Environmental Modelling & Software 87 (2017) 64–71

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Comparative analysis of discretization methods in Bayesian networks

Farnaz Nojavan A. [a, b, *], Song S. Qian [c], Craig A. Stow [d]

[a] Nicholas School of the Environment, Duke University, Durham, NC 27708, USA
[b] ORISE Fellow at the Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI 02882, USA
[c] Department of Environmental Sciences, The University of Toledo, Toledo, OH 43606, USA
[d] NOAA Great Lakes Environmental Research Laboratory, Ann Arbor, MI 48108, USA

ARTICLE INFO

Article history:
Received 22 December 2015
Received in revised form
17 October 2016
Accepted 26 October 2016
Available online 10 November 2016

Keywords:
Bayesian networks
Discretization
Environmental modeling
Equal interval
Equal quantile
Moment matching

ABSTRACT

A key step in implementing Bayesian networks (BNs) is the discretization of continuous variables. There are several mathematical methods for constructing discrete distributions, the implications of which on the resulting model has not been discussed in literature. Discretization invariably results in loss of information, and both the discretization method and the number of intervals determines the level of such loss. We designed an experiment to evaluate the impact of commonly used discretization methods and number of intervals on the developed BNs. The conditional probability tables, model predictions, and management recommendations were compared and shown to be different among models. However, none of the models did uniformly well in all comparison criteria. As we cannot justify using one discretization method against others, we recommend caution when discretization is used, and a verification process that includes evaluating alternative methods to ensure that the conclusions are not an artifact of the discretization approach.

1. Introduction

Bayesian networks (BNs) are probabilistic graphical models that consist of nodes and directed links depicting the dependencies among the variables in the model (Jensen, 2001). Probabilistic relationships among the variables are expressed using conditional probability tables (CPTs). BNs are promising tools to aid reasoning and decision making under uncertainty. The term Bayesian network was first introduced by Pearl (1982) and Spiegelhalter and Knill-Jones (1984) in the field of expert systems. Some of the early appearances of BNs in environmental modeling were by Varis and Kuikka (1997), Varis (1997), and Reckhow (1999).

Several distinct advantages of BNs make them popular for environmental modeling (Kelly et al., 2013). BNs' modularity enables integrating multiple ecosystem components or aspects of the problem (e.g. science network and management network in Johnson et al. (2010)). This is desirable in environmental modeling due to the complexity of natural ecosystems and the associated decision-making processes. Furthermore, BNs can accommodate various knowledge sources and data types such as expert knowledge, previous data from the same system or similar systems with a transparent definition of prior knowledge. Another methodological advantage is the suitability to both data-rich and data-poor ecosystems. BNs can be developed with minimal data in a data-poor ecosystem and as more data become available the model can be updated. Uncertainty is inherent in environmental models due to natural ecosystem variability, current knowledge of environmental processes, model structure uncertainty, data and observation (e.g., observation error, missing data), and computational restrictions. BNs explicitly represent uncertainty by conditional probability distributions for each node and the uncertainty is propagated through the model and presented in the final results. Finally, the capacity of BNs to incorporate new data or updated information using the Bayes' theorem makes them particularly valuable in the context of adaptive management of ecosystems.

The aforementioned advantages of BNs have resulted in many applications in the environmental sciences over the last decade, including natural resources management (McCann et al., 2006; Castelletti and Soncini-Sessa, 2007; Dorner et al., 2007; Farmani et al., 2009), ecological risk assessment (Borsuk et al., 2004; Pollino et al., 2007; Barton et al., 2008; Malekmohammadi et al., 2009), and

* Corresponding author. ORISE Fellow at the Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI 02882, USA.
E-mail address: farnaz.nojavan@duke.edu (F. Nojavan A.).

http://dx.doi.org/10.1016/j.envsoft.2016.10.007
1364-8152/

integrated models (Bromley et al., 2005; Croke et al., 2007; Johnson et al., 2010; Kragt et al., 2011). While BNs have many advantages, a current limitation in their practical implementation is that most software cannot accommodate continuous variables, thus binning or discretization is required for model development (Death et al., 2015).

Aguilera et al. (2011) examined 118 papers published between 1990 and 2010 related to the applications of BNs in the environmental sciences. Among these papers, 52.6% used discrete data and 30.7% used some form of discretization method to convert continuous data; however, 48.6% of the papers did not include any description about the discretization process, 25.7% used experts to discretize the continuous data into intervals, 2.9% used equal interval, and 2.9% used equal quantile, and 2.9% used the default method of the software (Aguilera et al., 2011).

Although discretization is common in a BN's implementation, it has the potential to result in loss of information, and the consequences in inference and decision-making have not been well-explored (Death et al., 2015). In this paper, we investigate how discretization may result in differing decisions, as various discretization methods lead to different characterization of the underlying continuous distribution. We used long-term water quality monitoring data from a large number of lakes in Finland and examined the well-studied relation among chlorophyll *a*, total phosphorus, and total nitrogen in lakes to evaluate the effects of discretization methods on the final model.

2. Material and methods

2.1. Study design

There are two decisions to be made when discretizing continuous data: (1) the discretization method and (2) the number of break points/intervals. We designed an experiment to assess the effect of three commonly used discretization methods and the number of break points/intervals on the resultant BNs. The BNs presented here are simple and consist of three nodes describing the relation among total nitrogen (N), total phosphorus (P), and chlorophyll *a* (Chl *a*) concentrations in Finnish lakes (Fig. 1).

Nine BNs were developed, each corresponding to one of the nine combinations of discretization methods and number of break points. The BNs were fitted to discretized data using the bnlearn package in R developed for structure learning, parameter learning, and inference (Scutari, 2010; Nagarajan et al., 2013; R Core Team, 2014). The code is available as part of an online supplementary material and the results presented here are reproducible (Nojavan et al., 2015).

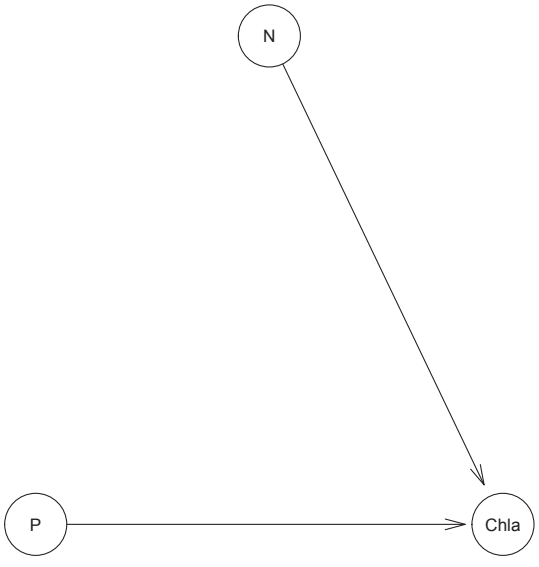

Fig. 1. Directed acyclic graph.

BNs are defined qualitatively as directed acyclic graphical (DAG) models with conditional probability tables (CPTs) depicting the quantitative dependencies among the variables. The DAG and CPTs represent the model's structure and parameters, respectively. The structure shows cause-effect relations of the underlying system. The structure of this paper's example is formed based on two criteria. Firstly, data availability is an important concern when developing empirical models. While nitrogen and phosphorus loading to the lakes are the root cause of the Chl *a* concentration variation, such data is not available in many cases. Finnish lakes (Malve and Qian, 2006) and the National Lakes Assessment (NLA) (USEPA, 2009) data sets are examples of such data availability imposed restrictions on model structure. Secondly, our goal here was to keep the example structure as simple as possible to illustrate the impact of discretization on model results. We will discuss the aforementioned two points in detail in Section 4.

The structure of this paper's example is based on the literature findings on the dependency of Chl *a* on N and P (Dillon and Rigler, 1974; Smith, 1982) applied to the Finnish lake data (Malve and Qian, 2006). The structure is specified using the modelstring function from the bnlearn package. The conditional probability table for each node is calculated as $Pr(y \in k|x_i)$, where x_i is the set of all parent nodes for y and k is the kth interval. N and P are considered root nodes, as they do not have any parents; hence, the CPTs for them is reduced to the prior probability distributions from the training data. The CPT for the Chl *a* node is computed by $Pr(Chla \in k|N, P)$. The CPT estimation is done using the bn.fit function from the bnlearn package.

2.1.1. Discretization methods

We use three commonly used discretization methods, each designed to capture certain features of the data distribution, to discuss the potential issues.

2.1.1.1. Equal interval. Equal interval is a discretization method in which the data are divided into equal length intervals. This method is ideal when the data distribution is roughly uniform. When the underlying distribution of the variable is not uniform or when outliers are present in the data, the equal interval method can be problematic (e.g., resulting in intervals with few observations). In our dataset, there are several unusually low and high nitrogen concentration values. Using three intervals, discretization with these extreme data points included, results in the following break points (on the logarithmic scale): 3.434, 5.665, 7.896, 10.127 (i.e., low if $N \in (3.434-5.665)$, medium if $N \in (5.665-7.896)$, and high if $N \in (7.896-10.127)$). The low nitrogen interval includes only ten observations (0.05% of all observations). In contrast, the discretization with the "outliers" removed results in the following break points: 4.500, 5.818, 7.137, 8.455. The definition of low nitrogen, on the logarithmic scale, changes from <5.665 to <5.818 (high from >7.896 to >7.137).

Additionally, many kinds of data, particularly pollutant concentration data, are right-skewed and are often log-transformed before analysis to make their distribution more symmetric (Koch, 1966; Ott, 1990). The equal interval discretization method is not invariant to nonlinear transformations, such as a log-transformation, where the relative spacing among observations is not preserved. Hence, the decision to log-transform the variables impacts the intervals and final results.

2.1.1.2. Equal quantile. Equal quantile (or equal frequency) discretization method divides data into groups of (approximately) equal sample size. This method is designed to capture distributions with one or more concentrated "modes". The equal quantile method, when performed unsupervised using a software, can result in assignment of the same value to different intervals, depending on how the discretization method is implemented, if there are multiple occurrences of the same value (Chen and Pollino, 2012). The order of the observations is important for this method; however, the relative spacing among observations is irrelevant. Hence, equal quantile discretization is not affected by non-linear transformations.

2.1.1.3. Moment matching. Moment matching discretization method matches the moments of the discretized distribution with the moments of the continuous data. This method is designed to capture the data distribution in a systematic way. The discretized data distribution will have the same moments as the continuous data distribution. The following set of equations is solved to find n break points X_1,X_2, …,X_n and associated discrete probability distribution P_1,P_2, …,P_n.

$$\begin{gathered} P_1X_1 + P_2X_2 + \ldots + P_nX_n = 1\text{s tmoment(mean)} \\ P_1(X_1 - mean)^2 + \ldots + P_n(X_n - mean)^2 = 2\text{nd moment(variance)} \\ \vdots \\ P_1(X_1 - mean)^{2n-1} + \ldots + P_n(X_n - mean)^{2n-1} = (2\text{n} - 1)\text{moment} \\ P_1 + P_2 + \ldots + P_n = 1 \end{gathered}$$

As the number of the moments being matched increases, the discrete distribution becomes a more accurate approximation of the continuous distribution; however, the number of equations in the moment matching approach increases linearly with the number of break points (2^*n-1). Our objective is then to find a point that is feasible to all these equations (i.e., constraints). This leads to an optimization problem that is nonlinear and non-convex. This class of optimization problems are generally very hard to solve, even numerically, as the number of variables and/or constraints increase. Because the solution approach essentially involves an exhaustive search of the entire space, the computational complexity of the problem grows exponentially with the number of break points. In particular, the problem becomes quickly intractable as the number of break points exceeds four. Furthermore, the increased number of intervals in BNs is an additional complexity as described below (see section 2.1.2).

2.1.2. Number of intervals

Ecological models using BNs typically include 2–10 intervals (Uusitalo, 2007); however, we chose three to five as the number of intervals, a more realistic range considering the restrictions commonly imposed by data availability and model complexity. It may seem that more intervals would better represent continuous data; however, the size of the conditional probability table for every node, calculated as the product of the number of intervals of that node and the number of intervals of each parent node, also increases. Even in a simple three node network such as our example (Fig. 1), the difference between three and five intervals for each variable results in calculating 27 (3^3) versus 125 (5^3) conditional probabilities for Chl *a*. Thus, a large data set is required to justify using many intervals, and even then some conditional probabilities may be based on relatively few observations. Therefore, although a model may be more precise as the number of intervals increases, the model is not necessarily more accurate (Marcot et al., 2006). We categorized the continuous data set into the following number of intervals and labeled them accordingly:

- Three (Low, Medium, and High)
- Four (Low, Medium, Medium High, and High)
- Five (Low, Medium Low, Medium, Medium High, and High)

2.2. Data

We used lake monitoring data from Finland reported by Malve and Qian (2006). The large number of lakes (≈2289) in Finland, coupled with long-term monitoring from 1988 to 2004 during July and August resulted in a rich data set (n = 19,247). Our example BN (Fig. 1) describes the relationships among N (μg L^{-1}), P (μg L^{-1}), and Chl *a* (μg L^{-1}). A scatterplot matrix indicates that these three variables are strongly correlated, thus are suitable for development of a BN that will quantify these dependencies (Fig. 2). The correlation between N and P concentration is the result of a common cause, nutrient loading; however, as discussed in Subsection 2.1 and Section 4, to demonstrate the impact of discertization on BNs we will keep the model structure simple. They are also approximately, marginally, normally distributed, with most mass near the center and narrow relatively symmetric tails.

2.3. Comparison

We use both graphical and numerical comparisons to illustrate inconsistencies in marginal distributions obtained using the three alternative discretization methods and differing number of intervals. Using graphical tools, we compare the resultant, discretized, marginal Chl *a* distributions with one another and with the empirical, marginal Chl *a* distribution. Our goal in comparing the marginal distributions is to show discretization changes the starting point, data set, of BN's development and this would impact the results as a consequence.

To examine the effect of the discretization method and the number of break points on predictive accuracy we used a cross-validation procedure. The data were randomly divided into two subsets for training and testing purposes; the training data held 90% of the observations and the testing subset contained 10%. The training dataset was used to fit the BN model and the resulting model was used to predict Chl *a* in the testing dataset. This process was repeated 10 times. Each time the differences between each model's predicted Chl *a* (represented by the midpoint of the predicted interval) and the observed Chl *a* were used to calculate three criteria, introduced by Marcot et al. (2006), for assessing the model's predictive accuracy. These criteria include:

1. Sum of squared errors (SSE) - sum of the squared difference between the predicted and the observed;
2. Model accuracy – the percent of the total number of cases for which the actual intervals and predicted intervals are equal, measured by a confusion matrix. A confusion matrix is a square matrix with the number of rows (1,⋯,*i*,⋯,*I*) and columns (1,⋯*j*,⋯*J*) the same as the number of intervals of the variable. The (*i*,*j*) element of the matrix represents the number of observations with an observed interval *i* and predicted interval *j*; and
3. The area under the receiving operating characteristic curve (AUC) – the probability of a true positive outcome (the proportion of actual observations which are correctly classified) versus a false positive outcome (accuracy of data classification) (Bradley, 1997). A model with perfect predictions would have an AUC equal to 1.

3. Results

3.1. Marginal Chlorophyll a distribution comparison

Following, we highlight the discrepancy among the marginal

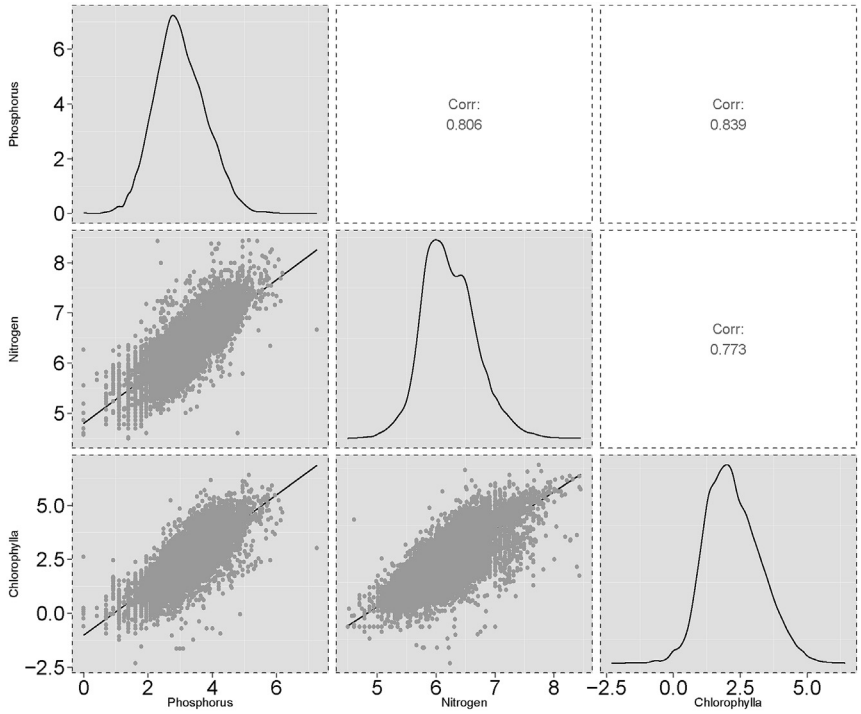

Fig. 2. Scatter plot matrix displays the Finnish lake data set of the variables chlorophyll *a*, nitrogen, and phosphorus in the logarithmic scale.

distributions that is the immediate result of discertization. We emphasize the differences among the marginal distributions as the BNs are developed based on the discretized data set. Hence, if we start with different data sets, the results would not be similar. Marginal distributions differ markedly among the three discretization methods, even when the same number of intervals is used for each method (Fig. 3). In no instance do the interval endpoints coincide (except for the Chl *a* extremes), and the proportion of observations within corresponding intervals differs considerably among methods. For example lower tail areas range from 3% (equal interval) to 40% (moment matching) for the three interval discretization, 0.4% (equal interval) to 26.2% (moment matching) for the four interval discretization, and 0.2% (equal interval) to 20% (equal quantile) for the five interval discretization. Middle intervals range from 33% (equal quantile) to 85% (equal interval) for the three interval discretization, and 20% (equal quantile) to 61.8% (equal interval) for the five interval discretization. Generally, the symmetry and narrow tails of the empirical distribution (Fig. 2) is reflected in the equal interval method, but is poorly captured in the equal quantile approach. The moment matching method tends to better reflect the empirical distribution as the number of intervals increases from three to five. Similar discrepancies are apparent when comparing the empirical distributions of phosphorus and nitrogen from the observed data (Fig. 2) to discretized distributions depicted in supporting material (Nojavan et al., 2015). As an example, probabilities of phosphorus in low, medium, and high change significantly among discretization methods (Figures in supplementary material (Nojavan et al., 2015)).

3.2. Conditional Chlorophyll a distribution comparison

Several examples of pronounced differences in the conditional probabilities of Chl *a* among the three discretization methods, given the states of parent nodes N and P, are highlighted in CPTs for the three-interval discretization (Tables 1–3). When N is “low” and P is “high” the equal interval method indicates that the probability of Chl *a* being “high” is 0.00 (Table 1), while the equal quantile method indicates that a high Chl *a* has a 0.18 probability (Table 2) and moment matching indicates a 0.43 probability (Table 3). Similarly, when N is “medium” and P is “low” the probabilities of Chl *a* being low are 0.03, 0.60, and 0.66 for equal interval (Table 1), equal quantile (Table 2), and moment matching (Table 3), respectively. And finally, when N and P are both “high”, the probability that Chl *a* is “low” is 0.00 for equal interval (Table 1), 0.54 for equal quantile (Table 2), and 0.76 for moment matching (Table 3), respectively.

We note that these discrepancies arise because the definitions of low, medium, and high, commonly used categories, differ among discretization methods resulting in a communication problem. For example, high Chl *a* is defined as concentrations greater than 3.51, 2.64, and 3.39 ($\mu g L^{-1}$ in log scale) using the equal interval, equal quantile, and moment matching methods, respectively (Fig. 3).

3.3. Prediction comparison

BNs discretized using different methods result in differing future predictions. We used the BNs generated from the training dataset (90% of the original data set) to predict the testing data set (the remaining 10% of the original data set). Confusion matrices summarize the predictive accuracy in terms of percent of correctly predicted observations (Table 4). Using the equal interval method, the resulting BN model predicts Chl *a* to be in “low”, “medium”, and “high” categories with probabilities 0%, 96%, and 4%, respectively. The model based on equal quantile method predicts Chl *a* to be in “low”, “medium”, and “high” categories with probabilities of 37%, 32%, and 31%, respectively. The moment matching model predicts Chl *a* to be in “low”, “medium”, and “high” categories with probabilities of 41%, 46%, and 13%, respectively. Our predictions change significantly from one method to the other. The discrepancy in predictions is initiated by different CPTs, as CPTs define the

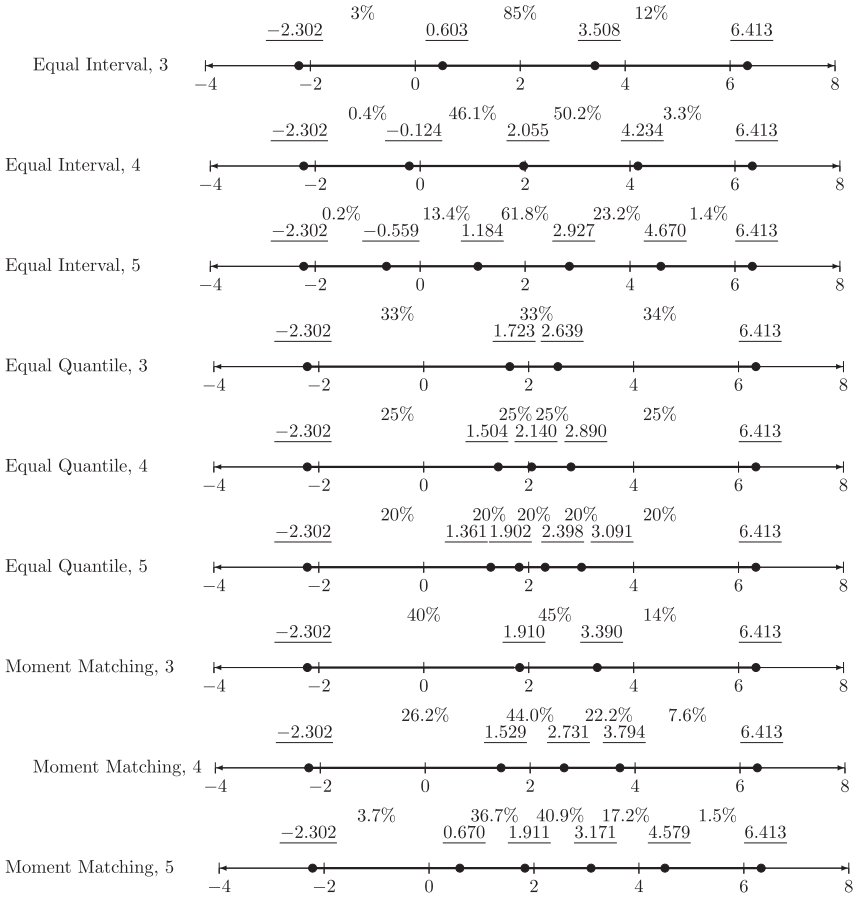

Fig. 3. The figure depicts original data for chlorophyll *a* concentrations from −2.302 to 6.411 discretized using the equal interval, equal quantile, and moment matching methods into three, four, and five intervals. The underlined numbers show the break points and the percentages show the frequency of observation in each interval.

Table 1
Conditional probability table for chlorophyll *a* node in the BN discretized using equal interval method. Each number represents the probability of chlorophyll *a* taking any of its discrete states, low (L), medium (M), and high (H), given the states of nitrogen and phosphorus. For example, the probability of log chlorophyll *a* concentrations being between 0.603 and 3.51 μg/L is 0.21 (the bold and underlined number) given that nitrogen concentrations are between 5.82 and 7.14 and phosphorus concentrations are between 4.83 and 7.24.

	Nitrogen / **Phosphorus**	L:[4.5,5.82] L:[0,2.41]	M:(2.41,4.83]	H:(4.83,7.24]
Chlorophyll *a*	L: [−2.3,0.603]	0.21	0.03	0.00
	M: (0.603,3.51]	0.79	0.97	1.00
	H: (3.51,6.41]	0.00	0.00	0.00
	Nitrogen / **Phosphorus**	L:[0,2.41]	M: (5.82,7.14] M:(2.41,4.83]	H:(4.83,7.24]
Chlorophyll *a*	L: [−2.3,0.603]	0.03	0.00	0.02
	M:(0.603,3.51]	0.97	0.86	**0.21**
	H:(3.51,6.41]	0.00	0.13	0.77
	Nitrogen / **Phosphorus**	L: L:[0,2.41]	M:(2.41,4.83]	H:(7.14,8.46] H:(4.83,7.24]
Chlorophyll *a*	L:[−2.3,0.603]	0.00	0.00	0.00
	M:(0.603,3.51]	1.00	0.29	0.11
	H:(3.51,6.41]	0.00	0.71	0.89

Table 2
Conditional probability table for chlorophyll *a* node in the BN discretized using equal quantile method. Each number represents the probability of chlorophyll *a* taking any of its discrete states, low (L), medium (M), and high (H), given the states of nitrogen and phosphorus. For example, the first number in the upper right of the table, 0.18, is the probability of chlorophyll *a* concentrations between −2.3 and 1.72 *μg/L* given that nitrogen concentrations are between 3.43 and 5.99 and phosphorus concentrations are between 3.3 and 7.24.

	Nitrogen / **Phosphorus**	L: [4.5,5.99] L: [0,2.64]	M: (2.64,3.3]	H: (3.3,7.24]
Chlorophyll *a*	L: [−2.3,1.72]	0.80	0.37	**0.18**
	M: (1.72,2.64]	0.20	0.57	0.57
	H: (2.64,6.41]	0.00	0.05	0.24
	Nitrogen / **Phosphorus**	L: [0,2.64]	M: (5.99,6.41] M: (2.64,3.3]	H: (3.3,7.24]
Chlorophyll *a*	L: [−2.3,1.72]	0.60	0.12	0.04
	M: (1.72,2.64]	0.39	0.67	0.38
	H: (2.64,6.41]	0.02	0.21	0.58
	Nitrogen / **Phosphorus**	L: [0,2.64]	M: (2.64,3.3]	H: (6.41,8.46] H: (3.3,7.24]
Chlorophyll *a*	L: [−2.3,1.72]	0.54	0.09	0.01
	M: (1.72,2.64]	0.43	0.64	0.12
	H: (2.64,6.41]	0.03	0.27	0.86

underlying relations among variables (Fig. 4). Further, the differing CPTs are the result of discertization methods and inconsistent definitions of “low”, “medium”, and “high” categories. We used multiple measures of performance to compare predictions for the testing data; no one discretization method outperformed the others consistently. Comparisons using SSE, model accuracy, and AUC as criteria offer no conclusive guidance that one method or number of break points consistently outperforms the others (see

Table 3
Conditional probability table for chlorophyll *a* node in the BN discretized using moment matching method. Each number represents the probability of chlorophyll *a* taking any of its discrete states, low (L), medium (M), and high (H), given the states of nitrogen and phosphorus. For example, the first number in the upper right of the table, 0.43, is the probability of chlorophyll *a* concentrations between −2.3 and 1.91 $\mu g/L$ given that nitrogen concentrations are between 4.5 and 6.12 and phosphorus concentrations are between 3.97 and 7.24.

	Nitrogen	L: [4.5,6.12]		
	Phosphorus	L: [0,2.77]	M: (2.77,3.97]	H: (3.97,7.24]
Chlorophyll *a*	L: [−2.3,1.91]	0.85	0.39	**0.43**
	M: (1.91,3.39]	0.15	0.60	0.36
	H: (3.39,6.41]	0.00	0.01	0.21
	Nitrogen		M: (6.12,7.02]	
	Phosphorus	L: [0,2.77]	M: (2.77,3.97]	H: (3.97,7.24]
Chlorophyll *a*	L: [−2.3,1.91]	0.66	0.09	0.01
	M:(1.91,3.39]	0.34	0.76	0.38
	H: (3.39,6.41]	0.00	0.15	0.61
	Nitrogen			H: (7.02,8.46]
	Phosphorus	L: [0,2.77]	M: (2.77,3.97]	H: (3.97,7.24]
Chlorophyll *a*	L:[−2.3,1.91]	0.76	0.09	0.02
	M:(1.91,3.39]	0.24	0.57	0.12
	H:(3.39,6.41]	0.00	0.34	0.86

Table 4
Confusion matrix for chlorophyll *a* in BN discretized using equal interval method and 3-interval. Each element of the matrix is the number of cases for which the actual interval is the row and the predicted interval is the column. The discrete states are low (L), medium (M), and high (H).

Equal Interval		**Predicted**		
		L: [−2.3,0.603]	M: (0.603,3.51]	H: (3.51,6.41]
Observed	L: [−2.3,0.603]	0	65	0
	M: (0.603,3.51]	0	1637	17
	H: (3.51,6.41]	0	165	59
Equal Quantile		**Predicted**		
		L: [−2.3,1.72]	M: (1.72,2.64]	H: (2.64,6.41]
Observed	L: [−2.3,1.72]	533	105	6
	M: (1.72,2.64]	180	389	107
	H: (2.64,6.41]	11	121	491
Moment Matching		**Predicted**		
		L: [−2.3,2.56]	M: (2.56,3.37]	H: (3.37,6.41]
Observed	L: [−2.3,1.91]	642	155	1
	M: (1.91,3.39]	154	657	71
	H: (3.39,6.41]	2	90	171

Marcot et al. (2006) and Subsection 2.3); generally, the differences are small. The results are summarized in Table 5. Because the data set is large (n = 19,248), the model is simple, the relationships among the variables are fairly straightforward, and each model has been optimally fitted, goodness-of-fit measures may not be a good basis to differentiate models (Qian and Cuffney, 2012). However, the models differ significantly in the resulting relations and CPTs as well as their management implications.

3.4. Management applications

BNs are tools for managers and policy makers to evaluate the impact of their pending decisions/policies on an ecosystem prior to implementation. Consider a case where the policy makers are assessing the impact of lowering phosphorus on Chl *a* concentrations.

Summarized results of low phosphorus on Chl *a*, discretized using different methods, point to very different conclusions (Table 6). The BN discretized using the equal interval method does not conclude that lowering phosphorus is effective in lowering Chla *a*, while the BN discretized using the equal quantile approach finds that lowering phosphorus would result in low Chl *a* concentrations 66% of the time. As another case consider managers targeting policies that would result in low/medium Chl *a* concentrations (avoid high Chl *a* concentrations). While the BNs discretized with equal quantile and moment matching methods recommend low to medium phosphorus and nitrogen concentrations, the BN discretized using equal interval recommends medium nitrogen and phosphorus concentrations (see Table 7 for a summary of results).

4. Discussion

Our goal in this paper was to investigate whether alternative discretization methods affect the conclusions reached using the resultant BNs. Discrepancies among the BNs are expected, as each method results in a different categorization of the predictor and response variables. However, the differences can be masked by commonly used category names such as “low”, “medium”, and “high”. Consequently, two BNs developed based on different discretizing methods may have the same structure but the meaning of the categories, as well as the meaning of conditional probabilities, will differ. Likewise, the predictions and decisions made based on the BNs developed using different discretizing methods are expected to differ because they are models with different meanings. We illustrated the differences by comparing nine BNs developed using three commonly used discretizing method and three commonly used number of break points in this study. The resulting CPTs changed from one method of discretization to the other. The CPTs provide the basis and define the relations in a BN; consequently, any calculation based on them would be different. The predictions were different among methods. Management recommendations were also different among the developed BNs (subsection 3.4). Hence, we call for a careful consideration of the underlying model and recognizing the limit of the BN imposed by the currently available software. BN models developed using different discretization methods should be considered and compared. If the results of such models are different, the choice of one method over the other should be justifiable. An alternative approach would be developing continuous BN models using Markov Chain Monte Carlo.

The example we used has a simple structure and we can easily understand the differences among the discretization methods - the differences are a result of different definitions of the categories. As discretization represents a simplification of the underlying functional relations, the optimal discretizing method is the one that can preserve the underlying functional relations. When the underlying relations are unknown, we expect that the optimal method is also unknown. In our example, we did not find a consistently best discretization method using three commonly used model comparison criteria. The equal interval method performs relatively well on all comparison criteria (Table 5); however, it does that by classifying most of the new data into “medium” category where most of the data lie (Table 4). Hence, other model performance criteria such as balanced accuracy and class-specific recall might be better measures of model goodness.

The deliberately simple structure, which includes only three variables, and is supported by a large data set, also indicates that the problems we have documented do not arise principally because many of the conditional probability intervals are data-sparse. As the number of variables and intervals increase in a BN, the size of the data set required to develop the BN also increases. Thus, we would expect more pronounced discrepancies among more complex models developed from limited data.

When results from a BN model are communicated with a

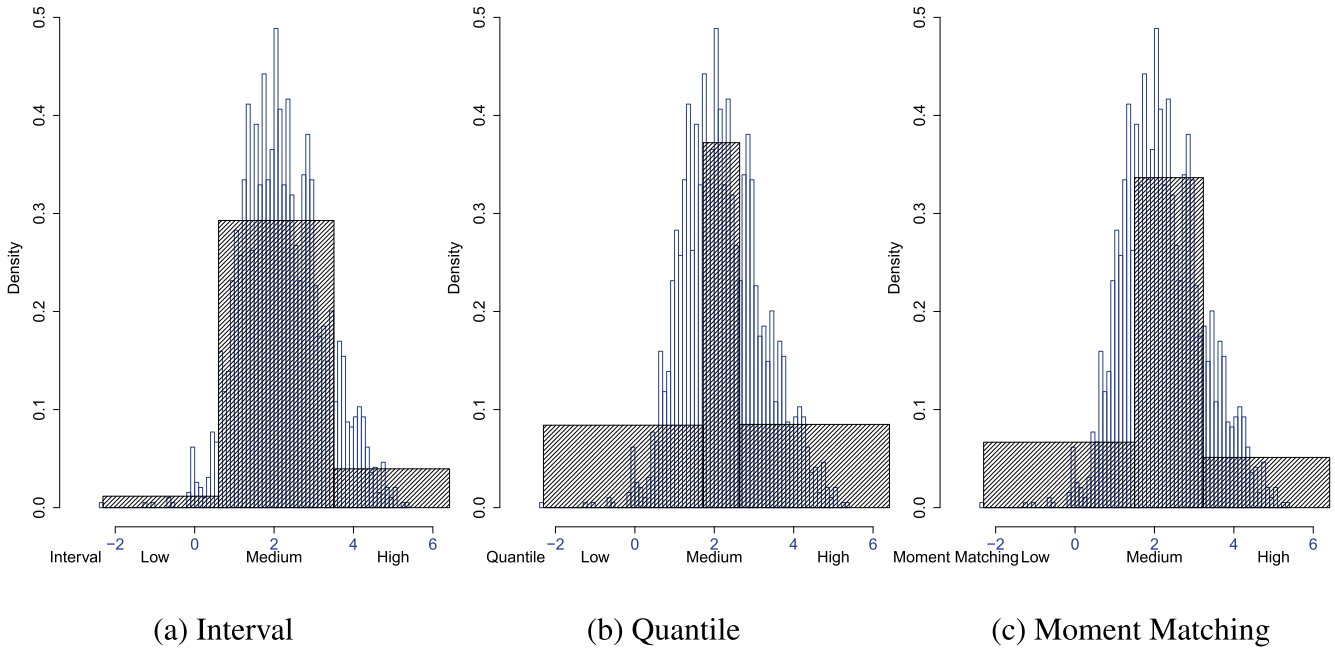

Fig. 4. The histogram shows the test data in continuous form. The overlaid cross-hatched histogram shows the model predictions for the test data in discrete form using the BNs developed by data discretized using equal interval (a), equal quantile (b), and the Moment Matching (c) methods, respectively. Comparison of continuous histogram with the predicted discretized shows the discrepancy between the observed and predicted in the test data. Comparison of histograms across the row shows the discrepancy between predictions due to method of discretization.

Table 5
Comparison of predictive accuracy among different discretization methods using SSE, Accuracy, and AUC as criteria with 3 intervals and 5 intervals.

	SSE	Accuracy	AUC
3 Intervals			
Equal Interval	1818.545	0.871	0.844
Equal Quantile	3681.574	0.728	0.824
Moment Matching	3228.382	0.751	0.807
4 Intervals			
Equal Interval	1962.192	0.700	0.803
Equal Quantile	2852.654	0.624	0.772
Moment Matching	2480.924	0.621	0.765
5 Intervals			
Equal Interval	1325.392	0.698	0.902
Equal Quantile	2141.934	0.549	0.725
Moment Matching	818.166	0.702	0.843

Table 6
Probability Table for Chlorophyll *a* under low phosphorus scenario for models discretized using three different methods.

	Chlorophyll *a*		
Method	Low	Medium	High
Equal Interval	0.07	0.93	0.00
Equal Quantile	0.66	0.32	0.02
Moment Matching	0.73	0.26	0.01

Table 7
Probability Table for phosphorus and nitrogen under a scenario where chlorophyll *a* concentrations do not exceed medium.

	Phosphorus			Nitrogen		
Method	Low	Medium	High	Low	Medium	High
Equal Interval	0.27	0.72	0.01	0.19	0.79	0.02
Equal Quantile	0.46	0.36	0.18	0.41	0.33	0.26
Moment Matching	0.44	0.49	0.07	0.48	0.48	0.04

manager, the underlying definition of categories can be overlooked. What is considered “high”, for example, depends on experience and an understanding of the system in question. High nutrient concentrations for one group of managers could be “medium” to another group. Examples of such differences can be found in a study by Kashuba et al. (2010), where 11% urbanization in a watershed is low in Southeast US but very high in state of Maine in north east US. As a result, we consider the discrepancy in the management recommendations as a major weakness of BNs based on discretized variables. Further, frequently used discretization methods embedded in BN software (e.g. equal interval and equal quantile) do not discretize data into intervals that are necessarily of interest to managers. For example, managers might be interested in predicting Chl *a* concentrations higher than a relevant water quality standard (e.g., 40 $\mu g/L$ in many states in the US), but the default discretization methods in BN software might not lead to a break point at the value of interest. As discussed in the results, the BN discretized using equal interval did not find the lowering phosphorus as effective as did the BN models based on the other two discretization methods. If the BN discretized with equal intervals was used to provide recommendations, then lowering phosphorus would be considered not cost-effective, whereas, this is only the result of discretization. We would caution against decisions based on models for which the outputs vary by the choice (of discretization method) that does not have justifiable scientific basis.

5. Conclusions

BNs are effective, valuable tools to quantify uncertainty in environmental modeling. We highlighted the main drawback of the BNs (discretization of continuous variables) and argued that discretization of continuous variables should be avoided, if possible. However, most currently available BN software requires discretization. Hence, when discretization is necessary, its consequences should be carefully evaluated. The results of different discretization schemes should be compared and used to learn more

about the system under study. Future work should focus on developing BNs software that can accommodate continuous variables with flexible functional forms, for example, by using the Gibbs Sampler as was done in Qian and Miltner (2015).

Acknowledgements

We would like to thank Elizabeth Albright and Marco Marani for constructive reviews of earlier versions of this manuscript. We would also like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the manuscript.

References

Aguilera, P., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environ. Model. Softw. 26 (12), 1376–1388.

Barton, D., Saloranta, T., Moe, S., Eggestad, H., Kuikka, S., 2008. Bayesian belief networks as a meta-modelling tool in integrated river basin management-pros and cons in evaluating nutrient abatement decisions under uncertainty in a norwegian river basin. Ecol. Econ. 66 (1), 91–104.

Borsuk, M.E., Stow, C.A., Reckhow, K.H., 2004. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecol. Model. 173 (2), 219–239.

Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognit. 30 (7), 1145–1159.

Bromley, J., Jackson, N.A., Clymer, O., Giacomello, A.M., Jensen, F.V., 2005. The use of hugin to develop bayesian networks as an aid to integrated water resource planning. Environ. Model. Softw. 20 (2), 231–242.

Castelletti, A., Soncini-Sessa, R., 2007. Bayesian networks and participatory modelling in water resource management. Environ. Model. Softw. 22 (8), 1075–1088.

Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environ. Model. Softw. 37, 134–145.

Croke, B., Ticehurst, J., Letcher, R., Norton, J., Newham, L., Jakeman, A., 2007. Integrated assessment of water resources: Australian experiences. Water Resour. Manag. 21 (1), 351–373.

Death, R.G., Death, F., Stubbington, R., Joy, M.K., Belt, M., 2015. How good are bayesian belief networks for environmental management? a test with data from an agricultural river catchment. Freshw. Biol. 60 (11), 2297–2309.

Dillon, P., Rigler, F., 1974. The phosphorus-chlorophyll relationship in lakes. Limnol. Oceanogr. 19 (5), 767–773.

Dorner, S., Shi, J., Swayne, D., 2007. Multi-objective modelling and decision support using a bayesian network approximation to a non-point source pollution model. Environ. Model. Softw. 22 (2), 211–222.

Farmani, R., Henriksen, H.J., Savic, D., 2009. An evolutionary bayesian belief network methodology for optimum management of groundwater contamination. Environ. Model. Softw. 24 (3), 303–310.

Jensen, F.V., 2001. Bayesian Networks and Decision Graphs. Springer, New York.

Johnson, S., Fielding, F., Hamilton, G., Mengersen, K., 2010. An integrated Bayesian network approach to Lyngbya majuscula bloom initiation. Mar. Environ. Res. 69 (1), 27–37.

Kashuba, R., Cha, Y., Alameddine, I., Lee, B., Cuffney, T.F., 2010. Multilevel Hierarchical Modeling of Benthic Macroinvertebrate Responses to Urbanization in Nine Metropolitan Regions across the Conterminous United States. Technical report. U. S. Geological Survey.

Kelly, R.A., Jakeman, A.J., Barreteau, O., Borsuk, M.E., ElSawah, S., Hamilton, S.H., Henriksen, H.J., Kuikka, S., Maier, H.R., Rizzoli, A.E., et al., 2013. Selecting among five common modelling approaches for integrated environmental assessment and management. Environ. Model. Softw. 47, 159–181.

Koch, A.L., 1966. The logarithm in biology 1. mechanisms generating the log-normal distribution exactly. J. Theor. Biol. 12 (2), 276–290.

Kragt, M., Newham, L.T., Bennett, J., Jakeman, A.J., 2011. An integrated approach to linking economic valuation and catchment modelling. Environ. Model. Softw. 26 (1), 92–102.

Malekmohammadi, B., Kerachian, R., Zahraie, B., 2009. Developing monthly operating rules for a cascade system of reservoirs: application of bayesian networks. Environ. Model. Softw. 24 (12), 1420–1432.

Malve, O., Qian, S.S., 2006. Estimating nutrients and chlorophyll a relationships in finnish lakes. Environ. Sci. Technol. 40 (24), 7848–7853.

Marcot, B.G., Steventon, J.D., Sutherland, G.D., McCann, R.K., 2006. Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. Can. J. For. Res. 36 (12), 3063–3074.

McCann, R.K., Marcot, B.G., Ellis, R., 2006. Bayesian belief networks: applications in ecology and natural resource management. Can. J. For. Res. 36 (12), 3053–3062.

Nagarajan, R., Scutari, M., Lebre, S., 2013. Bayesian Networks in R with Applications in Systems Biology. Springer, New York. ISBN 978–1461464457.

Nojavan, A.F., Qian, S.S., Stow, C.A., 2015. Supplementary material for "Comparative analysis of discretization methods in Bayesian networks" http://dx.doi.org/10.5281/zenodo.35174.

Ott, W.R., 1990. A physical explanation of the lognormality of pollutant concentrations. J. Air & Waste Manag. Assoc. 40 (10), 1378–1383.

Pearl, J., 1982. Reverend bayes on inference engines: a distributed hierarchical approach. In: AAAI, pp. 133–136.

Pollino, C.A., Woodberry, O., Nicholson, A., Korb, K., Hart, B.T., 2007. Parameterisation and evaluation of a bayesian network for use in an ecological risk assessment. Environ. Model. Softw. 22 (8), 1140–1152.

Qian, S.S., Cuffney, T.F., 2012. To threshold or not to threshold? that's the question. Ecol. Indic. 15 (1), 1–9.

Qian, S.S., Miltner, R., 2015. 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. Environ. Model. Softw. 69, 14–22. July 2015.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reckhow, K.H., 1999. Water quality prediction and probability network models. Can. J. Fish. Aquatic Sci. 56 (7), 1150–1158.

Scutari, M., 2010. Learning bayesian networks with the bnlearn r package. J. Stat. Softw. 35 (3), 1–22.

Smith, V.H., 1982. The nitrogen and phosphorus dependence of algal biomass in lakes: an empirical and theoretical analysis. Limnol. Oceanogr. 27 (6), 1101–1112.

Spiegelhalter, D.J., Knill-Jones, R.P., 1984. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. J. R. Stat. Soc. Ser. A General. 35–77.

USEPA, 2009. National Lakes Assessment: a Collaborative Survey of the Nation's Lakes. Technical Report EPA 841-R-09–001. U.S. Environmental Protection Agency (USEPA), U.S. Environmental Protection Agency, Office of Water and Office of Research and Development, Washington, D.C.

Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model. 203 (3), 312–318.

Varis, O., 1997. Bayesian decision analysis for environmental and resource management. Environ. Model. Softw. 12 (2), 177–185.

Varis, O., Kuikka, S., 1997. Joint use of multiple environmental assessment models by a bayesian meta-model: the baltic salmon case. Ecol. Model. 102 (2), 341–351.