Environmental Modelling & Software 40 (2013) 140–150

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

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

# Prediction analysis of a wastewater treatment system using a Bayesian network

Dan Li [a], Hai Zhen Yang [a], Xiao Feng Liang [b,*]

[a] College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China
[b] School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiaotong University, 800 Dongchuan Rd., Minhang District, Shanghai 200240, China

ARTICLE INFO

*Article history:*
Received 29 October 2011
Received in revised form
29 August 2012
Accepted 30 August 2012
Available online 19 September 2012

*Keywords:*
Prediction analysis
Modified sequencing batch reactor
Bayesian network
Inference

ABSTRACT

Wastewater treatment is a complicated dynamic process, the effectiveness of which is affected by microbial, chemical, and physical factors. At present, predicting the effluent quality of wastewater treatment systems is difficult because of complex biological reaction mechanisms that vary with both time and the physical attributes of the system. Bayesian networks are useful for addressing uncertainties in artificial intelligence applications. Their powerful inferential capability and convenient decision support mechanisms provide flexibility and applicability for describing and analyzing factors affecting wastewater treatment systems. In this study, a Bayesian network-based approach for modeling and predicting a wastewater treatment system based on Modified Sequencing Batch Reactor (MSBR) was proposed. Using the presented approach, a Bayesian network model for MSBR can be constructed using experiential information and physical data relating to influent loads, operating conditions, and effluent concentrations. Additionally, MSBR prediction analysis, wherein effluent concentration can be predicted from influent loads and operational conditions, can be performed. This approach can be applied, with minimal modifications, to other types of wastewater treatment plants.



## 1. Introduction

Wastewater treatment is a complicated process which is affected by several microbial, chemical, and physical factors. Real-time prediction analysis of a wastewater treatment system is difficult because of the complex biological reactions that vary with time and environmental conditions.

A series of activated sludge models have been used to predict the performance of wastewater treatment systems (Gernaey et al., 2004; Gujer et al., 1999; Henze, 2007; Henze et al., 1999) and neural network-based modeling has been used to capture the relationships between variables in complex wastewater treatment systems (Cote et al., 1995; Lee et al., 2005). However, activated sludge models involve a large number of reactions and the parameters are often difficult to measure, while neural networks are black-box models that tend not to show dependencies between variables. Given the nonlinearity, uncertainty, and dynamic features of the wastewater treatment process, an alternative modeling platform is needed.

Bayesian networks also called the Bayesian belief networks, are a powerful knowledge representation tool that deals explicitly with uncertainty (Jensen, 1996; Jensen and Nielsen, 2007; Pearl, 1986, 1995a). In the past few decades, Bayesian networks have been used in medical diagnoses (Berzuini et al., 1992; Shwe et al., 1991), military applications (Grois et al., 1998; Hautaniemi et al., 2000), ecological studies (Borsuk et al., 2006; Pollino et al., 2007; Stow et al., 2003; Young et al., 2011), environmental management (Borsuk et al., 2004; Bromley et al., 2005; Uusitalo, 2007; Varis and Keskinen, 2006), water resource management (Castelletti and Soncini-Sessa, 2007; Henriksen et al., 2007; Molina et al., 2010), environmental modeling (Aguilera et al., 2011; Chen and Pollino, 2012), and environmental change (Varis, 1995; Varis and Kuikka, 1997). Bayesian networks have a solid theoretical foundation, flexible inference capability, and convenient decision support mechanism. However, only a few articles have reported the application of Bayesian network in wastewater treatment plants. Chong and Walley (1996) described the use of a Bayesian network to analyze faults in an aerobic wastewater treatment plant and Sahely and Bagley (2001) developed a Bayesian network for diagnosing upsets in an anaerobic wastewater treatment system.

Modified Sequencing Batch Reactors (MSBR) is an advanced Sequencing Batch Reactor (SBR)-based wastewater treatment process. Despite the potential of the Bayesian network technique in the analysis of MSBR, its application for this purpose has not been reported to date. In this paper, we introduce the use of Bayesian networks for the analysis of MSBR. The objective of this work was to

* Corresponding author. Tel.: +86 02134207165; fax: +86 02134207163.
*E-mail address:* liangxiaofeng1021@gmail.com (X.F. Liang).

1364-8152/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2012.08.011

find a way to model and predict a wastewater treatment system based on MSBR via Bayesian networks. Our final objective was to set up an automatic real-time prediction and diagnosis system for a wastewater treatment system.

## 2. Materials and methods

### 2.1. Brief introduction of the Bayesian network

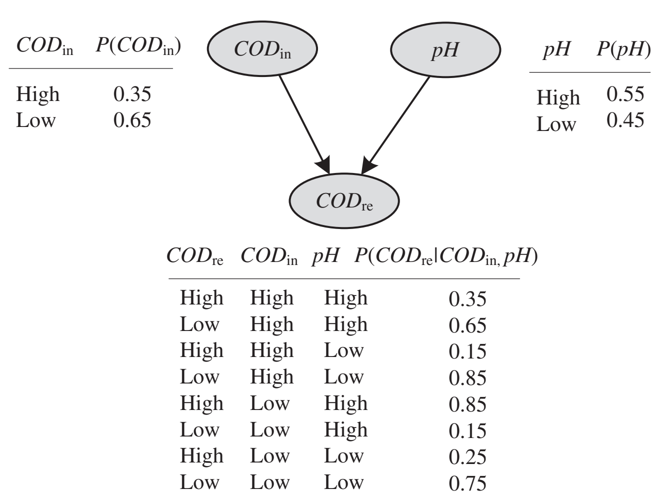
Bayesian networks are directed acyclic graphs comprising of nodes and directed edges connecting the nodes (Jensen, 1996; Pearl, 1995a). Each node represents a random variable and its associated probability distribution or conditional probability distribution. The Bayesian network can be defined as $N = (<V,E>,P)$, where $V$ is a set of nodes expressed as $V = \{V_1,V_2,\ldots,V_n\}$, $E$ is a set of arcs, and $P$ represents a set of conditional probability distributions. A probability distribution is a conditional probability distribution$P(X|.\pi(X))$ when the node is a leaf node. If the node is a root node (i.e., it has no parent nodes), the probability distribution is a marginal distribution, $P(X)$.

Fig. 1 shows an example of a Bayesian network for a simplified anaerobic treatment system, where the influent chemical oxygen demand (COD) concentration ($COD_{in}$), and $pH$ are the parent (i.e., root) nodes. The COD removal efficiency ($COD_{re}$) is a child node affected by its parent nodes. As shown in Fig. 1, no arc exists between the influent COD concentration and $pH$, indicating that these variables do not directly affect each other and are conditionally independent. For illustration, each node is assumed to have two states (high and low). The assumed conditional probabilities for the Bayesian network are also presented in Fig. 1. For example, the probability of $COD_{re}$ being in the high state is 0.35 when both $COD_{in}$ and $pH$ are in the high state.

### 2.2. Methodology for constructing the Bayesian network

Three methods can be used to construct a Bayesian network model (Jensen, 1996; Pearl, 1995b). First, domain experts (i.e., experts familiar with these systems) identify the important variables of the model, and Bayesian network experts determine the Bayesian network structure and parameter distributions based on the qualitative and quantitative relationships that the domain experts provide. Second, domain experts determine the important variables of the model, and Bayesian network experts determine Bayesian network structure based on the qualitative relationships that the domain experts provide. The Bayesian network parameters are then obtained by machine learning from the training data. Third, domain experts identify the important variables of the model. The Bayesian network structure and parameters are then determined by learning from the training data.

In this study, the second method was used to construct the Bayesian network for two reasons. First, the structure of a Bayesian network reveals the qualitative relationships between variables (Jensen and Nielsen, 2007) and previous studies (Yang, 2001) could be used to develop the initial Bayesian network structure for MSBR. Second, the Bayesian network parameters reveal the quantitative relationships between variables (Jensen and Nielsen, 2007) and given that only a minimal number of studies have quantitatively investigated MSBR systems, quantitative relationships could not be reliably obtained through expert opinion. Consequently, data mining was employed to derive the quantitative relationships used in the Bayesian network.




| $COD_{in}$ | $P(COD_{in})$ |
|---|---|
| High | 0.35 |
| Low | 0.65 |

| $pH$ | $P(pH)$ |
|---|---|
| High | 0.55 |
| Low | 0.45 |

| $COD_{re}$ | $COD_{in}$ | $pH$ | $P(COD_{re}\|COD_{in}, pH)$ |
|---|---|---|---|
| High | High | High | 0.35 |
| Low | High | High | 0.65 |
| High | High | Low | 0.15 |
| Low | High | Low | 0.85 |
| High | Low | High | 0.85 |
| Low | Low | High | 0.15 |
| High | Low | Low | 0.25 |
| Low | Low | Low | 0.75 |

**Fig. 1.** Example of a Bayesian network for a simplified anaerobic treatment system.

### 2.3. BN inference

The essence of Bayesian network inference lies in the computation of the posterior probability distribution of a set of random variables (given a specific Bayesian network structure and a set of observations). Observations are integrated into the network through alteration of the probability distributions underlying particular nodes and the probabilities of other nodes are then updated accordingly by the software. The computation is based on *a priori* probability when no observation is given. Denoting an observed variable as $E$, the posterior probability distribution of the query variable denoted by $Q$ is then expressed as

$$P(Q|E = e), \tag{1}$$

where $e$ is the value of the observed variable $E$.

Bayesian network inference can be classified as follows: (a) prediction inference, i.e., from causes to effects; (b) diagnostic inference, i.e., from effects to causes; (c) inter-causal inference, i.e., between causes of a common effect; and (d) mixed inference, i.e., combination of two or more of the above. In this study, mixed inference was used to analyze a wastewater treatment system.

A number of software packages, such as Bayes Net Toolbox (Murphy, 2001), BayesiaLab (http://www.bayesia.com/), Hugin Expert (http://www.hugin.com/), and Netica (http://www.norsys.com/), can be used for parameter learning and inference. In this research, BayesiaLab (5.0) was used in the computation of network parameters and prediction inference.

## 3. Development of a Bayesian network model for a wastewater treatment system based on MSBR

### 3.1. Description of MSBR

MSBR is an advanced wastewater treatment technology based on SBR (Yang, 2001). Over the years, MSBRs with four, five, six, seven, and nine pools have been developed in an effort to further improve the technology. In this work, a six-pool MSBR was employed. Fig. 2 shows the process flow of this MSBR.

The MSBR comprises of two functional areas, the anaerobic–anoxic–oxic and SBR functional areas. The sewage initially enters the anoxic zone, then enters the anaerobic zone before eventually entering the aerobic aeration zone. During this process, organic matter is decomposed by phosphate-accumulating organisms and denitrifying bacteria in the anaerobic and anoxic zones, thereby decreasing the COD concentration significantly. Simultaneously, nitrogen and phosphorus are partially removed. After the anaerobic–anoxic–oxic reaction process, the sewage alternatively

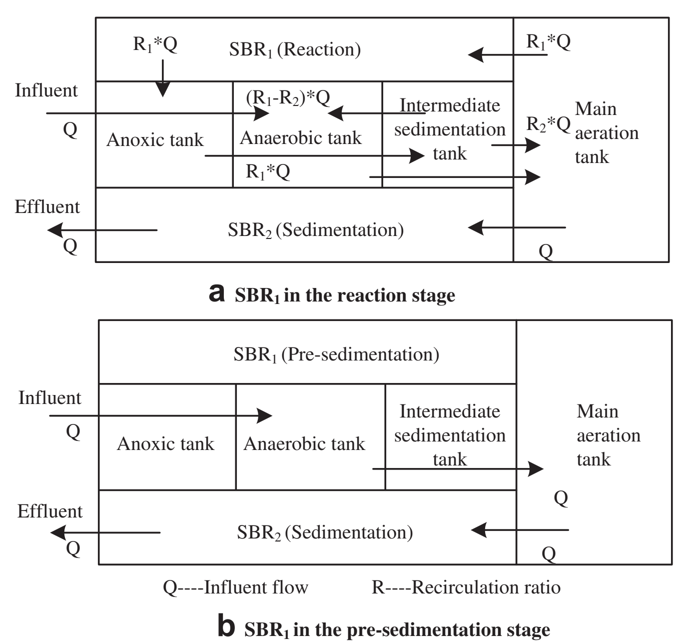



**Fig. 2.** Process flow of an MSBR.