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

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).

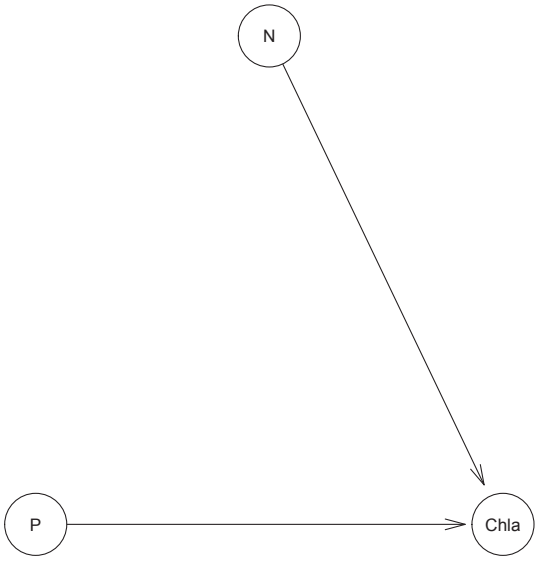

Fig. 1. Directed acyclic graph.

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.