Environmental Modelling & Software 69 (2015) 14–22

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A continuous variable Bayesian networks model for water quality modeling: A case study of setting nitrogen criterion for small rivers and streams in Ohio, USA

Song S. Qian [a, *], Robert J. Miltner [b]

[a] Department of Environmental Sciences, University of Toledo, 2801 W. Bancroft Street, MS# 604, Toledo, OH 43606-3390, USA
[b] Ohio Environmental Protection Agency, 4675 Homer-Ohio Lane, Groveport, OH 43125, USA

ARTICLE INFO

Article history:
Received 2 March 2015
Accepted 2 March 2015
Available online

Keywords:
Bayesian statistics
Nutrient criteria
Biological monitoring
Simulation
Ohio
Clean Water Act
Updating

ABSTRACT

We present a continuous variable Bayesian networks modeling framework that integrates the graphical representation of a Bayesian networks model with empirical model-developing approach. Our model retains the Bayesian networks model's graphical representation of hypothesized causal connections among important variables and employs conventional statistical modeling approaches for establishing functional relationships among these variables. The modeling framework avoids discretizing continuous variables and the resulting models can be updated over time when new data are available or updated using local data to develop a site-specific model. We illustrate the modeling approach using a data for establishing nutrient criteria in streams and rivers in Ohio, U.S.A.

1. Introduction

The Bayesian networks (BN) modeling approach is increasingly used in environmental modeling for supporting management decisions (see Chen and Pollino (2012) and Aguilera et al. (2011) for reviews). One of the most appealing feature of a BN model is likely its graphical approach for illustrating complicated connections among multiple components of a problem thereby facilitating the communication of scientific research to a wide range of stakeholders. Furthermore, the relatively simple graphical structure allows participation of stakeholders in model developments.

An oft-downplayed drawback of the BN approach is the need for discretizing continuous variables. Discretization is a problem not only because of the potential loss of statistical accuracy (Chen and Pollino, 2012), but also because of the difficulty in subsequent interpretation. The difficulty arises because the discretized variable is often categorized as, for example, *low*, *medium*, and *high*. Because there is no consensus on how to properly discretize a continuous variable, the resulting categories are often ambiguous and their meanings depend on the context. For example, Kashuba (2010) developed a BN model for assessing stream ecosystem status in southeast U.S. Using the equal frequency method, the variable defining watershed urban intensity (% developed land) is divided into four bins (*low* 0–11%, *medium low* 12–36%, *medium high* 37–65%, and *high* 66–100%). This division would likely be meaningless if applied to a different region. For example, almost all headwater stream watersheds in Maine, U.S.A. will fall into the category "low" (0–11%) (Susan Davies, 2009, personal communication). In developing a eutrophication BN model for an estuary in North Carolina, Nojavan et al. (2014) divided the variable total nitrogen concentration into three bins (<56, 56–334, and >334 μg/L) based on data from 2007 to 2011. Because the water quality of their study area has been improving since the mid-1990s, the category "low" would include more than half of the observations collected since 2012 (Nojavan, 2014).

Furthermore, how a continuous variable is discretized (discretization methods and number of bins) is directly linked to the subsequent models. Discretization methods and number of bins can change a BN model's structure when using structural learning algorithms (Alameddine et al., 2011), as well as the conditional probability tables when using a fixed model structure (Nojavan, 2014).

* Corresponding author. Tel.: +1 419 530 4230; fax: +1 419 530 4421.
E-mail addresses: song.qian@utoledo.edu (S.S. Qian), bob.miltner@epa.state.oh.us (R.J. Miltner).

http://dx.doi.org/10.1016/j.envsoft.2015.03.001
1364-8152/

Although recognized by many, the problem of discretization is often unmentioned in many applications of BN in the environmental modeling literature (Aguilera et al., 2011). In this paper, we present a framework for developing a BN model without discretizing continuous variables. Our aim is to develop a probabilistic modeling framework that can be used for supporting management and decision-making under uncertainty, specifically, for making operational decisions (Kelly et al., 2013). The approach borrows the basic model-construction strategy used in a BN model (Jensen, 2001), and expands the model specification and fitting using traditional data exploration and regression (Weisberg, 2005; Gelman and Hill, 2007; Qian, 2010) and the model updating using Bayesian computation [Markov chain Monte Carlo simulation, (Qian et al., 2003)]. The model structure is developed through a directed acyclic diagram (DAG) model (Lauritzen, 1996), with connections among nodes represented by empirical models.

We present the general framework of developing a continuous variable Bayesian networks (cBN) model in Section 2.1 and apply it to a data collected for establishing nitrogen criterion for Ohio's small rivers and streams in Section 2.2. We also discuss the process of updating the resulting model using local or regional data for developing local and regional nutrient criteria.

2. Methods

In a traditional environmental modeling domain, we often discuss the trade-off between a mechanistic model and an empirical model. A mechanistic model is a summary of the functional connections among multiple components of a real world problem, reflecting the causal relationship we know. These models are, in theory, suited for supporting management decision-making. But a mechanistic model is often overly complicated and may not be practical for proper model calibration because of the limited availability of appropriate data. Empirical or statistical models establish correlations, which do not necessarily reflect causal relations. Models such as DAG models represent a middle ground between the two approaches. The hypothetical relationships among relevant variables are presented using a graphical model and the functional forms of these relationships (e.g., differential equations) are represented using conditional probability tables. As a result, fast computing algorithms can be applied. However, as noted earlier, discretization can be a source of many problems. We propose a modeling framework to avoid discretization and yet retain the advantages of a BN model.

2.1. A continuous variable Bayesian network model

Graphical models such as the BN models (Pearl, 1986, 1988; Jensen, 2001) take full advantage of the conditional probability structure, not only in model formulation, but also in computation. However, a BN model is limited to using categorical variables. When continuous variables must be used, we either discretize them (BN) or we assume that the continuous variables are normal random variates and connections among nodes are linear [the structural equations model (Bollen, 1989; Grace, 2006)]. With the advent of the Gibbs Sampler (Gelfand and Smith, 1990), the computation requirements are less restrictive. The Bayesian inference software WinBUGS (Gilks et al., 1994; Lunn et al., 2000) (now OpenBUGS (Spiegelhalter et al., 2014)) and JAGS (Plummer, 2003) further popularize the use of Gibbs sampler. We can now conduct the same complex computing with continuous variables without using discretized conditional probability models, without assuming normality, and without being limited to linear models. Consequently, we can take advantage of both the causal relationship (the DAG model) and the available data to build a Bayesian network model using continuous variables.

The basic idea of our modeling approach is to replace the conditional probability tables (CPTs) for factor variables in a BN model with a series of conditional probability distributions for continuous variables. With CPTs, we use the conditional probability formula (the Bayes theorem) to quantify the functional relations among variables and the formula can be readily programmed. When using continuous conditional probability distributions, the computation can be implemented using the Gibbs sampler, or more generally Markov chain Monte Carlo (MCMC) simulation (Gilks et al., 1996; Qian et al., 2003).

We describe the computational strategy using a hypothetical model (Fig. 1(a)). In this simple DAG, two nodes have parent nodes (nodes X_1 and X_2 are parent nodes to Y_1 and nodes Y_1 and X_3 are parents to Y_2). Arrows in the DAG represent the directional dependency among the five variables (e.g., Y_1 is a function of X_1 and X_2). If we consider a DAG such as the one in Fig. 1(a) as a joint probabilistic model of the variables in all nodes, the DAG provides a causal structure which can be translated into conditional probability distributions such that the joint distribution can be simplified. Let V be the collection of all variables represented by the DAG ($V = \{X_1,X_2,X_3,Y_1,Y_2\}$), the joint distribution of V can be represented as $p(V) = \prod_{v\in V} p(v|\text{parents}[v])$, where $p(\cdot)$ represents a probability distribution function. For the model in Fig. 1(a), the joint distribution is $p(V) = p(X_1)p(X_2)p(Y_1|X_1,X_2)p(X_3)p(Y_2|Y_1,X_3)$. When all variables are categorical, these conditional probability distributions are represented by conditional probability tables, and the multiplication operation is done through the Bayes theorem.

When these variables are continuous, we can specify the marginal distributions of X_1,X_2 and X_3 directly (e.g., histograms of data) and the conditional distributions of $Y_1|X_1,X_2$ and $Y_2|Y_1,X_3$ empirically, using perhaps linear or nonlinear regression analysis. For example, if a regression model is used, we may denote the conditional distribution as $Y_1 = f(X_1,X_2,\alpha) + \varepsilon_1$, or $Y_1|X_1,X_2 \sim N(\mu_1,\sigma_1^2)$, where $f(\cdot)$ denotes a linear or nonlinear function of X_1 and X_2, $\mu_1 = f(X_1,X_2,\alpha)$ with coefficients α, and $\varepsilon_1 \sim N(0,\sigma_1^2)$ is the residual random variable. Likewise, exploratory analysis can lead to the conditional distribution of $Y_2|X_3,Y_1 \sim N(\mu_2,\sigma_2^2)$, where $\mu_2 = g(X_3,Y_1,\beta)$. Furthermore, variables in V can be divided into predictors (X's, observed without error) and response variables (Y's, observed with error). When data are available for all variables, the model is reduced to a problem of estimating the probability distribution of model coefficients (α and β) and error variances (σ_1^2,σ_2^2), and the graphical model in Fig. 1(a) is revised to a DAG representing the computational process (Fig. 1(b), where oval nodes are quantities to be estimated, rectangle nodes are variables with observations, and arrows representing the functional dependency). For example, the data node y_1 is a child node of μ_1 and σ_1^2, representing the distributional assumption $y_1 \sim N(\mu_1,\sigma_1^2)$; the parameter node μ_2 has three parent nodes — μ_1,x_3, and model coefficient vector β. If we use a linear regression model, the node μ_2 represents the linear model mean function $\mu_2 = \beta_0 + \beta_1\mu_1 + \beta_2 x_3$.

We note that variable Y_1 in Fig. 1(a) is a parent to Y_2, while in Fig. 1(b) the connection between the two variables are established through their respective means μ_1 and μ_2. In other words, we replace the CPTs in a BN model with functions for calculating μ_1 and μ_2 (f and g in Fig. 1(b)).

Just as eliciting CPTs is an important step of building a BN model, finding the likely functional forms of f and g is an important component for our modeling framework. We can derive the functional forms based on our substantive knowledge (e.g., mechanistic models) or based on empirical modeling through exploratory regression analysis (see Chapter 4 of Gelman and Hill (2007)). When using the empirical modeling approach, we use the DAG as a guide for building component models (finding the likely functional forms of f and g) one at a time. Once these functional forms are established, they can be linked to form the joint distribution of all variables. The resulting model is a continuous variable Bayesian network model (cBN). Model parameters (e.g., α,β in Fig. 1) should be estimated simultaneously using the Gibbs sampler. Once the joint distribution is quantified (all unknown parameters are estimated), statistical inference can be made through Monte Carlo simulations.

Borsuk et al. (2004) proposed a similar modeling approach. However, their model was implemented in Analytica (Lumina, 1997), which requires fixed model coefficients. As a result, conditional models (e.g., $p(Y_1|X_1,X_2)$ are fit independently and cannot be updated upon new observations (nor refit jointly).

2.2. The Ohio example

We illustrate the model-building process using data from wadeable streams and rivers in Ohio collected as part of the effort for setting nutrient criteria. The data is "cross-sectional" in that they represent multiple streams and rivers across the state of Ohio. The objective of the model is to establish a link between nutrient concentration and indicators of stream aquatic ecosystem condition. Through the link, we find the nutrient concentration distribution associated with an aquatic ecosystem that is likely to meet the designated use. Our approach follows the following steps:

1. Developing a conceptual model linking nutrient concentrations and other factors to stream aquatic ecosystems indicators (e.g., Fig. 1(a)),
2. Building empirical (e.g., regression) models among nodes (if such functional relationships are unknown) using available data and expressing these models in terms of conditional probability distributions (probability distribution of a child node conditional on its parent node(s)),
3. Revising the initial graphical model to connect data and unknown parameters (e.g., Fig. 1 (b)), and
4. Estimating all unknown parameters of the joint probabilistic distribution using the Gibbs sampler.

Once the joint distribution of all relevant variables (nodes) are quantified, we can derive the (conditional) distribution of nutrient concentration that is associated with acceptable stream ecosystem indicator values. This conditional distribution can be used to define acceptable nutrient concentrations.

Data collection and the initial nutrient criteria development process are reported by Miltner (2010). The objective of a nutrient criterion is to ensure that streams meet the designated use for aquatic life, which is measured by one or more macroinvertebrate metrics in Ohio. We use the Invertebrate Community Index (ICI) and the EPT taxa richness [EPT, number of Ephemeroptera (mayfly), Plecoptera (stonefly), and Trichopera (caddisfly) taxa in a sample (Ohio EPA, 1978)] in our illustration. Because stream macroinvertebrate community are affected by many other factors (e.g., stream flow, habitat condition, watershed land use, shading, etc.)