Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation xxx (xxxx) xxx
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original Articles

Use of a structure aware discretisation algorithm for Bayesian networks applied to water quality predictions

Helen J. Mayfield[a,*], Edoardo Bertone[b], Carl Smith[c], Oz Sahin[a,b,d]

[a] *Cities Research Institute, Griffith University, Queensland, Australia*
[b] *School of Engineering and Built Environment, Griffith University, Queensland, Australia*
[c] *UQ Business School, The University of Queensland, Queensland, Australia*
[d] *Griffith Climate Change Response Program, Griffith University, Queensland, Australia*

Received 16 May 2018; received in revised form 11 July 2019; accepted 14 July 2019
Available online xxxx

Abstract

Bayesian networks have become a popular modelling technique in many fields, however there are several design decisions that, if poorly made, can result in models with insufficient evidence to make good predictions. One such decision is how to discretise the continuous nodes. The lack of a commonly accepted algorithm for achieving this makes it a difficult task for novice data modellers. We present a structure aware discretisation algorithm that minimises the number of missing values in the conditional probability tables by taking into account the network structure. It also prevents users from having to specify the exact number of bins. Results from two water quality case studies in south-east Queensland showed that the algorithm has potential to improve the discretisation process over equal case discretisation and demonstrates the suitability of Bayesian networks for this field.

Keywords: Bayesian networks; Structure aware discretisation; Water treatment optimisation

1. Introduction

Bayesian Networks (BNs) are a type of probabilistic decision support tool [14], and are widely used in a variety of applications ranging from environmental management and ecology [7,20,21,27] to medicine [17,26]. While BNs offer a range of advantages such as the ability to deal with missing data [13], explicitly model interactions between the predictors [17], and incorporate expert opinions into the model parameters [14], there are several decisions that can affect model performance and increase the chance of spurious predictions and limit the usefulness of the technique to novice practitioners. Two of the key decisions are the structure of the network and the discretisation of continuous nodes.

BNs consist of a graphical component and an underlying data structure. The graphical component of a BN is a directed acyclic graph (DAG), which represents variables as nodes, and dependencies between variables as edges (or links). Nodes for continuous variables are often discretised into a number of states (or bins). Edges are directed

[*] Corresponding author.
E-mail address: helenmayfield@warpmail.net (H.J. Mayfield).

https://doi.org/10.1016/j.matcom.2019.07.005
0378-4754/© 2019 International Association for Mathematics and Computers in Simulation (IMACS). Published by Elsevier B.V. All rights reserved.

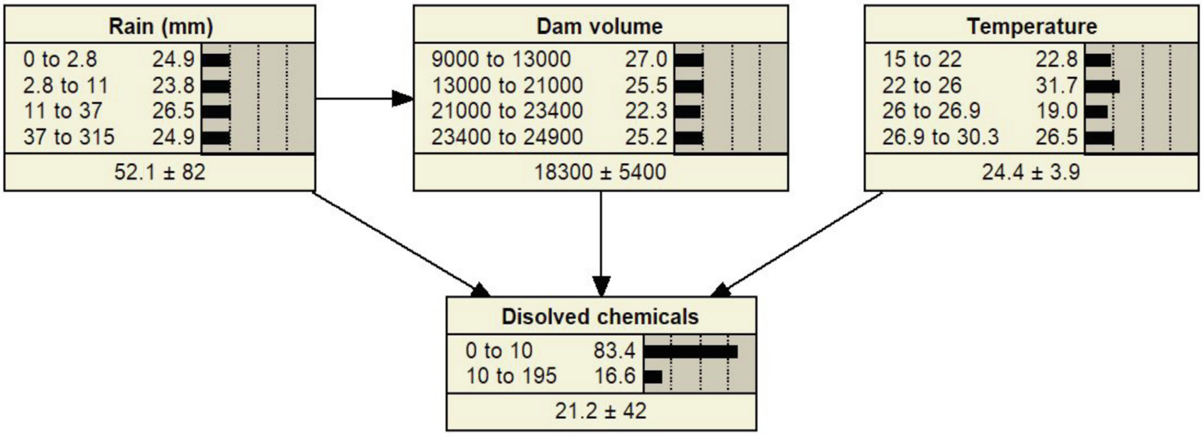

Fig. 1. An example BN for predicting dissolved chemicals (child node) based on rainfall, dam volume and temperature (parent nodes). There is also a causal link from rainfall (parent node) to dam volume (child node).

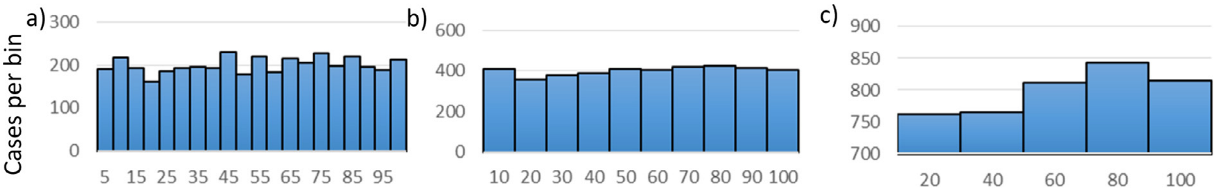

Fig. 2. Frequency distributions after discretisation of 4000 randomly generated numbers between 1 and 100 with (a) 20, (b) 10 and (c) 5 bins.

from the predictor variable (referred to as a parent node) to the dependent variable (referred to as the child node). An edge between a parent node and a child node can be interpreted as either full or partial causality, correlation or a functional relationship, depending on the underlying context. An example is shown in Fig. 1.

The data component of a BN takes the form of probability tables. For nodes without parents, these tables store unconditional probabilities, whereas nodes linked to parent nodes instead have conditional probability tables (CPTs). CPTs provide the probability of the child node being in a given state for any combination of parent node states. The size of the CPT is affected by the number of parent nodes as well as the number of bins in each node. While discretising continuous variables is not a theoretical requirement of BNs [24] it is a requirement of many popular BN software packages. However, discretisation of variables results in information loss, and if designed poorly can also reduce the predictive performance of a model [13]. Fig. 2 demonstrates alternative options for discretising 4,000 randomly generated numbers based on differing numbers of bins.

Datasets discretised to fewer bins are less able to represent the true distribution of the data [16], however, when applied to BNs, discretising a node into a large number of bins may require substantial (often unavailable) amounts of data to sufficiently complete the CPTs. Having either too few, or too many bins, can therefore lead to reduced predictive performance [12,19,25]. The trade-off between parsimony and precision is discussed in Marcot et al. [20] and Ratnapinda and Druzdzel [25] who noted that increasing the number of bins, while better describing the data distributions, does not automatically result in more accurate models.

Commonly used algorithms for discretising nodes within BNs include equal cases discretisation (ECD), which allocates an equal number of cases to each bin, equal range discretisation, which allocate an equal range of values to each bin, and statistical discretisation techniques such as splitting data based on quantiles or percentiles [25]. Other techniques include using expert opinion or cut-offs based on values of interest, such as ecological thresholds [18]. Importantly, each of these methods considers each node individually, ignoring the structure of the network and the resulting CPT composition and completeness. Excessive numbers of empty CPT values (where insufficient or no data is available to learn probabilities from) will prevent the network from generating accurate predictions.

No single technique has been proven as the go-to algorithm in every circumstance [30]. While some comparison studies have found that different methods on the same datasets yield similar results [18], others have found that the results differ substantially based on the method chosen [22]. Selecting an appropriate discretisation technique, as well as a suitable number of bins for each node is therefore essential for a well parameterised model. Although guidance is available [19,20], and there is some evidence indicating that the most suitable method should be selected based on characteristics of the data [22], making these decisions can still be a daunting task for those designing the network.

We propose that a good discretisation algorithm, if based on the data, should be flexible regarding the number of bins assigned, and the cut-off values for each bin. We further propose that to reduce unreliable results that may arise from missing values in the CPTs, the structure of the network should also be taken into account. In this paper we put forward a discretisation algorithm that accounts for both the distribution of the data and the structure of the network to select a suitable set of bins for each individual context. We evaluate this approach by applying a structure aware discretisation (SAD) algorithm to two water quality case studies, allowing an evaluation of both the algorithm itself and the general applicability of BNs to the scenarios presented in the case studies.

2. Materials and methods

2.1. Structure aware discretisation algorithm development

The objective of the structure aware discretisation (SAD) algorithm, implemented in R [1], is to ensure CPTs are as complete as possible by allowing flexibility in the final number of bins and the cut-off for each. The aim is to reach a compromise between having fewer empty CPT values, and maintaining a sufficient number of bins to reasonably model the distribution of each node. The algorithm consists of two stages (Fig. 3). After an initial discretisation round that reduces the number of bins until each contains a specified minimum number of cases, a second discretisation round is applied that also considers the structure of the BN. This round further reduces the number of bins to minimise the CPT combinations with insufficient cases.

In stage one, users specify the minimum number of cases to be allocated to each bin, as well as maximum number of allowed bins per node. For stage two, users additionally specify the minimum number of cases required in each CPT cell. Fixed nodes can also be defined with fixed, user-defined bins. Fixed nodes allow for bin values to be either specified by experts or based on relevant thresholds for the specific system being modelled (e.g. values from drinking water guidelines).

In stage one, continuous nodes (other than fixed nodes) are first discretised into bins independently of the network structure. During this first, structure-unaware stage of discretisation (SUD), missing values are removed and the remaining data for each node is split into a large number of temporary bins, each containing an equal number of cases. This initial number of cases is calculated by dividing the total number of sample points in the data by the number of temporary bins. If, after allocating the cases equally amongst the temporary bins, two bins have the same maximum value (which occurs if, for example, there are 20 cases with the same value and each bin has five cases), these are merged into a single bin. The algorithm then checks each bin, merging any containing fewer than the user-specified minimum number of cases. Merged bins are combined with either the preceding or following bin, depending on which of the two has the fewest cases. If the first or last bin does not contain the minimum number of cases, it is merged with the second or penultimate bin respectively. This process continues until each bin in the node contains the minimum number of cases. Bins are merged in order, starting from the one containing the fewest number of cases.

In the second, structure aware stage (SAD), the CPT of each node is checked for incomplete cells (i.e. having one or more combinations with fewer than the specified minimum number of cases). Starting from the first node with an incomplete CPT (the order is user-specified), the algorithm calculates which node to compress (either that node or one of its parents) to reduce the number of incomplete cells. The node selected will be whichever one has the greatest number of bins and is editable (i.e. not fixed or categorical, and with more than the minimum bins). The algorithm compresses the first bin in that node whose range falls within the incomplete cell in the CPT. The next incomplete CPT is then processed in the same fashion, with only a single change being made to each CPT before moving onto the next. This process is repeated until either all CPTs are complete or no further mergers are possible. The full code is provided in Online supplementary file B.

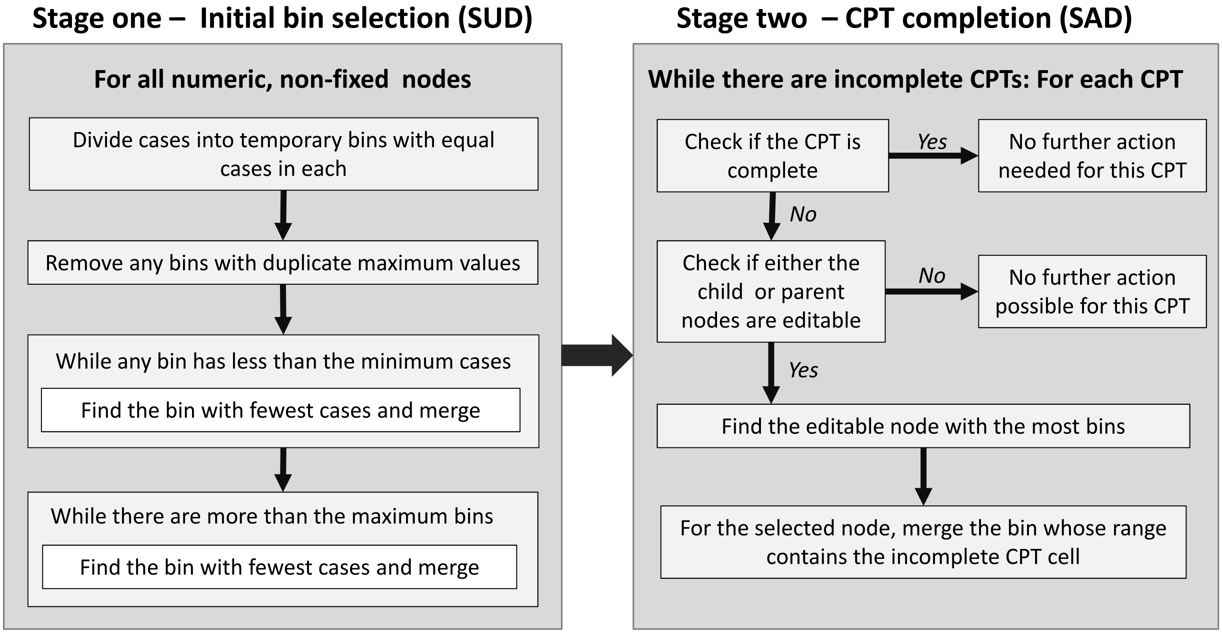
Stage one – Initial bin selection (SUD)
For all numeric, non-fixed nodes
Divide cases into temporary bins with equal cases in each
Remove any bins with duplicate maximum values
While any bin has less than the minimum cases
Find the bin with fewest cases and merge
While there are more than the maximum bins
Find the bin with fewest cases and merge
Stage two – CPT completion (SAD)
While there are incomplete CPTs: For each CPT
Check if the CPT is complete
Yes
No further action needed for this CPT
No
Check if either the child or parent nodes are editable
No
No further action possible for this CPT
Yes
Find the editable node with the most bins
For the selected node, merge the bin whose range contains the incomplete CPT cell

Fig. 3. Outline of the structurally unaware (Stage 1) and structurally aware (Stage 2) discretisation algorithms.

2.2. *Case study 1 — Prediction of taste and odour events in drinking water reservoirs*

Modelling taste and odour (T&O) events in drinking water reservoirs and water treatment plants (WTP) is important for water utility managers to help prevent and/or adequately cope with such events. The most common T&O compounds are geosmin and 2-Methylisoborneol (MIB). Modelling the production of these compounds is extremely challenging for a number of reasons. Firstly there are different potential sources, including algal blooms, vegetation, and standing timber [31]. Secondly, even if an algal bloom occurs, only specific species and strains can produce T&O compounds; with fewer than 3% of known cyanobacteria species confirmed as potential sources [2]. A key problem is therefore to predict which cyanobacteria strain will bloom. This is difficult, as each strain has different optimal growth conditions. Potential T&O predictors include, amongst others, water temperature, nutrients, chlorophyll-a, turbidity and light intensity [23,29,31]

High levels of geosmin and MIB have been recorded at the Capalaba WTP, which withdraws raw water from the adjoining Tingalpa Reservoir. Tingalpa Reservoir supplies water for the Redland region of South-East of Brisbane (Australia) and the WTP is owned and managed by the bulk water supplier Seqwater. The original capacity of 24,868 ML was reduced to 13,206 ML in 2014 because of safety reasons relating to the dam wall. A direct consequence of the dam lowering was a decline in water quality, with turbidity, nutrients and metals increasing after 2014 [6]. Between November 2015 and January 2016, geosmin levels peaked twice to above 160 ng/L, substantially higher than previous average peak concentrations of 15 ng/L [28].

Historical fortnightly water quality data from between 2011 and 2016 were made available through collaboration with Seqwater. Daily dam volume data were also collected. This was used to create a data-driven model to help understand peak geosmin events [6]. The model used a conceptual regression tree to provide an estimation of geosmin based on water temperature, cyanobacteria (*Microcystis aeruginosa* and *Dolichospermum circinale* only) and nitrogen [6]. Another data-driven model was later developed to estimate the amount of powdered activated carbon required to remove the resulting T&O [5].

Using this case study to evaluate the SAD algorithm allows an investigation of the potential to deploy an optimally discretised BN to predict T&O events in Capalaba WTP. The response variable (i.e. target node in the BN) was geosmin concentration; specifically, whether or not the geosmin levels were more than 10 ng/L. Among the predictors considered in this BN were cyanobacteria concentrations, nutrients (e.g. nitrogen, iron), turbidity and dam volume. The data sample contained 189 records, and the prevalence rate for a positive finding was 17%. Eight predictor variables were used.

2.3. Case study 2 — Forecasting lake manganese levels

A second case study was designed to compare an existing data-driven manganese (Mn) forecasting model presented in Bertone et al. [8] with BN models implemented using SAD. Advancetown lake, also known as Hinze dam, is the main source of potable water for the Gold Coast region of Australia. High soluble manganese concentrations in the water can cause aesthetic issues (in particular discolouration). Being able to understand the manganese cycle in the lake through data driven models can be highly beneficial, as it allows for these peak events to be predicted and proactively managed.

Seqwater aims to keep soluble manganese concentrations in the treated water lower than 0.02 mg/L [15]. If raw water concentrations exceed this threshold, pre-filter chlorination is required to oxidise the soluble manganese and remove it; if the concentrations are very high (above 0.18 mg/L), potassium permanganate is also added to help with manganese oxidation and removal. In Advancetown lake, a thermal stratification persists during most of the year, but is broken over winter [9], causing the water to circulate. During this time, mixing processes cause soluble manganese, usually confined in high concentrations in the deeper layers, to rise to the surface layer of the reservoir, where the water is typically drawn from [10].

Based on these dynamics, the main predictors for soluble manganese in the epilimnion (i.e. the more superficial layer) of Advancetown lake are water temperature, and potentially pH and dissolved oxygen. Each of these parameters is measured remotely and autonomously for the full water column every one to three hours by a vertical profiling system (VPS). Data for generating the models were sourced from these VPS records as well as historical local weather records as A VPS-based predictive model was developed to forecast soluble manganese concentrations one week ahead with acceptable accuracy [8].

Whereas the existing model predicts one specific value, a BN could potentially improve on the model outputs and associated uncertainty by estimating the chance of exceeding critical thresholds, such as 0.02 mg/L (i.e. use pre-filter chlorination) and 0.18 mg/L (i.e. use potassium permanganate). These are critical thresholds for treatment operations, and referencing the predictions to these could facilitate the deployment of the model. As well as the daily VPS data and weather data, additional inputs were also included for the BN, to check whether, optimising the discretisation process would help identify better correlations compared to the original model. This case study has 1190 samples and a prevalence rate of a positive finding of 10%. Six predictors were included and the target node was soluble manganese concentration in the epilimnion.

2.4. Evaluation

The SAD algorithm was compared with equal case discretisation (ECD) and SUD for both case studies. Three network structures were tested: a naïve BN in which all nodes have only the target node as a parent; a tree augmented BN (TAN) in which all nodes have the target node and at most one other node as a parent; and an expert-structured BN where links between nodes were determined by an expert with good knowledge of each case study. The structure for the TAN networks was learnt from the data using Netica [3]. For comparison, each structure was discretised three times, each allowing SAD to compress nodes down to a minimum of either 4, 8 or 15 bins per node. For ECD, this represented the exact number of bins generated. For SUD (stage one of SAD), this number is not used, as the minimum number of bins is instead determined by the distribution of the data and the minimum number of cases required in each bin. For SAD, reaching this minimum of bins triggers the algorithm to stop compressing the node further (fewer bins may be generated in stage one, in which case no further modifications are made to the node in stage two).

For case study one, the BNs were generated for a minimum of 10 cases per bin, with those discretised using SAD requiring a minimum of 4 cases per CPT cell. The second case study allowed for a minimum of 60 cases per bin and the SAD models further required a minimum of 5 cases per CPT cell. This represents a requirement for approximately 5% of cases in every bin. Both case studies allowed for a maximum of 20 bins at the end of the SUD round to avoid unwieldy CPTs being created in for the TAN and expert models. The primary metric used for evaluating model performance was the area under the receiver operating curve — AUC [11]. Values for the true skill statistic — TSS [4] are also provided in the online supplementary material. For TSS, the most probable outcome was selected (i.e. cases with a greater than 50% chance of having a geosmin value greater than 10 ng/L were considered as a positive result). Trial datasets were created using repeated random subsampling to create a

Table 1
Average number of bins per node generated by the SAD algorithm when set to a minimum of 4, 8 or 15 bins. Minimum and maximum values are shown in parenthesis. Note that the minimum number of bins generated in the first stage (SUD) may be less than the minimum specified for the second stage (SAD). In these instances nodes are not compressed further.

	Minimum bins	Naïve structure	TAN structure	Expert structure
	4	4.25 (4,5)	4.25 (4, 6)	4 (4, 4)
Case study 1 SUD Avg = 9 (min = 4, max = 12)	8	7.25 (4, 9)	7.25 (4, 9)	7.13 (4, 8)
	15	9 (4, 12)	9 (4, 12)	9 (4, 12)
	4	5.67 (4, 11)	4 (4, 4)	4 (4, 4)
Case study 2 SUD Avg = 11.67 (min = 7, max = 13)	8	8.33 (7, 11)	10.5 (7, 8)	10.5 (7, 8)
	15	11.67 (7, 13)	11.67 (7, 13)	11.67 (7, 13)

Table 2
Average AUC values over 20 trials for BNs discretised to a minimum of 4, 8 or 15 bins. Standard deviations given in parentheses. The number of minimum buckets does not affect the SUD algorithm.

		Case study 1			Case study 2		
		4	8	15	4	8	15
	ECD	0.86 (0.10)	0.91 (0.06)	0.9 (0.07)	0.94 (0.01)	0.97 (0.01)	0.97 (0.01)
TAN	SUD	0.83 (0.08)	0.83 (0.08)	0.83 (0.08)	0.98 (0.01)	0.98 (0.01)	0.98 (0.01)
	SAD	0.86 (0.08)	0.84 (0.08)	0.83 (0.08)	0.94 (0.01)	0.97 (0.01)	0.98 (0.01)
	ECD	0.88 (0.08)	0.9 (0.07)	0.92 (0.06)	0.93 (0.01)	0.96 (0.01)	0.97 (0.01)
Naïve	SUD	0.91 (0.05)	0.91 (0.05)	0.91 (0.05)	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)
	SAD	0.89 (0.07)	0.89 (0.07)	0.91 (0.05)	0.96 (0.01)	0.97 (0.01)	0.97 (0.01)
	ECD	0.79 (0.1)	0.82 (0.07)	0.83 (0.07)	0.94 (0.02)	0.93 (0.03)	0.87 (0.05)
Expert	SUD	0.88 (0.06)	0.88 (0.06)	0.88 (0.06)	0.92 (0.03)	0.92 (0.03)	0.92 (0.03)
	SAD	0.87 (0.06)	0.87 (0.05)	0.88 (0.06)	0.92 (0.02)	0.92 (0.03)	0.92 (0.03)

series of 20 training–testing data pairs. In each pair 75% of the data was used for training and 25% for testing. The Netica C application programming interface [3] was used to automate the trials and calculate metric values.

To compare the performance of the BNs against the original models, a selection of historical data was used for each case study. For case study 1 (geosmin prediction at the Capalaba WTP), a random, selection of historical data, including both threshold exceedances and low values, was used to compare predicted exceedance probabilities against actual measured data. This was done to provide a more practical test of the models and evaluate their potential as a useful geosmin prediction tool in support of water treatment decision-making.

For case study 2 (manganese levels in Advancetown lake), predictions from the original conceptual data-driven models were compared against a selection of the BNs generated in this study by comparing absolute difference between real Mn levels and the model predictions. As the original conceptual data-driven model provides a binary output (i.e. if Mn $>$ 0.02 mg/L then chance of being above 0.02 mg/L = 100%) the error will be either 0 or 100. For the BNs, the predicted probability of an event occurring (i.e Mn $>$ 0.02 mg/L) can be used. For example, if the BN predicts a 65% chance of an event which did not occur then the error would be 65%). The mean error for each of the models was then compared

3. Results

The number of bins generated by SAD for each structure is given in Table 1.

AUC results over the 20 trials are given in Table 2. While differences in model performance were generally small, using the SUD algorithm improved model performance in 10 of the 18 cases. The deployment of SAD improved the performance of SUD in 2 cases. Results for TSS are provided in the online supplementary material.

Sample bins derived from each algorithm using data from case study 1 and a TAN BN structure (in which each predictor node has the target node and at most one other node as parent) are shown in Fig. 4.

Of interest in Fig. 4 is that the values for *TotalmicroDoli* (i.e., the total amount of *Microcystis aeruginosa* and *Dolichospermum circinale* combined) are approximately the same for each of the three algorithms. This is due to the variable being highly skewed, with over 70% of the values having a value of 0.

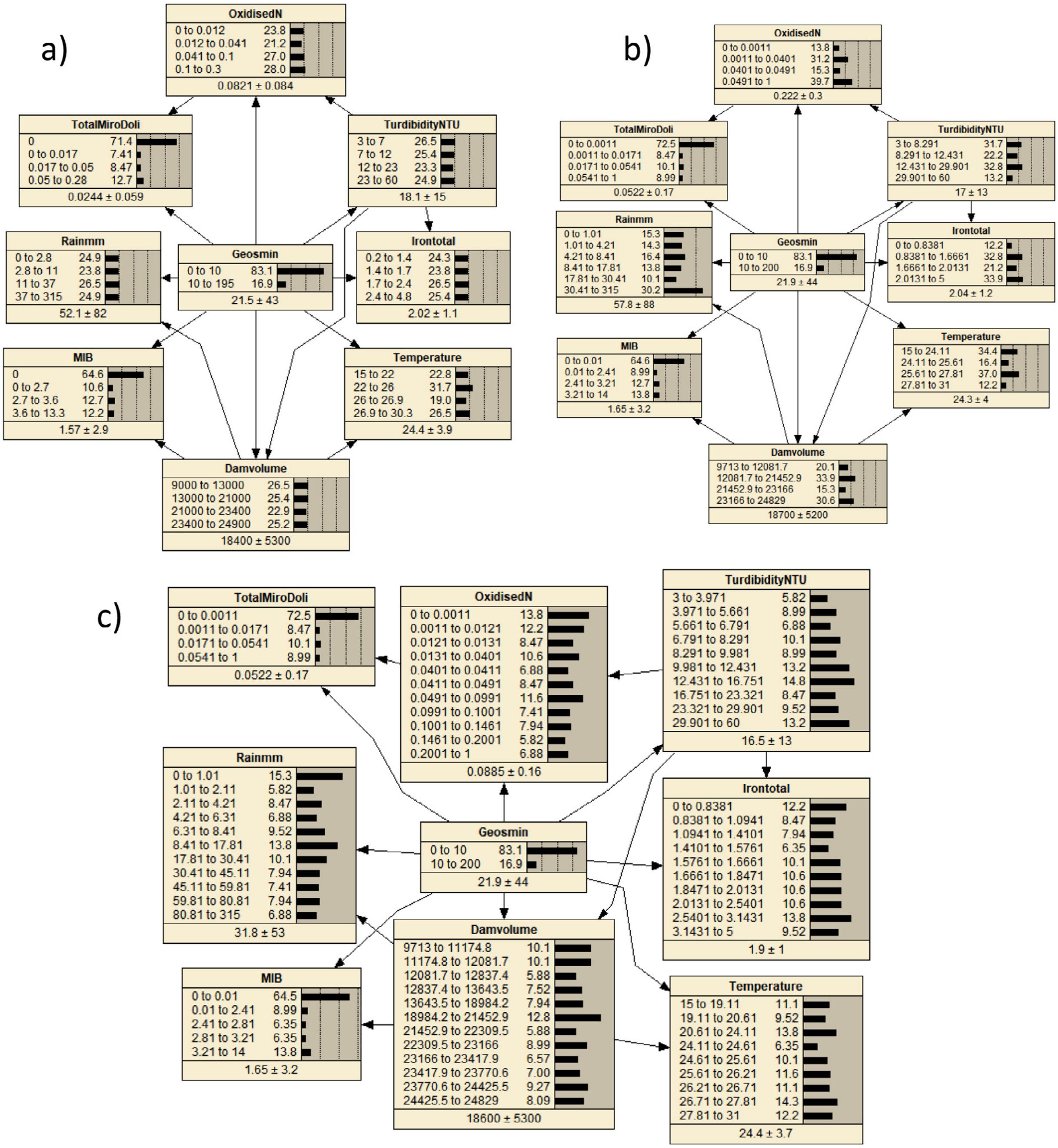

Fig. 4. Discretisation for a TAN BN resulting from (a) equal cases set to 4 bins, (b) structurally aware discretisation set to a minimum of 4 bins and (c) structurally unaware discretisation. SUD was set to a minimum of 10 cases per bin and SAD was set to 4 cases per CPT combination.

The predicted probabilities of geosmin exceeding 10 ng/L for several of the tested BNs are compared against real data in Table 3. Most of the numbers are very close to either 0% or 100%, with only the 1/3/2016 results showing mid-range probabilities, at least for the expert BNs. In general the BNs are extremely accurate and could be effectively used for prediction and proactive management of geosmin events.

Table 3
Case study 1: Probability of geosmin threshold exceedance based on historical scenarios for 9 BNs. The number next to the discretisation algorithm code indicates the maximum number of bins. SUD is not constrained to a maximum number of bins so it is marked as “X”. Instances where the 10 ng/L threshold were exceeded are shown in bold.

Date	Geosmin measured [ng/L]	P(geosmin > 10 ng/L)								
		TAN			Naïve			Expert		
		ECD15	SUDX	SAD15	ECD15	SUDX	SAD15	ECD4	SUDX	SAD4
8/9/11	0	16.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	3.3%
29/12/11	**22.2**	**96.4%**	**100.0%**	**100.0%**	**95.6%**	**4.1%**	**4.1%**	**8.3%**	**16.7%**	**9.1%**
1/3/12	4.8	3.0%	0.0%	0.0%	2.4%	0.0%	0.0%	23.5%	0.0%	0.0%
6/9/12	2.5	1.3%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	3.3%
4/2/14	2.9	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
10/11/15	**31.2**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
29/12/15	**169**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
31/12/15	**195**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
1/3/16	**30.7**	**95.1%**	**98.5%**	**98.5%**	**95.6%**	**99.9%**	**99.9%**	**23.5%**	**66.7%**	**42.9%**

Table 4
Comparison of absolute difference between real Mn levels and model predictions.

Year	Performance metric	Data-driven crisp prediction model	Expert SUD	Expert SAD
2015	Probability Mn < 0.02 mg/L	57.1%	23.7%	43.2%
2015	Probability Mn > 0.18 mg/L	5.0%	3.9%	4.2%
2015	Median crisp prediction	0.02 mg/L	0.04 mg/L	0.04 mg/L
2017	Probability Mn < 0.02 mg/L	45.5%	26.6%	43.1%
2017	Probability Mn > 0.18 mg/L	0%	0.0001%	7.3%
2017	Median crisp prediction	0.05 mg/L	0.022 mg/L	0.026 mg/L

The comparison of absolute difference between real Mn levels and model predictions for the original data-driven model and the expert structured SUD and SAD BNs (Table 4), show that while the BNs had a higher average error than the original models for the 2015 data, they had a lower error when predicting whether Mn levels would be above the critical threshold values. For both 2015 and 2017, BNs discretised with the SAD algorithm performed worse than those discretised using only the SUD algorithm.

4. Discussion

While differences were generally small, the SUD algorithm improved the AUC score compared to ECD in over half of the BNs tested. These results are consistent with those of Nojavan et al. (2017) [22] who found that no single discretisation method was uniformly ideal across all BNs. Even in cases where performance was not improved in every variation of the BN (i.e. four, eight or 15 bins), there are still advantages in using the SUD model presented here. For example, in the naive BN of case study 1, the ECD approach led to AUC values in between 0.88 and 0.92 (Table 2). However it is impossible for the user to know in advance how many bins are needed for each node. By applying the SUD algorithm in this instance, the average AUC was 0.91 without the user having to decide a priori how many bins should be defined. This reduces the need to run additional trials to determine a suitable number of bins for the model structure.

The lower than expected improvement from stage two of the algorithm (SAD), where only two of the 18 models showed improvement, is possibly a result of the small sample size of the dataset for the two case studies (189 and 1190 cases respectively). This may have inhibited SAD in two ways. Firstly, few cases meant that fewer bins were generated in stage one (SUD), leaving limited potential improvement for round two. For example, under the settings tested (minimum 10 points per bin) the SUD algorithm in case study 1 resulted in at most 12 bins for any node. This meant that when restricted to at least 15 bins, SAD was unable to make any further changes to the nodes. Secondly, fewer cases mean that CPTs are less likely to be complete after the initial stages, resulting in many nodes being reduced to the minimum number of bins.

In some instances, SAD continues to merge nodes without improving model parameterisation. For example, if two parent nodes are fixed, and some combination of these two nodes have insufficient cases, SAD would continue to

Please cite this article as: H.J. Mayfield, E. Bertone, C. Smith et al., Use of a structure aware discretisation algorithm for Bayesian networks applied to water quality predictions, Mathematics and Computers in Simulation (2019), https://doi.org/10.1016/j.matcom.2019.07.005.

compress the child node until the minimum number of bins is reached, despite it not being possible to complete the CPT. A number of improvements are planned in order to contain this effect, including ignoring CPT combinations for fixed nodes where completing the CPTs would be impossible, and allowing a small percentage of each CPT to remain incomplete.

The flexibility to define a range of bins, rather than an exact number is a major advantage of the SAD and SUD algorithms over the ECD algorithm, in that the user does not need to know *a priori* how many bins are required to represent the distribution of the data or how few bins are needed to ensure the network has sufficient data to learn from. Although some input parameters are required, the algorithms themselves determine the number of bins needed to meet these criteria, or to get as close as possible to them. SAD generally produces fewer bins to ensure that all node have sufficient cases in the CPT. Conversely, the SUD algorithm generally produces more bins (Table 1, Fig. 4), resulting in better model resolution. By taking the structure of the network into account, SAD is able to suggest bins that more closely match the data than ECD (by allowing more bins when data permits), while maintaining a compromise with the number of incomplete CPTs.

While both SUD and SAD algorithms allow initial flexibility in the number of bins represented in the model, users still need to specify the minimum number of cases in each bin for SUD and the minimum cases per CPT combination for SAD. There are several different ways for deciding these values, and the appropriate values will depend on the context in which the BN is being used. Where a high rate of confidence is required and large amounts of data are available, specifying the number of cases needed for 95% confidence based on the sample size may be appropriate.

For datasets with a low prevalence rate of target values (for example in BNs that are designed to assist in detecting rare events), more cases per bin may be needed to allow CPTs to be effectively parameterised. For example a dataset with a 10% prevalence rate would need at least 10 cases in each bin to be able to represent this rate. Where insufficient data exist to complete CPTs to the desired specifications, users may consider instead reducing the complexity of the network structures, which will reduce the size of the CPTs, and therefore the amount of data required to complete them. For exceptionally rare events, and for scenarios that have not occurred in the past, data will always be insufficient to empirically determine probabilities, so expert opinion (i.e. subjective probabilities) would still be required to complete CPTs.

The results found here indicate that BNs are a suitable modelling technique for water quality analysis, and several useful observations can be made in regard to this. When applied to a random set of historical data (Table 3) the TAN structure leads to more accurate predictions than the other two structures, with expert structured BNs being the least accurate. One plausible explanation for this is that in the expert structured models several key predictors, such as total iron, are not directly linked to the target node, which reduces their potential influence on the predictions. A sensitivity analysis on the naïve and TAN models showed total iron was the second most important input parameter. This shows the importance of subject experts being properly informed of the implications of model structure and being made aware of the key relationships in the data. It should also be noted here that although the expert models may be less accurate, they still have great value as causal models for exploring the dynamics of a system, compared to the purely data driven structures.

5. Conclusion

A new, structure aware discretisation algorithm was proposed for BNs, and tested in case studies related to water quality prediction. For the majority of the BNs evaluated, those discretised using SUD performed as well as, or better than, those using the ECD algorithm; however the distinction was less clear when compared to BNs discretised using SAD. These results show that the SAD algorithm (in particular the first stage) has the potential to maintain or improve BN performance compared to default algorithms such as ECD, whilst providing more flexibility in the design decisions by avoiding the need for the user to specify the exact number of bins. In many cases, BNs discretised with the new algorithms were also able to match or improve the predictive performance of the existing models presented in the case studies. Allowing users the flexibility to define a range of bins, rather than an exact number helps to overcome one of the major challenges faced by novice BN designers and helps to prevent the resulting BNs from generating spurious predictions.

Acknowledgements

The Authors would like to thank the Griffith University's Cities Research Institute for funding this project and Seqwater for sharing the data for the case studies presented in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.matcom.2019.07.005.

References

[1] R Core Team, R: A language and environment for statistical computing, in R Foundation for Statistical Computing, . 2017: Vienna, Austria.

[2] American Water Works Association, Algae: source to treatment. 57. 2011: American Water Works Association.

[3] Norsys Software Corp, Netica Bayesian Belief Network software. 2013: https://www.norsys.com.

[4] O. Allouche, A. Tsoar, R. Kadmon, Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS), J. Appl. Ecol. 43 (2006) 1223–1232.

[5] E. Bertone, C. Chang, P. Thiel, K. O'Halloran, Analysis and modelling of powdered activated carbon dosing for taste and odour removal, Water Res. 139 (2018) 321–328.

[6] E. Bertone, K. O'Halloran, Analysis and modelling of taste and odour events in a shallow subtropical reservoir, Environments 3 (3) (2016) 22.

[7] E. Bertone, O. Sahin, R. Richards, A. Roiko, Extreme events, water quality and health: A participatory Bayesian risk assessment tool for managers of reservoirs, J. Cleaner Prod. 135 (2016) 657–667.

[8] E. Bertone, R.A. Stewart, H. Zhang, M. Bartkow, C. Hacker, An autonomous decision support system for manganese forecasting in subtropical water reservoirs, Environ. Model. Softw. 73 (2015) 133–147.

[9] E. Bertone, R.A. Stewart, H. Zhang, K. O'Halloran, Analysis of the mixing processes in the subtropical Advancetown Lake, Australia, J. Hydrol. 522 (2015) 67–79.

[10] E. Bertone, R.A. Stewart, H. Zhang, K. O'Halloran, Statistical analysis and modelling of the manganese cycle in the subtropical Advancetown Lake, Australia, J. Hydrol.: Reg. Stud. 8 (2016) 69–81.

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

[12] I. Celik, U. Ghia, P.J. Roache, C.J. Freitas, H. Coleman, P.E. Raad, Procedure for estimation and reporting of uncertainty due to discretization in CFD applications, J. Fluids Eng.-Trans. ASME 130 (7) (2008).

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

[14] N. Fenton, M. Neil, Risk Assessment and Decision Analysis with Bayesian Networks, CRC Press, New York, 2013.

[15] P.M. Kohl, S.J. Medlar, Occurrence of Manganese in Drinking Water and Manganese Control, American Water Works Association, 2006.

[16] D. Landuyt, S. Broekx, R. D'Hondt, G. Engelen, J. Aertsens, P.L.M. Goethals, A review of Bayesian belief networks in ecosystem service modelling, Environ. Model. Softw. 46 (2013) 1–11.

[17] C.L. Lau, H.J. Mayfield, J.H Lowry, C.H. Watson, M. Kama, E.J. Nilles, C.S. Smith, Unravelling infectious disease eco-epidemiology using Bayesian networks and scenario analysis: A case study of leptospirosis in Fiji, Environ. Model. Softw. 97 (Supplement C) (2017) 271–286.

[18] P. Lucena-Moya, R. Brawata, J. Kath, E. Harrison, S. ElSawah, F. Dyer, Discretization of continuous predictor variables in Bayesian networks, Environ. Model. Softw. 66 (C) (2015) 36–45.

[19] B.G. Marcot, Common quandaries and their practical solutions in Bayesian network modeling, Ecol. Model. 358 (2017) 1–9.

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

[21] H. Mayfield, C. Smith, M. Gallagher, M. Hockings, Use of freely available datasets and machine learning methods in predicting deforestation, Environ. Model. Softw. 87 (2017) 17–28.

[22] A.F. Nojavan, S.S. Qian, C.A. Stow, Comparative analysis of discretization methods in Bayesian networks, Environ. Model. Softw. 87 (2017) 64–71.

[23] J. Parinet, M.J. Rodriguez, J.-B. Sérodes, Modelling geosmin concentrations in three sources of raw water in Quebec, Canada, Environ. Monit. Assess. 185 (1) (2013) 95–111.

[24] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 2014.

[25] P. Ratnapinda, M.J. Druzdzel, Learning discrete Bayesian network parameters from continuous data streams: What is the best strategy? J. Appl. Log. 13 (4, Part 2) (2015) 628–642.

[26] H.M. Semakula, G. Song, S.P. Achuu, S. Zhang, A Bayesian belief network modelling of household factors influencing the risk of malaria: A study of parasitaemia in children under five years of age in sub-Saharan Africa, Environ. Model. Softw. 75 (2016) 59–67.

[27] C.S. Smith, A.L. Howes, B. Price, C.A. McAlpine, Using a Bayesian belief network to predict suitable habitat of an endangered mammal – The Julia Creek dunnart (Sminthopsis douglasi), Biol. Cons. 139 (3) (2007) 333–347.

[28] I.M. Suffet, D. Khiari, A. Bruchet, The drinking water taste and odor wheel for the millennium: beyond geosmin and 2-methylisoborneol, Water Sci. Technol. 40 (6) (1999) 1–13.

[29] N. Sugiura, M. Utsumi, B. Wei, N. Iwami, K. Okano, Y. Kawauchi, T. Maekawa, Assessment for the complicated occurrence of nuisance odours from phytoplankton and environmental factors in a eutrophic lake, Lakes Reserv.: Res. Manage. 9 (3–4) (2004) 195–201.

[30] L. Uusitalo, Advantages and challenges of Bayesian networks in environmental modelling, Ecol. Model. 203 (3–4) (2007) 312–318.

[31] H. Uwins, P. Teasdale, H. Stratton, A case study investigating the occurrence of geosmin and 2-methylisoborneol (MIB) in the surface waters of the Hinze Dam, Gold Coast, Australia, Water Sci. Technol. 55 (5) (2007) 231–238.

Please cite this article as: H.J. Mayfield, E. Bertone, C. Smith et al., Use of a structure aware discretisation algorithm for Bayesian networks applied to water quality predictions, Mathematics and Computers in Simulation (2019), https://doi.org/10.1016/j.matcom.2019.07.005.