 2008). As is a trace element of particular interest from the perspective of water quality assessment. A better understanding of As in river systems is essential for water quality modeling and water resources management. Nevertheless, hydro-geochemical processes are usually very complex and highly nonlinear, in which high degrees of spatial and temporal variability exist. Modeling complicated processes with unknown factors is a very challenging task.

There are basically two approaches for modeling: the theory-driven (conceptual and physically-based) approach; and the data-driven (empirical and statistical) approach. Theory-driven models represent general internal sub-processes and physical mechanisms, and their parameters are usually site-specific and are generally assumed as a lumped representation of basic characteristics. When building water quality models, it, however, requires extensive surveys and vast information on various hydrological sub-processes to calibrate models and compute final results. Data-driven models are commonly implemented with techniques developed in areas such as statistics, soft computing, computational intelligence and machine learning, and they tend to explore and establish the relationships between historical inputs and outputs. Artificial neural networks (ANNs), a class of data-driven techniques, have been recognized as an alternative tool to traditional methods for modeling dynamic nonlinear systems, where input–output mechanisms may not precisely exhibit. ANNs have been applied with success in many fields, such as hydrological systems (May and Sivakumar, 2009; Alvisi and Franchini, 2011; Rajaee, 2011; Adeloye et al., 2012; Cavalcante et al., 2013), groundwater issues (Nikolos et al., 2008; Chang et al., 2010), air pollution prediction (Heo and Kim, 2004), and water quality assessment (Singh et al., 2009).

An important strength of ANNs is to infer complex relationships without prior knowledge of a system. Noori et al. (2009) indicated that high-dimensional, irrelevant, redundant or noisy variables might be meaningless and the influential degrees of variables might not explicitly exhibit in observed data sets. An appropriate selection of variables can enhance the effectiveness and the domain interpretability of an inference model, which can be beneficial to improve prediction performance and provide a more effective predictor through reducing the number of variables. Consequently, the selection of input variables that are the most relevant to outputs is a crucial step in modeling ANN applications, especially in environmental science, which usually needs to handle extremely complex nonlinear relations between parameters with available monitoring data limited in size. In addition to the construction of estimation models, this study will also focus on selecting the subsets of features that are useful to improve prediction performance and enhance the understanding of the underlying concepts in the models.

In feature selection, it has been recognized that a combination of good features does not necessarily lead to good performance. That is to say, "the m best features are not the best m features" (Peng et al., 2005). There are indirect or direct means to reduce the redundancy among features and select features with the minimal redundancy. Factor analysis is a methodology that uses a subset of variables for illustrating the variability among these variables. The information of the interdependency among variables in a dataset can be obtained and then be used to reduce the dimension of the data set. Therefore, factor analysis is a tool to turn high-dimensional problems into problems with simpler structures and it has been implemented in hydrogeological systems, biosciences, and other applied sciences that deal with large numbers of variables (Love et al., 2004; Bandalos and Boehm-Kaufman, 2009). The Gamma test (GT) is a nonlinear factor selection tool and is widely used to assess the input–output relationship in a numerical data set for identifying the best combination of model inputs (Tsui et al., 2002; Noori et al., 2011; Chang et al., 2013).

In this study, we propose a systematical process that incorporates the GT into the adaptive network-based fuzzy inference system (ANFIS) to form the ANFIS(GT) model for effectively identifying the important factors affecting As concentration and reliably estimating As concentration in the Huang Gang Creek (Taipei, Taiwan) based on limited hydrological and water quality data collected at environmental monitoring stations in the river basin. The behavior analysis of important factors affecting As concentration is further conducted for delivering a better understanding of the complex composition of As pollution in the Huang Gang Creek (a hot spring creek).

2. Methodologies

The purpose of this study is to model As concentration based on limited water samples bearing inherent rigorous complex relations through selecting a subset of features that are useful to build a good predictor. The proposed approach comprises two parts: extract important factors affecting As concentration by the GT; and configure an estimation model of As concentration by the ANFIS coupled with cross-validation techniques. The architecture of the proposed approach is illustrated in Fig. 1, and a brief introduction of the methods used in this study is given as follows.

2.1. Factors selection through the Gamma test (GT)

Selecting the most relevant variables is usually suboptimal for building a predictor, particularly for variables that are redundant. In contrast, a subset of useful variables may exclude many redundant variables. The GT was initially presented by Aðalbjörn et al. (1997) and Končar (1997) and has been adopted to determine the best set of the most relevant variables from a list of possible model inputs (Noori et al., 2010). This study implements the GT to obtain important clues about the features of the data set used for estimating As concentration in the study watershed. The Gamma statistic (Γ) of the GT is an estimate of the variance of noise and is used as an index to assess the performance of input combinations. A Γ closer to zero produces a more suitable output. A total of $(2^m - 1)$ Gamma values are calculated for all possible combinations of m variables. The primary criterion for choosing important variables depends on the probability distribution of Gamma values. The probability of occurrence for variables x_i in the interval of interest is computed by Eq. (1):

$$P(x_i) = \frac{n_i}{n} \quad i = 1, 2, \ldots, m \tag{1}$$

where m is the number of variables, n is the total number of input combinations in the interval of interest, and n_i is the number of the occurrence of the ith variable in the interval of interest. In this study, good and poor input combinations are defined as the combinations associated with the intervals (0%, 10%) and (90%, 100%), respectively. A variable is considered important if it simultaneously possesses a high-frequency of occurrence in good input combinations and a low-frequency of occurrence in poor input combinations (Durrant, 2001; Chang et al., 2013).