A Bayesian network-based data analytical approach to predict velocity distribution in small streams

Onur Genc and Ali Dag

ABSTRACT

Developing a reliable data analytical method for predicting the velocity profile in small streams is important in that it substantially decreases the amount of money and effort spent on measurement procedures. In recent studies it has been shown that machine learning models can be used to achieve such an important goal. In the proposed framework, a tree-augmented Naïve Bayes approach, a member of the Bayesian network family, is employed to address the aforementioned two issues. Therefore, the proposed study presents novelty in that it explores the relations among the predictor attributes and derives a probabilistic risk score associated with the predictions. The data set of four key stations, in two different basins, are employed and the eight observational variables and calculated non-dimensional parameters were utilized as inputs to the models for estimating the response values, u (point velocities in measured verticals). The results showed that the proposed data-analytical approach yields comparable results when compared to the widely used, powerful machine learning algorithms. More importantly, novel information is gained through exploring the interrelations among the predictors as well as deriving a case-specific probabilistic risk score for the prediction accuracy. These findings can be utilized to help field engineers to improve their decision-making mechanism in small streams.

Key words | Bayesian belief networks, ecological modeling, model validation, velocity distribution

Onur Genc (corresponding author)
Department of Civil Engineering,
Melikşah University,
Kayseri,
Turkey
E-mail: *ogenc@meliksah.edu.tr*

Ali Dag
Department of Industrial & Systems Engineering,
Auburn University,
Auburn,
Alabama,
USA

INTRODUCTION

Water resources and the factors that have an impact on them should be considered for water resource management and conservation. Therefore, active monitoring of flow properties is required for ecological balance and socio-economic reasons. Velocity profile, discharge, energy and momentum coefficients, and shear stress are the main characteristics of the most important flow properties that need to be identified carefully. Particularly, determination of the velocity profile is essential to estimate the other above-mentioned properties in a river cross section. Some difficulties arise during the collection of point velocities in river cross sections since the roughness characteristics and water depth tend to change with time from one section to another along the flow direction. The procedure of velocity measurements also requires time, money, and effort (Ardiclioglu *et al.* 2012).

From the past to the present, a number of traditional methods and data mining models have been proposed for dealing with difficulties in obtaining velocity distributions, in calculating the other flow properties mentioned above. For many years, the power and Prandtl–Von Karman universal velocity distribution laws were the most widely used to determine the velocity distribution. However, it is accepted that river hydrodynamic problems, which have complex structures, cannot be solved through these two equations (Ardiclioglu *et al.* 2012). Recently, an entropy concept was presented to solve the uncertainty problems in open channel flows by Chiu (1988, 1989, 1991). Then, many researchers followed up his method to identify the velocity distribution in open channels (Moramarco *et al.* 2004; Farina *et al.* 2014; Greco & Mirauda 2015). To apply the entropy method,

doi: 10.2166/hydro.2015.110

three parameters must be known M, h, and u_{max}. The entropy parameter M can be calculated using the relationship between the mean, U_m, and maximum velocities, u_{max} in Equation (1). h indicates water depth in the measured vertical of river cross section.

$$\Phi(M) = \frac{U_m}{u_{max}} = \frac{e^M}{e^M - 1} - \frac{1}{M} \quad (1)$$

Researchers have encountered many problems, such as lost time and effort in implementation of the classical methods as well as the current ones. The data mining methods have been satisfactorily utilized to solve the problems in hydraulics and water resources engineering (Kisi *et al.* 2012; Azamathulla & Jarrett 2013; Genc *et al.* 2014, 2015). Wu *et al.* (2008) have successfully applied prediction of the water level through various data-derived models including linear regression, the nearest-neighbor method, ANN (artificial neural network), and support vector regression. They report that the proposed distributed support vector regression model can perform river flow prediction better in comparison with other models. Taormina & Chau (2015) studied neural network river forecasting models for the prediction of future streamflow discharges in the Shenandoah River watershed, Virginia, USA. They introduced the concept that a multi-objective fully informed particle swarm optimization algorithm is found to provide better performing models for the prediction of discharge. Li *et al.* (2015) modeled the daily streamflow forecasting through stepwise-clustered hydrological inference (SCHI) model. Their study was a first attempt to estimate the daily flow using this method. They revealed that the SCHI model had a superior performance to predict it.

Bayesian networks (BNs) are one of the core components of modeling systems frequently used in ecological systems (Watson *et al.* 2004; Ames *et al.* 2005; Borsuk *et al.* 2006; Pollino *et al.* 2007; Hamilton *et al.* 2015; Lucena-Moya *et al.* 2015). These studies in the literature report that Bayesian belief networks (BBNs) provide the incorporation and representation of uncertain information by identifying causal dependencies between model inputs in terms of probabilistic relationships (Borsuk *et al.* 2004; Hamilton *et al.* 2015).

Specifically related with the applications of BN in water management related areas (i.e., river management, lake management, etc.), Adriaenssens *et al.* (2004) performed a study to assess the success of prediction of macroinvertebrate taxa in rivers. Correctly classified instances and Cohen's Kappa (K) were employed to assess the predictive capacity of the classification models. Their results showed that careful feature selection as well as sensitivity analysis (SA) are the main mechanics that can potentially improve models' performances and related outcomes, for practical use in river restoration management. In a similar study, Reckhow (1999) employed BN to predict the surface water quality and assessment. The rationale behind employing BN is their reasonable structure that aims to handle the complex probabilistic behavior of the variables in the ecosystem.

However, the existing machine learning algorithms have not taken into account the relations among the predictor variables, which might be crucially important in understanding the mechanics of the influential factors of velocity distributions. Furthermore, the existing researches have not considered the probability of accuracy for each prediction. In other words, they have not paid enough attention to the risk of not being able to classify the point velocities correctly, although their major concern was the correct classification.

In the proposed study, a tree-augmented network (TAN)-based BBN model has been utilized to not only predict the velocity distributions but also uncover the hidden relations among the attributes as well as model the probability of prediction risk. In this sense, the aim of developing the velocity predictor is to visualize the conditional interdependency among the explanatory variables and to derive case-specific risk measures that are associated with the predictions. Therefore, in this study, a novel approach, which has never been employed in the associated field, has been utilized to reach the aforementioned goals. To the best of our knowledge, the targeted issues have not been questioned and handled in the existing literature.

METHODOLOGY

In the proposed methodology, a probabilistic acyclic graph-based data analytic approach is proposed to predict the

velocity distributions in small streams. The procedure starts with a data collection and preparation phase, and the details of this phase are discussed in the subsequent paragraphs. The second step employs a five-fold cross validation approach to prepare five training samples and five mutually exclusive test samples to provide robustness and to prevent the potential bias that can be caused by data splitting issues. In the third step, tree-augmented BN is employed (1) to predict the velocity distribution and (2) to identify the conditional relations between the independent predictors. The fourth step involves a SA, to observe the relative contribution of each variable to the prediction. In the fifth and final step, the prediction results that were obtained through applying BN are compared with the other data mining models that were widely used in similar research problems.

The advantage of the TAN model utilized in the proposed study over the other machine learning models is that it provides a case-specific probability of outcome and the detailed understanding of conditional relations among the independent variables. These two outcomes are two novel contributions of the proposed methodology to the existing literature, as discussed in the 'Introduction'. These main steps of the proposed study are discussed in detail in the subsequent sections.

Data sources and study area

In this study, a historical instantaneous point velocity data set, collected in two different basins in Turkey, was employed to determine the velocity profile in small streams. The records of flow measurements in four different stations during the period 2005–2010 were utilized for obtaining the velocity distribution conducted in this study (Table 1). The study area is characterized by a semi-arid climate. Barsama Şahsenem and Bünyan stations are on a tributary (called Sarımsaklı stream) of the Kızılırmak River that flows through a 1,355 km waterway from Sivas in the center of Turkey to the Black Sea at the northern tip of Turkey. Sosun station is on the Zamanti River, situated in the southeast of Turkey, that drains to the Mediterranean Sea (Figure 1).

This study focuses on relatively small and shallow streams, where the discharges, Q, vary between 0.29 and 3.93 m^3/s. The maximum water depths, H_{max}, change between 0.26 and 0.86 m in measured cross sections and the water surface width, T, values vary between 2.3 and 9.0 m. The value of the mean velocity, U_m, cross-section area, A, aspect ratio, T/R, and slope of the water surface, S_{ws}, in the measured cross section are summarized in Table 1. Froude, Fr, and Reynolds, Re, numbers which are presented in Table 1 demonstrate that all the flow measurements were performed under subcritical and turbulent flow conditions. All maximum and minimum values of these mentioned properties are in bold font in Table 1.

An acoustic Doppler velocimeter (ADV), which can record instantaneous three-dimensional (x, y, and z) flow velocities at a single point with a relatively high frequency, is used to measure point velocities during flow measurements. These flow measurements are performed by measuring the velocity of particles in a remote sampling volume based upon the Doppler shift effect (Voulgaris & Trowbridge 1998; Mclelland & Nicholas 2000). Based on the water surface width, measured cross sections were split into n number of slices, which are given in Table 1, column 12. The point velocities were determined in the vertical direction starting about 4 cm from the streambed for each vertical. These transactions were repeated every 2 cm from this point to free water surface for every vertical slice. The ADV was kept in a fixed position while field studies were carried out.

The frequency histogram of 2,184 point velocity values, u, is plotted in Figure 2. Most of the point velocity values (1,794) are located between the values of 0.2 and 0.9 m/s. The minimum point velocity is about 0.0004 m/s and the maximum velocity is about 1.90 m/s in the data set.

As was mentioned in earlier sections, the variable of interest in the proposed method is *point velocities*. The existing studies in the literature have forecast *point velocity* by handling it as a continuous variable (Genc & Dag 2016). BN models can be applied as a prediction method only if the outcome is a discrete variable. Therefore, in the proposed method, *point velocity* values are converted into discrete classes. 'Discretization' is the term that is used for the process of converting a continuous variable into a limited number of classes. Discrete values, as opposed to continuous one, are preferred by many researchers in

Table 1 | Main flow properties for all measurements

Stations (1)	Dates d/m/y (2)	Q m^3/s (3)	U_m m/s (4)	A m^2 (5)	H_{max} M (6)	T m (7)	T/R (8)	S_{ws} (9)	Re ($\times 10^6$) (10)	Fr (11)	n (12)
Barsama_1	28/05/2005	1.810	0.890	2.23	0.39	8.3	34.00	0.0091	0.76	0.481	7
Barsama_2	19/05/2006	2.440	1.051	2.03	0.40	**9.0**	35.20	0.0036	0.94	0.531	7
Barsama_3	19/05/2009	**3.930**	**1.214**	2.11	0.45	**9.0**	29.70	0.0094	**1.47**	**0.578**	9
Barsama_4	31/05/2009	0.970	0.590	2.67	**0.26**	8.4	**45.40**	0.0092	0.40	0.333	8
Barsama_5	24/03/2010	1.510	0.806	2.79	0.38	8.6	34.40	0.0097	0.61	0.417	4
Barsama_6	18/04/2010	2.150	0.865	2.48	0.38	8.8	22.10	**0.0120**	0.85	0.421	5
Şahsenem_1	29/03/2006	0.816	0.354	2.04	0.72	6.0	26.80	0.0059	0.47	0.350	5
Şahsenem_2	20/10/2007	0.718	**0.214**	2.32	0.66	5.4	21.90	0.0061	0.46	0.298	9
Şahsenem_3	22/03/2008	0.792	0.301	**3.24**	0.72	6.0	22.10	0.0037	0.49	0.314	9
Şahsenem_4	03/05/2008	0.613	0.405	1.64	0.85	5.4	25.10	0.0045	0.39	0.307	9
Şahsenem_5	11/10/2008	0.667	0.426	1.87	**0.86**	5.5	22.00	0.0046	0.44	0.303	9
Şahsenem_6	08/11/2008	0.732	0.286	2.48	0.79	5.6	19.60	0.0064	0.51	0.282	10
Bünyan_1	24/06/2009	0.788	0.600	1.40	0.28	4.0	7.00	0.0020	0.71	0.133	7
Bünyan_2	08/02/2010	0.434	0.529	1.36	0.32	4.0	7.50	0.0030	0.40	**0.084**	7
Bünyan_3	27/09/2009	0.636	0.565	1.40	0.33	3.9	8.20	0.0022	0.50	0.113	6
Bünyan_4	04/04/2010	1.082	0.518	1.18	0.32	4.0	7.30	0.0018	0.78	0.140	4
Bünyan_5	16/05/2010	1.188	0.536	1.24	0.32	4.0	7.00	0.0024	0.85	0.147	4
Bünyan_6	20/06/2010	0.708	0.516	1.40	0.34	3.9	7.30	0.0010	0.53	0.103	4
Sosun_1	19/05/2009	0.886	0.561	1.58	0.62	3.2	7.49	0.0032	0.84	0.227	6
Sosun_2	31/05/2009	**0.294**	0.285	1.03	0.43	3.0	9.49	**0.0016**	**0.32**	0.144	5
Sosun_3	24/03/2010	0.338	0.327	1.03	0.45	2.9	8.85	0.0026	0.37	0.156	5
Sosun_4	18/04/2010	0.529	0.541	**0.98**	0.54	**2.3**	**6.53**	0.0034	0.67	0.235	5

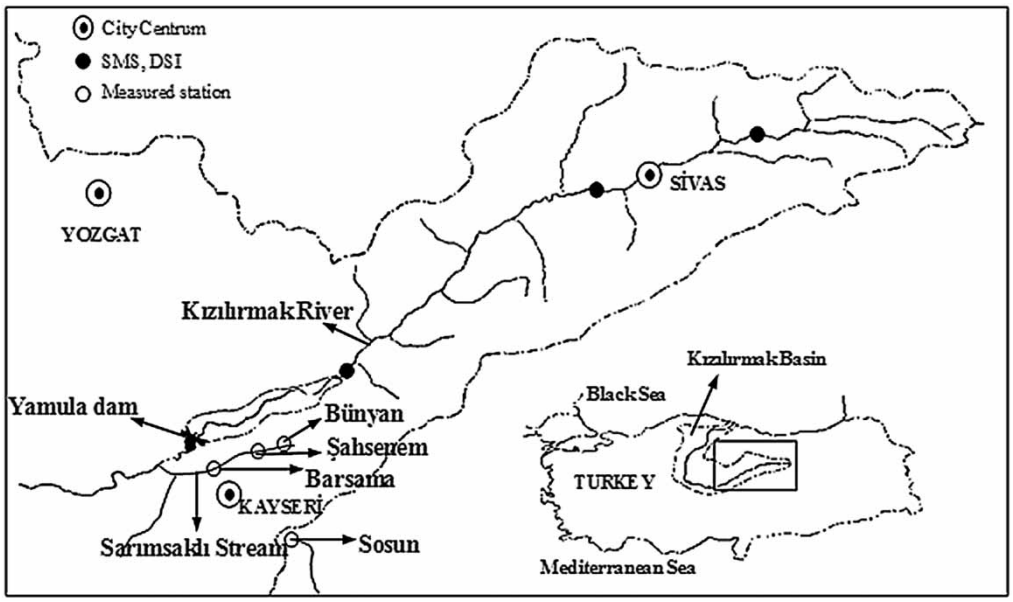

Figure 1 | Location of the study area and measurement stations (Ardiclioglu *et al.* 2012).

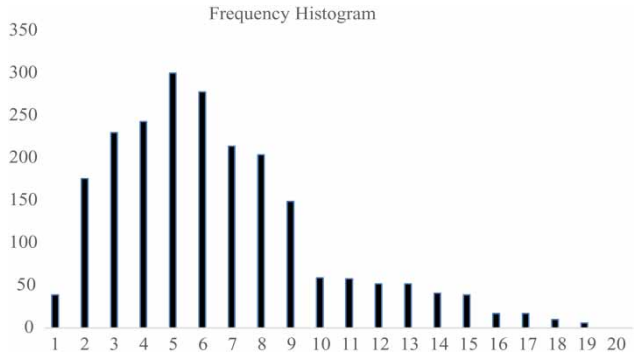

Figure 2 | The frequency histograms of the target variable, *u*.

prediction problems because of various advantages such as the following:

- Discrete values are easy to understand, use, and explain (Liu *et al.* 2002).
- Many algorithms spend less time on learning discrete values rather than continuous ones (Dougherty *et al.* 1995).
- The data set, itself, can be simplified and the data size can be reduced by conversion from continuous to discrete values.

Also, depending on the problem to be solved, the number of classes can be another important issue that needs to be determined. There are cases where the binary classification does not provide so much information (as in our case), since the outcome should be predicted as close as possible to its actual value. Therefore, in our study, the outcome variable is converted into ten classes using the following breakpoints (Table 2).

By applying the above-mentioned procedure, the problem has now become a multinomial classification problem, where the outcome variable has ten possible ordinal values, ranging from 1 to 10. Therefore, the learning algorithm will predict which class the *point velocity* belongs to by taking the eight predictor variables' values into account.

K-fold cross validation

K-fold cross validation is a method that was developed to provide objective performance measures by enhancing the model robustness and minimizing the potential bias that can be caused by random sampling (Kohavi 1995). In such a setting, the entire data set is randomly split into k mutually exclusive subsets of approximately equal size. Each time, the model is tested on one data split by using the remaining *k-1* as training splits. This procedure repeats until every k parts are used as a test set. The estimation of the overall performance of the model, which includes ten parts, is calculated by taking the average performance of k individual subsets as follows (Olson & Delen 2008):

$$\mathrm{CV} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{PM}_i \tag{2}$$

where CV stands for cross validation, k is the number of splits that is used, and PM is the performance metric used for evaluating the model performance. In this proposed study, to estimate the performance of a BN model, a stratified five-fold cross validation is employed. In stratified cross validation, the data set is randomly split into parts such that every split contains approximately an equal number of outcome labels. Existing research has shown that stratified cross validation tends to provide lower bias and lower variance when compared to regular k-fold cross validation (Kohavi 1995).

The feasible alternatives to the 'five-fold cross validation' could have been obtained by changing the number of folds. There is a trade-off between the number of samples (number of cases) and the number of folds employed. For example, if 'three-fold cross validation' had been employed in such a study, the learning algorithms would have learned the existing classes in a much deeper manner since each class would be represented with more learning samples. However, the risk of having bias among the folds would be higher when compared to the 'five-fold cross validation' case. Similarly, if 'ten-fold cross validation' had been employed in such a study, the learning algorithms would have learned the existing classes in a more superficial manner since each class would be represented with less learning samples. However,

Table 2 | Multinomial class ranges

Class no.	1	2	3	4	5	6	7	8	9	10
Range	<0.2	>0.2 < 0.3	>0.3 < 0.4	>0.4 < 0.5	>0.5 < 0.6	>0.6 < 0.7	>0.7 < 0.8	>0.8 < 1.0	>1.0 < 1.4	>1.4

the risk of having bias among the folds would be less when compared to the 'five-fold cross validation' case. Therefore, based on our past experience (depending on the sample size and number of classes used in the study) and the domain expert knowledge in the related field, it has been decided that employing 'five-fold cross validation' would have been the most reasonable option in such a case.

Deployment of BBN

A BBN is a directed acyclic graph that consists of interconnected variables, which present the probabilistic relations/dependencies among the predictor variables as well as the outcome variable. The nodes in the network represent the variables and the arcs represent the causal relations between these variables (Pearl 1985). It is a widely used data mining technique to model the complex relations among the attributes as well as to help reasoning under uncertainty (Anderson 1986). BBNs have been extensively used in prediction problems such as medicine, finance, civil engineering, etc. (Hearty *et al.* 2009; Landuyt *et al.* 2013; Sanford & Moosa 2013; Dag *et al.* 2015).

In the BBN model, Pa_{x_i} is the set of parents for each x_i, the BN chain rule can be expressed as (Koller & Friedman 2009):

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|Pa_{x_i}). \quad (3)$$

The Naïve Bayes network model is a simple model that can be employed to learn the structure by assuming a conditional independence between each predictor variables with the given target outcome. Given the attributes used in the model, Bayes rule is employed to compute the probability of target/outcome value, thus the target value that has the highest probability is chosen for the structure (Friedman *et al.* 1997). Tree-augmented Naïve Bayes (TAN) method is a relaxation of the Naïve Bayes classifiers, in which the outcome variable is a parent for each predictor attribute. However, in the TAN model, each predictor may have at most one other variable, with which it has some conditional relations. In other words, there might be one more arc that comes from another variable (parent node) in the network. That is, the class variable has no parents, however it is one of the parents of each variable along with at most one other variable. Inferences can be made on TAN by looking at arcs and nodes in the presented figure. An arc between two nodes represents the conditional interrelation between these two variables. The node where the arc exits is considered a parent variable to the other one, which the arc enters. In such a network, the contribution of the child node in predicting the class variable C is dependent on the value of the parent node. To exemplify, in Figure 3, X_3 and the class/target variable C are the parents of X_1. That is, in predicting the class variables, the contribution of X_1 is dependent on the value of X_3. In a classical TAN structure:

$$Pa_{x_i} = \{C, x_{\xi(i)}\}, \quad (4)$$

where $\xi(i)$ is the tree function over $x_1, \ldots, x_n$, and Pa_{x_i} is the set of parents for each x_i. Class variable (C) has no parents:

$$Pa_C = \varnothing. \quad (5)$$

An optimization problem arises to find the best tree where the objectives are to maximize the logarithmic likelihood of $\xi(i)$, and to build a maximum likelihood tree to find a maximal weighted spanning tree in a graph (Chow & Liu 1968). The TAN construction algorithm can be presented below as shown in Drakos *et al.* (2007).

I. Compute conditional mutual information function for each (i, j) pairs:

$$I_P(x_i{:}x_j|C) = \sum_{x_i, x_{j,C}} P(x_i, x_j, C) \log \frac{P(x_i, x_j|C)}{P(x_i|C)P(x_j|C)}, \; i \neq j \quad (6)$$

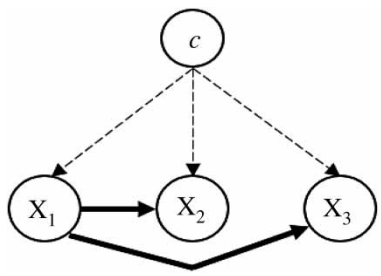

Figure 3 | TAN structure.

The information amount that x_j provides about x_i when the class variable is known is provided with this function:

II. Construct a complete undirected graph and use the conditional mutual information function to annotate the weight of an edge that connects x_i to x_j.
III. Build a maximum weighted spanning tree.
IV. Convert the undirected graph to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.
V. Construct a TAN model by adding a vertex labeled by C and adding an arc from C to each x_i.

Variable selection is a critical issue that needs to be handled carefully in employing data mining models for prediction problems. In building a BN, the number of the variables plays a crucial role in that BBN uses acyclic graph to visualize the relations among these attributes. Therefore, in the cases where the data set has a crowded set of potential predictors, the variables that should be incorporated in the model should be decided carefully to simplify the visual network. This selection can be made either by the domain expert or the use of variable selection algorithms. In our case, the eight predictor variables that need to be collected are carefully selected by the field experts who have many years of experience in the associated field. After having decided on the variables that need to be incorporated in the model, the network can be learnt and built without any explicit knowledge of domain experts by using a comprehensive data set (as in this current case). Therefore, researchers have paid a considerable amount of interest to learning from the data itself (Lucas *et al.* 2004). Our objective in the proposed study is to learn a TAN network, where predictor variables ($x_1, \ldots, x_n$) ($n = 8$) can be defined as a set of attributes/predictors. These predictors were obtained through a data collection procedure that was discussed in detail in the section 'Data sources and study area'.

Performance metrics

In binary classification problems, there are several performance criteria that can be measured to evaluate the performance of the models employed. These metrics are sensitivity, specificity, area under the receiver operating characteristic curve, F measure, G measure, etc., while for predicting the continuous target, R and R^2 measures are commonly employed for comparison. As for the multinomial classification problems, where the outcome has more than two classes to be predicted, there is no such standard evaluation criterion.

In our proposed study, *accuracy rate (AR)* was used as the prediction performance of the BBN model. It represents the ratio of total correct classifications to total number of samples in a model. Therefore, the bigger AR the better classification performance the model has. However, the performance of the model should be compared to the expected percentage of correct classification if the prediction was made in a random manner. For our case, if the assignment was made randomly per each case, the expected correct classification rate would be 10%, since the outcome has ten different classes. Therefore, one should judge the prediction performance with respect to the number of classes that the outcome has.

In our proposed study, two different rates were used to determine the classification performance of the model: the *exact* hit rate (only calculates the cases where the outcome is exactly correctly predicted) and *1-away* hit rate (it also counts the immediate neighbor predictions). The mathematical details of the hit rates can be seen as in Equations (7)–(9):

$$\mathrm{AR} = \frac{\text{Number of samples correctly classified}}{\text{Total number of samples}}. \tag{7}$$

$$\mathrm{AR}_{\mathrm{Exact}} = \frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{c}} \mathrm{w}_i. \tag{8}$$

$$\mathrm{AR}_{1-\mathrm{Away}} = \frac{1}{\mathrm{n}}\left((\mathrm{w}_1+\mathrm{w}_2)+\sum_{\mathrm{i}=2}^{\mathrm{c}-1}\mathrm{w}_{\mathrm{i}-1}+\mathrm{w}_{\mathrm{i}}+\mathrm{w}_{\mathrm{i}+1}+(\mathrm{w}_{\mathrm{c}-1}+\mathrm{w}_{\mathrm{c}})\right) \tag{9}$$

where the total number of classes is represented by c, N represents the total number of samples, and $\mathrm{w_i}$ represents the total number of samples classified as class i.

CASE STUDY AND DISCUSSION

Based on the discussion in the 'Methodology' section, a BBN (TAN) model has been employed to predict the

value of the point velocities on a multi-nominal scale. The results that were obtained through using BBN have also been compared with some of the popular prediction techniques that have been widely used in the literature. The remainder of this section is organized as follows. In the following section 'BN results', the multi-nominal classification results are presented by using a ten-by-ten matrix, since the dependent variable has ten categories. In this subsection, the network structure that visualizes the interactions among the independent factors is presented in addition to SA results. Finally, in the section 'Comparison to other methods', the results that were obtained through using other prediction models were compared with BN results.

The notations used in this study can be presented in the following forms. y/T is the ratio of lateral distance from channel wall (y) to the water surface width (T). z/H indicates the ratio of vertical distance from channel bed (z) to the water depth (H). T/H shows the ratio of water surface width (T) to the water depth (H). T/R is the aspect ratio, the ratio of surface width (T) to hydraulic radius (R). z/T denotes the ratio of vertical distance from the channel bed (z) to the water surface width (T). z/y represents the ratio of vertical distance from the channel bed (z) to the distance from channel wall (y). S_{ws} is the water surface slope. U_{sh} is the water surface velocity. u shows the measured point flow velocity (target variable).

BN results

Classification results

As has been discussed in previous sections, our BBN model aims to predict the point velocities in one of the ten multi-nominal categories. The performance of the model is measured through using two accuracy metrics. The first one is a conservative metric that represents the exact and correct classified cases. For our situation, where the dependent variables are categorized into ten classes, the field engineers would be satisfied if they could predict within one on either side. Therefore, the 1-away correct classification rates are also reported as the second measure, which is not as conservative as the first one. Our aggregated five-fold cross validation BBN results in a confusion matrix which is presented in Table 3. A confusion matrix is widely used in presenting the classification results. Actual and predicted classes are represented in columns and rows, respectively. The samples that are exact-correctly classified are represented in the diagonal cells from the upper left corner to lower right, as they are highlighted in Table 3. For example, the number of cases that are classified as class-1 while in fact they belong to class-1 are denoted in the upper left corner. Therefore, the numbers in diagonal cells represent the correct classification while others represent the misclassifications. In Table 3, the cells that are immediate neighbors to the diagonal cells represent the

Table 3 | Confusion matrix for the aggregated five-fold BBN classification results

		Actual categories										**Avg.**
Predicted categories		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
	1	93	63	21	22	7	3	4	1	0	0	
	2	75	82	52	28	18	4	2	3	0	0	
	3	15	46	58	37	18	5	3	1	0	0	
	4	17	26	68	108	72	27	7	11	0	0	
	5	6	5	18	71	91	44	10	1	0	0	
	6	4	0	9	14	39	42	25	6	0	0	
	7	0	0	3	4	17	58	90	58	5	0	
	8	3	2	8	5	3	8	35	60	19	0	
	9	2	7	7	10	12	24	28	66	159	34	
	10	0	0	0	0	0	0	0	0	21	54	
Number of samples		**215**	**231**	**244**	**299**	**277**	**215**	**204**	**207**	**204**	**88**	
Exact		**0.43**	**0.35**	**0.24**	**0.36**	**0.33**	**0.20**	**0.44**	**0.29**	**0.78**	**0.61**	**0.38**
1-away		**0.78**	**0.83**	**0.73**	**0.72**	**0.73**	**0.67**	**0.74**	**0.89**	**0.98**	**1.00**	**0.79**

misclassified samples by only a single class, which were used in calculating 1-away classification accuracies. In Table 3, the summary statistics on the accuracy measurements (both for exact and 1-away) of both for each individual category and the overall prediction accuracy are presented separately.

Since the dependent variable has been categorized into ten classes, it is expected that a random guess (prediction) about a sample has a 10% chance to correctly classify that sample. In other words, if a modeler randomly predicts the outcomes without using any mathematical model, the expected average for exact-correctly classified cases is 0.10, while in our model the average is 0.38, as can be seen from Table 3. These results indicate that the performance that the BBN model achieved has improved about 300% over a random guess. Not only the exact-correctly classified cases but also 1-away prediction results have shown a great performance in classifying the point velocities into ten multinomial categories by an average of 0.79. This basically means, almost eight out of ten predictions are either exactly-correct or only one class away from the target class.

As is emphasized in various sections of the proposed study, the main goal of this study is not only to predict the velocity distributions but also to investigate: (1) which of these predictors have conditional relations with each other, and if we can visualize these relations and (2) if we can derive a score for each predicted sample that gives the probability of accurate prediction?

To achieve the aforementioned goals, TAN has been employed since its structure perfectly fits the projected goals. The only way to apply such method is that the target outcome should be a discrete outcome, which was not the case in our original data set. Therefore, there were two options from which we could choose: (1) converting the continuous outcome to a binary variable; however, for those cases that need to be predicted as closely as possible, this option was not reasonable; and (2) converting the continuous outcome to a number of intervals such that these intervals should be satisfying the field engineers. Therefore, dividing the target outcome into ten intervals has created ten classes. Thus, two performance metric criteria that can handle such a situation have been employed accordingly, and are explained in great detail in 'Performance metrics'.

Having said that, to our knowledge, there is not any single study that has applied TAN to the related field. The studies in the existing literature (Adriaenssens *et al.* 2004; Ames *et al.* 2005; Kisi *et al.* 2012; Azamathulla *et al.* 2013; Genc *et al.* 2014) have rather handled the related problems by considering the target as a continuous outcome. Therefore, the performance criteria that were employed by the existing literature are measures such as R, MAE (mean absolute error), R^2, etc., which are widely used for continuous predictions. Therefore, the performance criteria that have been employed in the proposed method are not comparable with the existing studies on velocity predictions in streams.

SA

SA is a method to observe the cause and effect relationship between the predictors and the outcome that have been widely applied in machine learning models. It is used to identify the relative contribution of each variable to the prediction performance of the model. The basic idea in performing SA is to compare the model's classification performance when it includes a specific variable with the case when it does not include that specific variable. Therefore, the corresponding change in the output provides the amount of change in the model performance (Principe *et al.* 2000). This process is repeated for each and every input variable. A more detailed explanation about SA can be found in Dag *et al.* (2015).

Therefore, in order to observe the relative contribution of each independent variable towards classifying the point velocities, SA is performed. The results of this phase are weighted between 0 and 1, which are summarized in Figure 4. As can be seen from Figure 4, the variables *T/H*,

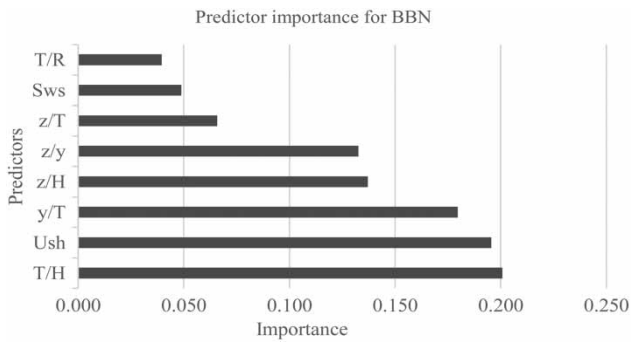

Figure 4 | Sensitivity analysis.

U_{sh}, and y/T have a much higher contribution in predicting the point velocities in small streams, especially when compared to the variables such as T/R, S_{ws}, and z/T.

Visual network structure of the model

One of the main advantages of employing TAN is that the interrelations between the predictors and the outcome (point velocity), and the impact of these relations on prediction can be investigated. In other words, the TAN structure not only predicts the outcome, but it also visualizes the inter-variable relations and measures how these relations affect the prediction probability. In our proposed study, since five-fold cross validation has been employed to model BBN structure, there are five mutually exclusive test sets used to visualize the network structure. Therefore, it would not be practical to present the network structures of all these five individual model networks in this study. For this reason, the fifth fold is selected to represent the network structure of the TAN model in that its classification performance is the closest one to the average of the entire set of five folds. It should be noted that the difference between the folds are minor in terms of prediction results and network structure. The probabilistic acyclic graph obtained by using the fifth fold is provided in Figure 5.

The variables and parameters, which are presented in this figure, can be identified as y/T, where y is the lateral distance from channel wall to measured point and T is the water surface width. z/h: z shows the vertical distance from channel bed and h indicates the water depth in measured vertical. T/H represents the ratio of water surface width (T) to the water depth (H). z/T and z/y are the non-dimensional parameters which were derived using observational values. T/R, S_{ws}, and U_{sh} are the variables that were presented in Table 1. As was explained in the related section, the presence of an arrow from an independent variable (parent) to another (child) implies that the impact of the child predictor on the *point velocity* is dependent on the value of the parent. Also, the dependent variable has many children while it has no parents in the network. There are several relationships that can be discovered that intuitively make a great deal of sense. There are several interesting patterns that could be inferred from the figure by closely examining the parent–child relationships in the given figure.

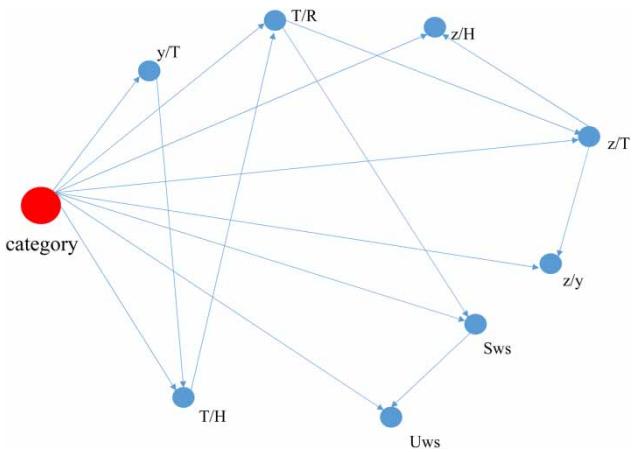

Figure 5 | The visual structure of TAN model.

For example, the impact of every predictor on the outcome is somehow affected by either one or more predictor(s). In other words, there is no predictor in the system such that its impact on the outcome is independent. Another expected inference could be that the contribution of predictor T/H in predicting the outcome variable depends on predictor y/T. This means that the interrelation between these two predictors would have an effect on the point velocity. This inference makes absolute sense since the prediction becomes more challenging in the case of wide and shallow streams. Therefore, the relation between T and H is influential on the prediction capability of the model. Also, the number of measured verticals used in splitting the cross sections is in direct proportion to the prediction. The arrow from z/T to z/H implies that the value of z/H has a particular impact on the point velocity. However, this impact is also dependent on the value of z/T. This inference makes absolute sense for the same reason. Similarly, the contribution of predictor U_{sh} has conditional interdependency with the predictor S_{ws}. The reason for that is, as the S_{ws} (slope of free water surface) increases, predicting the point velocity value would be more difficult since high S_{ws} causes high uncertainty in predicting the outcome.

There are several other inferences that can be made from the given figure, which makes absolute sense to field engineers. Also, we should note that the machine learning algorithms cover the hidden, unexpected, and uncovered patterns, as well. Therefore, there might be some unexpected

findings that should be investigated further in a prospective study.

The information that was extracted from the network can be further analyzed in a more detailed manner by a future study. More specifically, the findings can provide field engineers with significant insights on the variables that have impact on point velocity.

Comparison to other methods

As was emphasized in the earlier sections, the main goal of the current study is not only to predict the point velocities in small streams but rather to investigate the relations among the attributes and derive a probabilistic score for each prediction. Having said that, we have compared the BBN model's prediction results with other widely used data mining models that could be utilized for similar problem scenarios. Specifically, we have employed ANN, multinomial logistic regression (MLR) and decision trees due to the high performance that they have shown in preliminary analysis. What follows is a brief description for these methods.

ANNs are based on a computational model that is inspired by the human nervous system. The basic element of an ANN is a neuron. A neural network is a structure of many such neurons connected in a systematic way (Modeler 2012). Neural networks are typically organized in layers. Input variables are loaded into the network via the input layer. ANNs combine input variables in order to process information. An ANN accepts many inputs, learns them, applies a (usually nonlinear) transfer function in hidden layers, and presents the result in the output layer. Each layer is connected to the previous and the following layers. The weights of connections between the neurons are associated with layers, which obtain the strength of influence one neuron has on another (Modeler 2012). In recent studies, many researchers used ANNs in water resources, hydraulics, and hydrology (Yang & Chang 2005; Kisi *et al.* 2013; Genc *et al.* 2014). In our ANN model, the multilayer perceptron algorithm has been utilized using one hidden layer due to high accuracy achieved in the preliminary analysis.

MLR is a statistical technique that generalizes logistic regression to multiclass problems based on values of input fields (Greene 1993). Models can be produced by MLR when the target field is a set field with more than two possible values. In logistic regression, each data set is recoded as a group of numeric fields. MLR was utilized to estimate the probabilities of the different possible outputs of a categorically distributed dependent variable, given a data set of independent variables (Modeler 2012). MLR has been commonly employed in hydraulics and water resources in recent years by researchers (Mamitimin *et al.* 2015; Phillips *et al.* 2015). A stepwise MLR has been employed in our model via making trial and error based experiments.

Classification and regression tree (C&R-T) model is one of the tree-based characterization and estimation methods (Genc *et al.* 2015). The training set is divided into similar parts in the application of this model. The main aim of utilizing the C&R-T is to have subgroups with similar output variables. Some type of the node impurity measure is employed to measure the similarity. The data are partitioned into two subsets in C&R-T. Each subset should be more homogeneous than in the previous subset. It also allows one to specify the prior probability distribution in a classification problem (Modeler 2012).

In the subsequent section, the results obtained through using MLR, C&RT, and ANNs are presented by incorporating five-fold cross validation into each model by also using the same performance criteria, which were discussed in 'Performance metrics'. The aggregated five-fold cross validated results obtained are presented in Table 4.

As can be seen from both Tables 3 and 4, tree-augmented Bayes network outperformed the other data mining models in both exact and 1-away performance metrics, except for ANNs. The ranking (from high to low) of the models in terms of both exact and 1-way correct classification results are ANN, BBN (TAN), C&RT, and MLR. Recall that the main goal of this study is not to compare the data mining models in terms of prediction performances, but rather to investigate the conditional relations among the attributes and to derive probabilistic risk scores for each prediction for each test sample. The exact and 1-away classification results' means are illustrated in Figure 6. Comparison of the advantages and disadvantages of C&RT, ANN, and MLR models are presented in Table 5.

Table 4 | Confusion matrixes for data mining models representing the aggregated five-fold cross validated results

			Actual Categories										Avg.
MLR	Predicted categories		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
		1	**101**	57	28	23	12	10	4	0	0	0	
		2	56	**76**	76	50	28	16	4	2	0	0	
		3	6	28	**34**	34	24	10	5	2	0	0	
		4	17	27	58	**101**	79	39	21	14	1	0	
		5	17	21	32	54	**61**	41	20	9	1	0	
		6	6	10	5	9	30	**24**	15	5	2	0	
		7	5	2	1	10	19	47	**67**	43	3	0	
		8	7	10	10	18	23	25	62	**96**	21	0	
		9	0	0	0	0	1	3	6	36	**154**	30	
		10	0	0	0	0	0	0	0	0	22	**58**	
	Exact		**0.47**	**0.33**	**0.14**	**0.34**	**0.22**	**0.11**	**0.33**	**0.46**	**0.75**	**0.66**	**0.35**
	1-away		**0.73**	**0.70**	**0.69**	**0.63**	**0.61**	**0.52**	**0.71**	**0.85**	**0.97**	**1.00**	**0.71**
C&RT	Predicted categories		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
		1	**64**	39	11	10	0	0	0	0	0	0	
		2	77	**103**	80	45	19	10	0	7	0	0	
		3	19	9	**11**	5	5	0	2	0	0	0	
		4	34	30	69	**119**	62	11	3	1	0	0	
		5	16	40	62	102	**157**	99	44	20	0	0	
		6	2	0	0	5	11	**10**	6	7	4	0	
		7	2	3	5	5	16	73	**132**	124	27	0	
		8	0	2	3	2	3	3	5	**8**	10	0	
		9	1	5	3	6	4	9	12	40	**140**	50	
		10	0	0	0	0	0	0	0	0	23	**38**	
	Exact		**0.30**	**0.45**	**0.05**	**0.40**	**0.57**	**0.05**	**0.65**	**0.04**	**0.69**	**0.43**	**0.36**
	1-away		**0.66**	**0.65**	**0.66**	**0.76**	**0.83**	**0.85**	**0.70**	**0.83**	**0.85**	**1.00**	**0.76**
ANN	Predicted categories		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
		1	**120**	57	22	12	7	5	2	2	0	0	
		2	50	**98**	58	9	3	2	0	0	0	0	
		3	34	52	**80**	58	9	1	2	0	0	0	
		4	5	12	52	**129**	77	25	10	10	1	0	
		5	5	9	27	67	**114**	61	16	9	0	0	
		6	1	2	2	11	42	**66**	34	9	3	0	
		7	0	0	1	7	11	34	**86**	40	12	0	
		8	0	0	1	5	10	18	43	**99**	15	0	
		9	0	1	1	1	4	3	11	38	**155**	34	
		10	0	0	0	0	0	0	0	0	18	**54**	
	Exact		**0.56**	**0.42**	**0.33**	**0.43**	**0.41**	**0.31**	**0.42**	**0.48**	**0.76**	**0.61**	**0.46**
	1-away		**0.79**	**0.90**	**0.78**	**0.85**	**0.84**	**0.75**	**0.80**	**0.86**	**0.92**	**1.00**	**0.84**

FUTURE RESEARCH AND CONCLUSION

The main goal of the proposed research methodology is to develop a machine learning based mathematical model to observe the conditional relations among explanatory variables as well as to obtain a probabilistic score for prediction accuracy. To achieve these goals, a TAN network based machine learning method was developed by incorporating the five-fold cross validation concept, to classify the point velocities on a multinomial scale. A large data set that was collected between 2005 and 2010 from four different cross sections in Turkey, which includes eight potential predictors, was utilized in our study. By analyzing the data through the methodology used in this study, the following research questions have been addressed:

(a) What are the important contributors in predicting the point velocities in small streams?

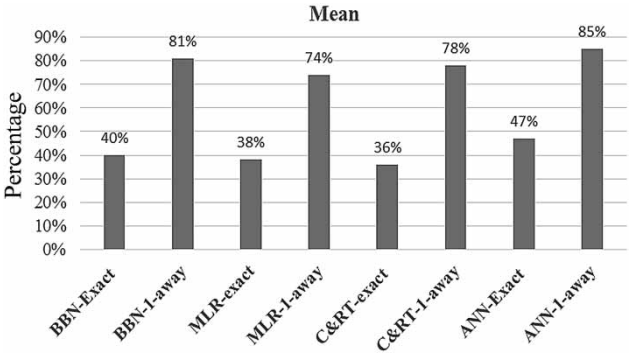

Figure 6 | Comparison of the classification models based on accuracy level.

(b) Which of these predictors have conditional relations with each other, and can we visualize these relations?
(c) Can we derive a score for each predicted sample that gives the probability of accurate prediction?

Therefore, the proposed method considers a probabilistic graphical network model, which has not been investigated in the relevant literature. This graphical network not only predicts the velocity but also observes the relations among the variables. Although the main goal is not only to predict but rather to observe the relations, TAN network has outperformed widely used popular prediction algorithms except for ANN. The findings which have been obtained through probabilistic graphical approach can provide valuable insights to field engineers for them to utilize in decision-making processes.

The methodology offered by the current research can be applied to any related field. The reason for that is that a very similar procedure can be applied to achieve such similar goals. The sequential steps, from data acquisition to classification of the samples, can be employed for a related problem where the outcome needs to be classified and the conditional relations need to be visualized.

It should be noted that using different training and test samples, the results obtained could have been different. This causes uncertainty and, in turn, becomes a concern that needs to be handled. However employing k-fold cross validation concept would decrease the possible differences between the performance levels. Recall, there are k different results, which have been obtained from k different test sets, which have been averaged while computing the final classification measure. Since, there are five different sets utilized in the proposed study, the potential risk of having different scores has been decreased satisfactorily.

Table 5 | Comparison of the advantages and disadvantages of C&RT, ANN and MLR models

C&RT		ANN		MLR	
Advantages	**Disadvantages**	**Advantages**	**Disadvantages**	**Advantages**	**Disadvantages**
Easy and intuitive interpretation	Error occurred in higher splits are propagated down because of the dependency among the subsequent splits	Can handle complex data sets, where there are non-linear relations among the attributes	High computational effort and time	In the case of high dimension, it may achieve more reliable results	Needs relatively high sample size in order to achieve reliable results
Robust to outliers, less effort for data preparation and usually faster at predicting the unknown records	Tree construction may badly be affected by non-predictor variables (such as measurement ID, etc.)	The relations and interactions among the predictor variables can be detected and handled in predicting the outcome	Risk of being stuck in local minima and black-box models (not explainable)	Not black box. Results are interpretable	Continuous outcomes cannot be handled
Tree performance is not affected from the complex relations among the predictors, also can handle missing data	Potential risk of sub-tree overproduction and higher risk of being error-prone in the existence of too many variables	Efficient on imprecise data sets	Prone to over-fitting	The interactions can be embedded into the model	Has relatively high risk when modeling non-linearity when compared to ANN models

A number of limitations in our data analytical study need to be mentioned. We could have incorporated a ten-fold cross validation concept instead of five-fold, which could increase the robustness and unbiasedness across the models. The main reason for that is each fold should have a sufficient number of classes for each of them to be trained well enough by the prediction algorithms. (Recall that, we have ten classes (from 1 to 10) to classify.) Owing to the limited number of samples collected (2,184 samples), employing five-fold cross validation would provide more samples for each fold, which in turn would prevent the over-fitting issues in the training phases. Second, in the comparison section, we could have presented the results from other machine learning algorithms that we have employed in the preliminary data analysis phase. However, presenting them all would have been unnecessary for readers and would take up a great deal of space. Third, more variables could have been incorporated to improve the model accuracy. However, based on the field engineers' suggestions as well as the relevant literature published, the eight variables that were used in this study have been found to be the most relevant and common ones.

A novel approach to predict the point velocities in small streams on a multinomial scale has been developed. Our approach can be easily extended to other relevant problem scenarios such as forecasting *discharge and shear stress* in small streams.

Potential research that could be carried out in the near future is that a graphical user interface can be developed that automatically predicts the point velocities by running machine learning based classification models as the background. Another possible improvement that could be made is to develop *voting based ensembling techniques,* which classify the target outcome based on the votes obtained from various models, that can be employed to classify the point velocities.

ACKNOWLEDGEMENTS

The corresponding author would like to thank the Scientific and Technological Research Council of Turkey (TUBITAK), since this paper was written during the period of time in which he was supported by this council to pursue a one-year visiting scholarship program at Auburn University, Auburn, Alabama/USA.

REFERENCES

Adriaenssens, V., Goethals, P. L. M., Charles, J. & De Pauw, N. 2004 Application of Bayesian belief networks for the prediction of macroinvertebrate taxa in rivers. *Int. J. Limnol.* **40** (3), 181–191.

Ames, D., Neilson, B., Stevens, D. & Lall, U. 2005 Using Bayesian networks to model watershed management decisions: an East Canyon Creek case study. *J. Hydroinform.* **7**, 267–282.

Anderson, J. R. 1986 Knowledge compilation: the general learning mechanism. In: *Machine Learning: An Artificial Intelligence Approach*, Vol. 2 (R. S. Michalski, J. G. Carbonell & T. M. Mitchell, eds). Morgan Kaufmann, San Francisco, CA, USA, pp. 289–310.

Ardiclioglu, M., Genc, O., Kalin, L. & Agiralioglu, N. 2012 Investigation of flow properties in natural streams using the entropy concept. *Water Environ. J.* **26**, 147–154.

Azamathulla, H. Md. & Jarrett, R. D. 2013 Use of gene-expression programming to estimate Manning's roughness coefficient for high gradient streams. *Water Resour. Manage.* **27** (3), 715–729.

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

Borsuk, M. E., Reichert, P., Peter, A., Schager, E. & Burkhardt-Holm, P. 2006 Assessing the decline of brown trout (Salmo trutta) in Swiss rivers using a Bayesian probability network. *Ecol. Modell.* **192** (1), 224–244.

Chiu, C. L. 1988 Entropy and 2-D velocity distribution in open channel. *J. Hydraul. Eng.* **114** (7), 738–756.

Chiu, C. L. 1989 Velocity distribution in open channel flow. *J. Hydraul. Eng.* **115** (5), 576–594.

Chiu, C. L. 1991 Application of entropy concept in open channel flow study. *J. Hydraul. Eng.* **117** (5), 615–627.

Chow, C. & Liu, C. 1968 Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory* **14** (3), 462–467.

Dag, A., Topuz, M. K., Oztekin, A., Bulur, S. & Megahed, F. M. 2015 A probabilistic data-driven methodology to score heart transplant survival. Decision Support Systems (under review).

Dougherty, J., Kohavi, R. & Sahami, M. 1995 Supervised and unsupervised discretization of continuous features. In: *Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, Los Altos, CA. pp. 194–202.

Drakos, S. G., Kfoury, A. G., Gilbert, E. M., Long, J. W., Stringham, J. C., Hammond, E. H. & Renlund, D. G. 2007 Multivariate predictors of heart transplantation outcomes in the era of chronic mechanical circulatory support. *Ann. Thorac. Surg.* **83** (1), 62–67.

Farina, G., Alvisi, S., Franchini, M. & Moramarco, T. 2014 Three methods for estimating the entropy parameter m based on a decreasing number of velocity measurements in a river cross-section. *Entropy* **16** (5), 2512–2529.

Friedman, N., Geiger, D. & Goldszmidt, M. 1997 Bayesian network classifiers. *Mach. Learn.* **29** (2–3). 131–163.

Genc, O. & Dag, A. 2016 A machine learning-based approach to predict the velocity profiles in small streams. *Water Resour. Manage.* **30** (1), 43–61.

Genc, O., Kisi, O. & Ardiclioglu, M. 2014 Determination of mean velocity and discharge in natural streams using neuro-fuzzy and neural network approaches. *Water Resour. Manage.* **28**, 2387–2400.

Genc, O., Gonen, B. & Ardiclioglu, M. 2015 A comparative evaluation of shear stress modeling based on machine learning methods in small streams. *J. Hydroinform.* **17** (5), 805–816.

Greco, M. & Mirauda, D. 2015 An entropy based velocity profile for steady flows with large-scale roughness. In: *Engineering Geology for Society and Territory*, Vol. 3 (G. Lollino, M. Arattano, M. Rinaldi, O. Giustolisi, J.-C. Marechal & G. E. Grant, eds). Springer International Publishing, Heidelberg, pp. 641–645.

Greene, W. H. 1993 *Econometric Analysis*, 5th edn. Prentice Hall, New York, pp. 720–723.

Hamilton, S. H., Pollino, C. A. & Jakeman, A. J. 2015 Habitat suitability modelling of rare species using Bayesian networks: Model evaluation under limited data. *Ecol. Modell.* **299**, 64–78.

Hearty, P., Fenton, N., Marquez, D. & Neil, M. 2009 Predicting project velocity in XP using a learning dynamic Bayesian network model. *IEEE Trans. Softw. Eng.* **35** (1), 124–137.

Kisi, O., Bilhan, O. & Emiroglu, M. E. 2012 ANFIS To estimate discharge capacity of rectangular side weir. *Proc. ICE-Water Manage.* **166** (9), 479–487.

Kisi, O., Bilhan, O. & Emiroglu, M. E. 2013 Anfis to estimate discharge capacity of rectangular side weir. *P I Civil Eng-Wat M* **166**, 479–487.

Kohavi, R. 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI* **14** (2), 1137–1145.

Koller, D. & Friedman, N. 2009 *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, USA.

Landuyt, D., Broekx, S., D'hondt, R., Engelen, G., Aertsens, J. & Goethals, P. L. 2013 A review of Bayesian belief networks in ecosystem service modelling. *Environ. Modell. Softw.* **46**, 1–11.

Li, Z., Huang, G., Han, J., Wang, X., Fan, Y., Cheng, G. & Huang, W. 2015 Development of a stepwise-clustered hydrological inference model. *J. Hydrol. Eng.* **20** (10), 04015008.

Liu, H., Hussain, F., Chew, T. L. & Dash, M. 2002 Discretization an enabling technique. *Data Mining Knowl. Discov.* **6**, 393–423.

Lucas, P. J., van der Gaag, L. C. & Abu-Hanna, A. 2004 Bayesian networks in biomedicine and health-care. *Artif. Intell. Med.* **30** (3), 201–214.

Lucena-Moya, P., Brawata, R., Kath, J., Harrison, E., El Sawah, S. & Dyer, F. 2015 Discretization of continuous predictor variables in Bayesian networks: an ecological threshold approach. *Environ. Modell. Softw.* **66**, 36–45.

Mamitimin, Y., Feike, T., Seifert, I. & Doluschitz, R. 2015 Irrigation in the Tarim Basin, China: farmers' response to changes in water pricing practices. *Environ. Earth Sci.* **73** (2), 559–569.

Mclelland, S. J. & Nicholas, A. P. 2000 A new method for evaluating errors in high-frequency ADV measurements. *Hydrol. Process.* **14** (2), 351–366.

Modeler, I. S. 2012 *Algorithms Guide*. IBM Corporation, Chicago.

Moramarco, T., Saltalippi, C. & Singh, V. P. 2004 Estimation of mean velocity in natural channels based on Chiu's velocity distribution equation. *J. Hydrol. Eng.* **9** (1), 42–50.

Olson, D. L. & Delen, D. 2008 *Advanced Data Mining Techniques*. Springer Science & Business Media, Berlin, Heidelberg.

Pearl, J. 1985 Bayesian networks: a model of self-activated memory for evidential reasoning. In: *Proceedings of the 7th Conference of the Cognitive Science Society*, University of California, Irvine, CA, pp. 329–334.

Phillips, J., Cripps, E., Lau, J. W. & Hodkiewicz, M. R. 2015 Classifying machinery condition using oil samples and binary logistic regression. *Mech. Syst. Signal Process.* **60**, 316–325.

Pollino, C. A., White, A. K. & Hart, B. T. 2007 Examination of conflicts and improved strategies for the management of an endangered Eucalypt species using Bayesian networks. *Ecol. Modell.* **201** (1), 37–59.

Principe, J. C., Euliano, N. R. & Lefebre, W. C. 2000 *Neural and Adaptive Systems: Fundamentals Through Simulations*. John Wiley and Sons, New York.

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

Sanford, A. & Moosa, I. 2013 Operational risk modelling and organizational learning in structured finance operations: a Bayesian network approach. *J. Oper. Res. Soc.* **66** (1), 86–115.

Taormina, R. & Chau, K. W. 2015 Neural network river forecasting with multi-objective fully informed particle swarm optimization. *J. Hydroinform.* **17** (1), 99–113.

Voulgaris, G. & Trowbridge, J. H. 1998 Evaluation of the Acoustic Doppler Velocimeter (ADV) for turbulence measurements. *J. Atmos. Ocean. Technol.* **15** (1), 272–289.

Watson, T., Christian, C., Mason, A., Smith, M. & Meyer, R. 2004 Bayesian-based pipe failure model. *J. Hydroinform.* **6**, 259–264.

Wu, C. L., Chau, K. W. & Li, Y. S. 2008 River stage prediction based on a distributed support vector regression. *J. Hydrol.* **358** (1), 96–111.

Yang, H. C. & Chang, F. J. 2005 Modelling combined open channel flow by artificial neural networks. *Hydrol. Process.* **19**, 3747–3762.

First received 25 May 2015; accepted in revised form 7 September 2015. Available online 23 October 2015