ORIGINAL ARTICLE

Prediction of longitudinal dispersion coefficient in natural rivers using a cluster-based Bayesian network

Mohamad Javad Alizadeh[1] · **Hosein Shahheydari**[1] · **Mohammad Reza Kavianpour**[1] · **Hamid Shamloo**[1] · **Reza Barati**[2]

Received: 22 September 2016 / Accepted: 26 December 2016 / Published online: 17 January 2017

Abstract The longitudinal dispersion coefficient is a key element in determining the distribution and transmission of pollution, especially when cross-sectional mixing is completed. However, the existing predictive techniques for this purpose exhibit great amounts of uncertainty. The main objective of this study is to present a more accurate model for predicting longitudinal dispersion coefficient in natural rivers and streams. Bayesian network (BN) approach was considered in the modeling procedure. Two forms of input variables including dimensional and dimensionless parameters were examined to find the best model structure. In order to increase the performance of the model, the clustering method as a preprocessing data technique was applied to categorize the data in separate groups with similar characteristics. An expansive data set consisting of 149 field measurements was used for training and testing steps of the developed models. Three performance evaluation criteria were adopted for comparison of the results of the different models. Comparison of the present results with the artificial neural network (ANN) model and also well-known existing equations showed the efficiency of the present model. The performance of dimensionless BN model 30% is more than dimensional ones in terms of the root mean square error. The accuracy criterion was increased from 70 to 83% by performing clustering analysis on the BN model. The BN-cluster model 43% is more accurate than ANN model in terms of the accuracy criterion. The results indicate that the BN-cluster model give 16% better results than the best available considered model in terms of the accuracy criterion. The developed model provides a suitable approach for predicting pollutant transport in natural rivers.

Keywords Bayesian network · Clustering analysis · Longitudinal dispersion coefficient · Natural stream

Introduction

Although the dispersion and mixing of pollutants take place in all three dimensions (i.e., longitudinal, vertical and lateral dimensions) of natural rivers and streams, the longitudinal dispersion is the dominant process. Accurate estimation of the longitudinal dispersion coefficient (K) is required in several applied hydraulic problems such as environmental engineering, river engineering, estuaries problems, intake designs and risk assessment of the injection of hazardous contaminants into river flows (Fischer et al. 1979; Liu 1977; Deng et al. 2001; Azamathulla and Wu 2011; Azamathulla and Ghani 2011; Sahay 2011; Tutmez and Yuceer 2013; Altunkaynak 2016; Najafzadeh and Tafarojnoruz 2016). The estimation of this coefficient is complicated due to irregularities of natural channels in shape and bed configuration and therefore in their

✉ Hosein Shahheydari
ashahheydari@mail.kntu.ac.ir

Mohamad Javad Alizadeh
mjalizadeh@mail.kntu.ac.ir

Mohammad Reza Kavianpour
kavianpour@kntu.ac.ir

Hamid Shamloo
hshamloo@yahoo.com; hshamloo@kntu.ac.ir

Reza Barati
r88barati@gmail.com; reza.barati@modares.ac.ir

[1] Faculty of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran

[2] Faculty of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, Iran

hydraulic conditions. Quantifying these bathymetric parameters is a challenging matter, where the related data for many rivers are not available.

K can be estimated experimentally (e.g., Perucca et al. 2009; Wang and Huai 2016), theoretically (e.g., Deng et al. 2001, 2002; Seo and Baek 2004; Wang and Huai 2016) and empirically (e.g., Seo and Cheong 1998; Swamee et al. 2000; Zeng and Huai 2014; Disley et al. 2015). A direct estimation of the dispersion coefficient by experimental means requires expensive and time-consuming tracer studies and/or limits to rectangular flumes data (Etemad-Shahidi and Taghipour 2012). The theoretical determination of K is also difficult due to the lack of the knowledge of transverse profiles of both velocity and depth of the flow (Deng et al. 2001). Hence, a large number of studies have been focused on developing empirical models for the estimation of K. Most of the predictive equations and models are developed based on four easily measurable variables including channel width (W), shear velocity (U_*), cross-sectional average velocity (U) and flow depth (H) (e.g., Kashefipour and Falconer 2002; Seo and Cheong 1998). The results of previous predictive equations differ and contain some amount of uncertainty, as will be shown.

In past few years, artificial intelligence techniques, such as artificial neural network (ANN), adaptive neuro-fuzzy inference system and genetic algorithm (GA) have shown promising performance in predicting longitudinal dispersion (Tayfur 2009; Toprak and Cigizoglu 2008). However, in some cases, the results showed significant variation. Also, excluding great values of W/H and K [$W/H > 50$ and $K > 100$ (m^2/s)] introduced more disadvantage to the ANN and most of the existing predictive equations. For example, the root mean square error (RMSE) of ANN model of Tayfur and Singh (2005) is 193.0 (m^2/s) in test step, while RMSE of this model is 19.3 when the experimental data with $K > 100$ are not included. As the longitudinal dispersion coefficient is a key element to determine the fate of pollution in rivers, therefore, an appropriate estimation of the coefficient has many applications in practical and engineering problems. Also, the development of a more accurate predictive model to cover the extreme values of K is of great interest.

Recently, Bayesian networks (BNs) have been successfully applied for hydrological processes including uncertain nonlinear and complex relationship among variables. Farmani et al. (2009) applied an evolutionary Bayesian belief network methodology for optimum management of groundwater contamination. They showed that the BN approach can help when dealing with uncertainties in decision making pertaining to human behavior. Xu et al. (2006) conducted a research study which applied a Bayesian regularized back-propagation neural network model for trend analysis, acidity and chemical composition of precipitation in North Carolina using precipitation chemistry data. Spatiotemporal drought forecasting within Bayesian networks has been carried out by Madadgar and Moradkhani (2014). The study demonstrated that Bayesian networks are suitable for probabilistic drought forecasting and have potential to improve drought characterization in different applications. Matheussen and Granmo (2015) proposed a snow accumulation and melt model formulated as a dynamic Bayesian network (DBN). They encoded uncertainty explicitly and trained the DBN using Monte Carlo analysis, carried out with a deterministic hydrology model under a wide range of plausible parameter configurations.

On the other hand, it was found that applying some data preprocessing techniques (clustering, wavelet transformation, etc.) linked with the main models can improve the efficiency of the forecasting (Alizadeh and Kavianpour 2015; Nourani et al. 2009). As the data applied in this study vary in a wide range, therefore, applying the clustering approach to recognize and group the subsets of the river data with similar pattern can be helpful in the modeling procedure. Moreover, the wavelet transform is applied to investigate and de-noise the temporal variation of the data. The purpose of clustering is to introduce different series of data from a large data set and produce a brief representation of a system's behavior.

The main objective of this study is to employ a Bayesian network methodology to present a more accurate model for predicting longitudinal dispersion coefficient in natural rivers and streams. In this regard, a wide range of field data including channel width (W), shear velocity (U_*), cross-sectional average velocity (U) and flow depth (H) were applied in the BN models. In order to increase the accuracy, clustering analysis as a preprocessing data technique was coupled with the BN model. The performance of the BN models was finally compared with those of the existing empirical equations and the developed ANN model as well.

Concept and background

The pollution can be dispersed longitudinally, transversely and vertically by advection and dispersion processes. Once the cross-sectional mixing is completed, the longitudinal dispersion becomes the most important process. In this case, the one-dimensional (1D) dispersion equation is widely used for unsteady non-uniform flow (Sahin 2014). The general form of this equation, advection–diffusion equation, can be described as (Taylor 1953);

$$\left(\frac{\partial C}{\partial t}\right) + U\left(\frac{\partial C}{\partial x}\right) = \frac{\partial}{\partial x}\left(K\frac{\partial C}{\partial x}\right) \tag{1}$$

where C is the cross-sectional average concentration (kg/m^3); U is the cross-sectional average velocity (m/s); x is the direction of the mean flow; t is time in seconds (s); and K denotes the longitudinal dispersion coefficient (m^2/s) in the x direction.

Equation 1 holds only after the so-called initial period is reached (Deng et al. 2001; Noori et al. 2011). Solutions of this equation can be obtained with appropriate initial and boundary conditions. However, the most challenging issue is the estimation of the longitudinal dispersion coefficient, which has highly nonlinear nature in natural rivers and streams. Some of the most influencing parameters for accurate estimation of the longitudinal dispersion coefficient are: density, viscosity, channel width, flow depth, mean velocity, shear velocity, bed slope, bed roughness, horizontal stream curvature (sinuosity) and bed shape factor (Guymer 1998; Seo and Cheong 1998). In this regard, many statistical models, analytical solutions and experimental works have been developed to estimate K in natural rivers.

Elder expanded Taylor's method for an open channel of infinite width (Elder 1959). Based on experimental measurements and supposing a logarithmic distribution for the velocity profile in the vertical direction, he suggested:

$$K = 5.93HU_* \tag{2}$$

where H is the depth of flow and U_* is the bed shear velocity. In this equation, the transverse variation in the velocity profile was not taken under consideration. It leads an underestimated prediction of K, because the transverse shear is more important than the vertical one in most natural streams.

Afterward, Fischer (1975) and Seo and Cheong (1998) presented the following equations, respectively:

$$\frac{K}{HU_*} = 0.011\left(\frac{U}{U_*}\right)^2\left(\frac{W}{H}\right)^2 \tag{3}$$

$$\frac{K}{HU_*} = 5.915\left(\frac{U}{U_*}\right)^{1.428}\left(\frac{W}{H}\right)^{0.62} \tag{4}$$

Kashefipour and Falconer (2002) used 81 sets of field data in USA and based on the dimensional and regression analysis developed Eqs. (5) and (6):

$$K = 10.612HU\left(\frac{U}{U_*}\right) \tag{5}$$

$$K = \left[7.428 + 1.775\left(\frac{W}{H}\right)^{0.62}\left(\frac{U_*}{U}\right)^{0.572}\right]HU\left(\frac{U}{U_*}\right) \tag{6}$$

They suggested that for open-channel flows with W/H greater and <50, Eqs. (5) and (6) can be applied for practical applications, respectively.

Toprak and Cigizoglu (2008) demonstrated that an accurate prediction of K can be obtained by using ANN models. Using genetic algorithm, Sahay and Dutta (2009) proposed Eq. (7) and Tayfur (2009) presented Eq. (8) for predicting K.

$$\frac{K}{HU_*} = 2\left(\frac{U}{U_*}\right)^{1.25}\left(\frac{W}{H}\right)^{0.96} \tag{7}$$

$$K = 0.91Q + 9.94 \tag{8}$$

in which Q is the flow discharge.

Etemad-Shahidi and Taghipour (2012) derived the following descriptions for longitudinal dispersion coefficient based on model tree approach:

$$\frac{K}{HU_*} = 15.49\left(\frac{W}{H}\right)^{0.78}\left(\frac{U}{U_*}\right)^{0.11}; \quad \text{if } \frac{W}{H} < 30.6 \tag{9}$$

$$\frac{K}{HU_*} = 14.12\left(\frac{W}{H}\right)^{0.61}\left(\frac{U}{U_*}\right)^{0.85}; \quad \text{if } \frac{W}{H} > 30.6 \tag{10}$$

Li et al. (2013) employed differential evolution for prediction of K in natural streams. The application revealed that the proposed approach is better than other expressions. The equation can be expressed as follows:

$$\frac{K}{HU_*} = 2.282\left(\frac{W}{H}\right)^{0.7613}\left(\frac{U}{U_*}\right)^{1.4713} \tag{11}$$

Zeng and Huai (2014) developed a new equation for predicting K as:

$$K = 5.4\left(\frac{W}{H}\right)^{0.7}\left(\frac{U}{U_*}\right)^{0.13}HU \tag{12}$$

They showed that Eq. (12) can predict a longitudinal dispersion coefficient well, especially for rivers within $20 < W/H < 100$.

Disley et al. (2015) applied field data of small, steep streams in Ontario to develop a predictive equation for longitudinal dispersion coefficient. Their predictive equation relates the longitudinal dispersion coefficient to hydraulic and geometric parameters of the stream and the Froude number. Using multiple regression analysis, they presented the following equation:

$$\frac{K}{HU_*} = 3.563\left(\frac{U}{\sqrt{gH}}\right)^{-0.4117}\left(\frac{W}{H}\right)^{0.6776}\left(\frac{U}{U_*}\right)^{1.0132} \tag{13}$$

Methodology

Bayesian network (BN)

A Bayesian network, which is probabilistic based, is a set of nodes representing random variables and a set of links for connecting the nodes in an acyclic manner. The BN is a powerful modeling tool for complex problems because it can describe numerous relevant factors simultaneously and express their relationship effectively and provides a mechanism to incorporate many kinds of prior information and expert knowledge into learning to solving problems with many uncertainties (Antal et al. 2004; Leu and Bui 2016). In the BN, a link from A and B to C means that variables A and B are the parent and C is the child. The dependencies are quantified by conditional probabilities for each node given its parents in the network. Bayesian networks can be used for two kinds of random variables: continuous chance nodes with a continuous infinite state space (Fig. 1a) and discrete chance nodes with a discrete finite state space (Fig. 1b). For the discrete chance nodes, the function describing how the node depends on its parents is a conditional probability table.

In a BN analysis, for n number of mutually exclusive parameters X_i ($i = 1, 2,\ldots n$), and a given observed data Y, the updated probability is computed by:

$$p(X_i|Y) = \frac{p(Y|X_i) * p(X_i)}{\sum_j p(Y|X_j)p(X_j)} \tag{14}$$

where $p(X_i|Y)$ demonstrates the posterior probability occurrence of X given the condition that Y occurs, $p(X)$ denotes the prior probability occurrence of X, $p(Y)$ denotes the marginal (total) probability occurrence of Y and is effectively constant since the obtained data are in hand, and $p(Y_i|X)$ refers to the conditional probability occurrence of Y given that X occurs too (often viewed in this sense as the likelihood distribution) (Pearl 1988). Learning and inferences can be drawn upon professional algorithm in the BN. More details about the mathematical background of the BN approach can be found in Jensen (2001) and Malekmohammadi et al. (2009).

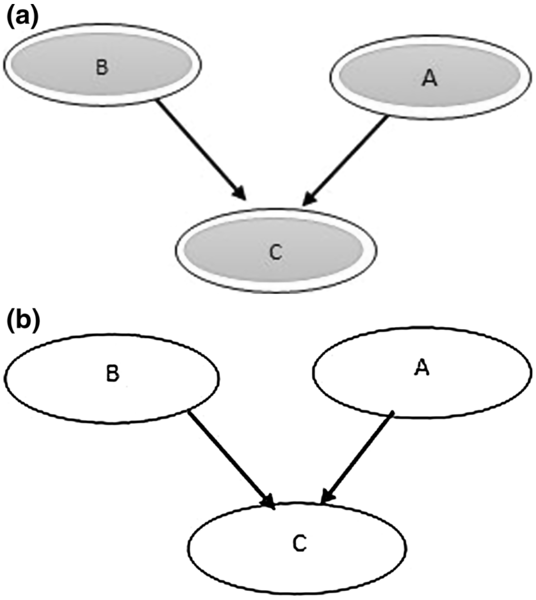


Fig. 1 Bayesian networks applications: **a** the network with continuous chance nodes, **b** the network with discrete chance nodes

Artificial neural network (ANN)

Over the past decade, artificial neural networks have been widely used to deal with complex and nonlinear problems, especially those with unknown relationship between input variables and output parameter (Azid et al. 2014). They were extensively used for hydrological modeling, and it was demonstrated that they are superior over the conventional regression-based models. Therefore, in this study, they have been employed as an alternative to compute longitudinal dispersion coefficient and to compare the results with those of BN models. ANN is an approximation function mapping inputs to outputs. A typical network consists of three layers of neurons, namely input, hidden and output, in which each neuron acts as an independent computational element. Input layer is defined as a layer of neurons receiving inputs (I_i) directly from outside the network. Layer of a network that is not connected to the network output (O) called hidden layer and layer whose output is passed to the world outside the network is output layer. In this study, the log-sigmoid (logsig) transfer function was used to calculate the layer's output from the net input. Figure 2 shows a schematic layout of ANN.

Different parameters can affect the accuracy of ANN models such as number of hidden layers and hidden neurons, training algorithm and data selection. In this study, the Levenberg–Marquardt back-propagation algorithm which is a second-order nonlinear optimization technique has been used in training of the ANN. This algorithm is usually faster and more reliable than any other back-propagation techniques (Hagan and Menhaj 1994; Ham and Kostanic 2001). The training data consist of a set of

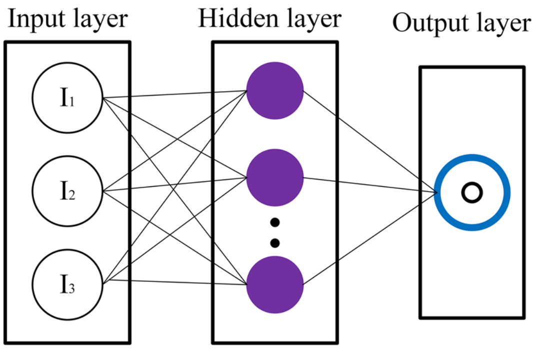


Fig. 2 Schematic layout of ANN

N training samples (x_s, t_s), where s is the sample number, t_s is the target value, and x_s is the N-dimensional input vector for the Sth training sample and y_s is the M-dimensional output vector of the trained network for the Sth sample.

In this study, the best topology (architecture) of the ANN used for modeling longitudinal dispersion coefficient compromises of two input variables of W/H, U/U_*, 8 neurons in the hidden layer and K/U_*H as the target variable.

The general performance of the ANN is measured by the mean square error (MSE):

$$E = \frac{1}{N}\sum_{S=1}^{N} E_s = \frac{1}{N}\sum_{S=1}^{N}\sum_{i=1}^{N}[t_s(i) - y_s(i)]^2 \tag{15}$$

Data clustering approach

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other clusters. Different methods can be applied for data clustering (e.g., *k-means*, fuzzy C-means and subtractive clustering). In this study, *k-means* approach first used by MacQueen (1967) was employed in order to cluster the input and output data. It is easily implemented and understandable and is popular for cluster analysis in data mining. The process of clustering by *k*-means method is illustrated in Fig. 3. Regarding Fig. 3, the method initializes with a number of clusters and a center for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest center. Afterward, the procedure is repeated by re-calculating new centroids in order to achieve the desired criterion.

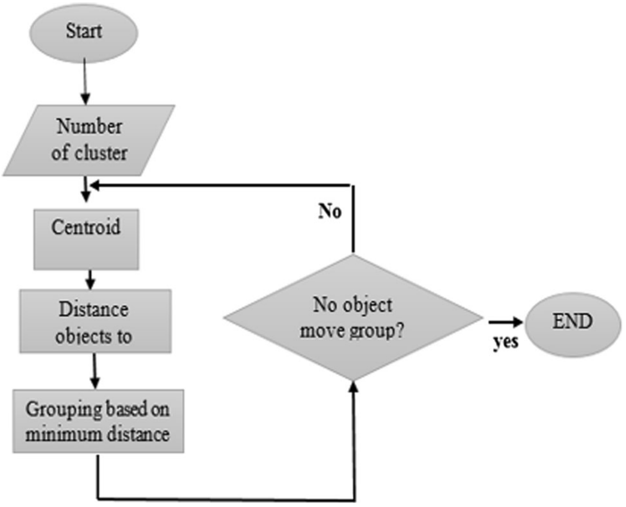


Fig. 3 Flowchart of the *k*-means method

Given a set of observations ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\boldsymbol{S} = \{S_1, S_2, \ldots, S_k\}$ to minimize the within-cluster sum of squares (sum of distance functions of each point in the cluster to the k center). In other words, its objective is to find:

$$\min \sum_{i=1}^{K}\sum_{x\in s_i} x - \mu_i^2 \tag{16}$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

Data and statistical analysis

In the present study, a wide range of field data set including 149 samples measured in different rivers (collected by (Etemad-Shahidi and Taghipour 2012) was used for applying cluster-based Bayesian network in the estimation of the longitudinal dispersion coefficient. The parameters, captured by monitoring process, contain geometric and hydraulic characteristics such as channel width, channel depth, average velocity, shear velocity and longitudinal dispersion coefficient. These data include longitudinal characteristics for different rivers which are reported by different researchers. It is necessary to mentioned that the dimensionless parameters were used to predict the longitudinal dispersion coefficient.

Table 1 shows the statistical analysis of the data sets including minimum (min) and maximum (max), average (mean) and standard deviation (SD). Data to be used for training step should be sufficiently large to cover the possible known variations of the important parameters in the problem domain. An attempt was made to select data in a way that the testing data follow the normal distribution of training data. For training and testing steps, 120 and 29 data samples were, respectively, selected. Moreover, the minimum and maximum values of the target variable were used in the training set. Figures 4, 5 and 6 illustrate the variations of the testing and training data with W/H, U/U_* and K/U_*H, respectively.

Modeling strategy

To find the most accurate predictive BN model, two types of BN models will be considered based on applying dimensional and dimensionless parameters. These strategies were carried out to examine how the input variables can affect the model's performances. In this way, three modeling strategies will be examined. In the first strategy, four parameters including W, H, U and U_* were inserted as input variables in the BN structure to predict dispersion coefficient in dimensional form. Two other strategies were

Table 1 Summary statistical analysis of the total, training and testing data sets

Parameter	Min	Max	Mean	SD
Total data sets				
W (m)	1.4	711.2	60.1	91.1
H (m)	0.14	19.94	1.55	2.12
U (m/s)	0.03	1.73	0.47	0.32
U_* (m/s)	0.002	0.553	0.082	0.063
K_x (m^2/s)	0.2	891.9	73.5	137.7
W/H	2.2	403.8	47.5	47.1
U/U_*	0.77	19.87	6.94	4.56
K/U_*H	3.1	5500.0	745.7	947.6
Training data sets				
W (m)	1.4	537.4	58.5	80.0
H (m)	0.14	8.90	1.51	1.60
U (m/s)	0.03	1.73	0.49	0.34
U_* (m/s)	0.002	0.553	0.087	0.068
K_x (m^2/s)	0.2	891.9	79.9	150.4
W/H	2.2	403.8	47.1	50.3
U/U_*	0.77	19.87	6.84	4.47
K/U_*H	3.1	5500.0	700.5	884.8
Testing data sets				
W (m)	10.0	711.2	66.2	127.0
H (m)	0.30	19.94	1.72	3.53
U (m/s)	0.10	0.68	0.37	0.17
U_* (m/s)	0.006	0.170	0.063	0.031
K_x (m^2/s)	1.4	237.2	48.4	62.6
W/H	10.4	138.5	49.0	31.6
U/U_*	1.21	19.06	7.35	4.93
K/U_*H	48.7	5500.0	922.0	1151.4

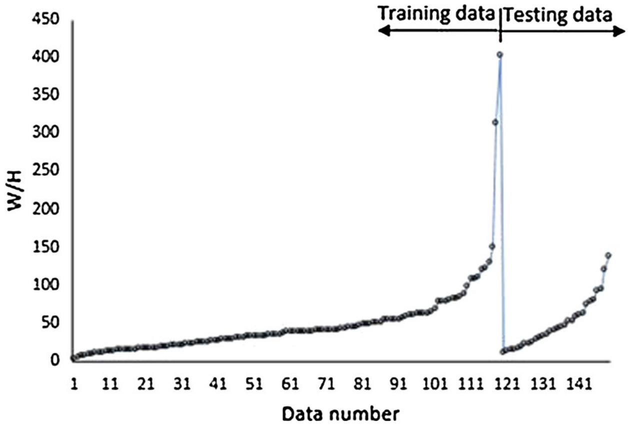


Fig. 4 Variations of training and testing data with $\frac{W}{H}$

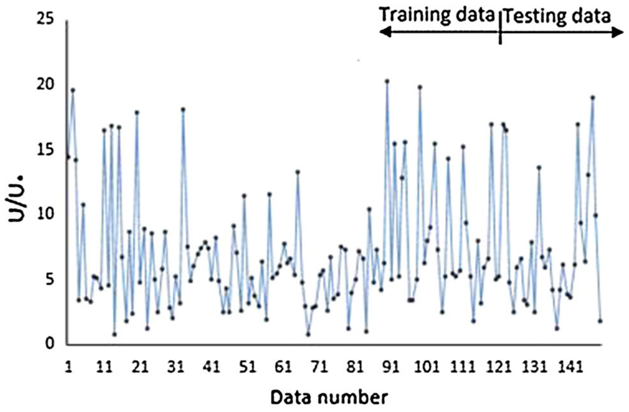


Fig. 5 Variations of training and testing data with $\frac{U}{U_*}$

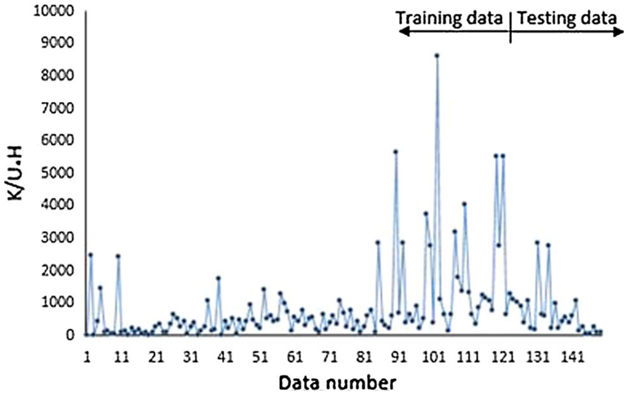


Fig. 6 Variations of training and testing data with $\frac{K}{U_*H}$

developed based on two dimensionless parameters of W/H and U/U_* for predicting dimensionless parameter of K/U_*H. One of them was developed based on clustering Bayesian network, and the other one was established without data clustering.

Figure 7 shows the structure of BN models for predicting dimensional and dimensionless dispersion coefficient. In Fig. 7a, b, by considering existing numeric data, the continuous BN was applied while, in Fig. 7c, discrete network was used as the data were in clusters 1–10 for each node.

In the case of clustering approach, the numeric data were divided into 10 clusters and these clusters were introduced to the main model. The choice of 10 clusters was determined through a trial and error procedure. In this regard, different clusters were examined by k-means clustering approach to provide the best results. In the cluster-based BN model, the target value is computed using mathematical expectation, $E[X]$:

$$E[X] = \frac{\sum_{i=1}^{10} x_i p_i}{\sum_{i=1}^{10} p_i} \tag{17}$$

where x_i is the average of ith cluster and p_i is the probability of the cluster.

Furthermore, an ANN model with the same input and output variable was developed. The ANN model was

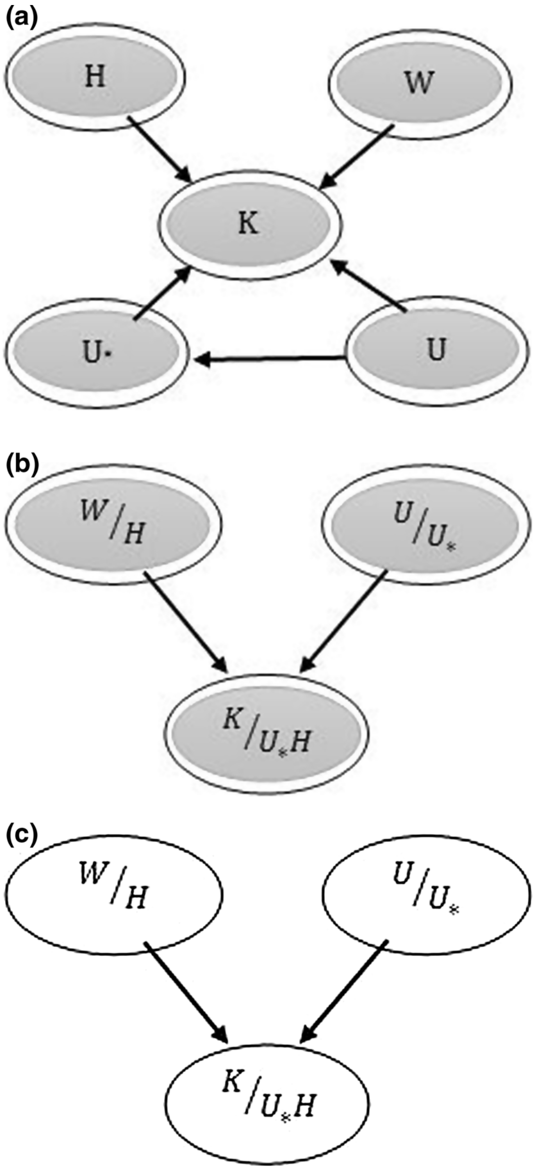


Fig. 7 Modeling strategies: **a** BN for dimensional data, **b** BN for dimensionless data **c** BN-cluster data of the discrete network

trained by Levenberg–Marquardt algorithm with 8 neurons in the hidden layer, which were assigned through a trial and error procedure. The data sets were normalized using the following equation:

$$Q_{\mathrm{i}}' = \frac{Q_{\mathrm{i}} - Q_{\min}}{Q_{\max} - Q_{\min}} \tag{18}$$

where X_{i}' is the normalized data, x_i is the observed value of the variable, and $x_{\max}$ and $x_{\min}$ are the maximum and minimum of the parameter, respectively.

In all BN and ANN models, 80% of the data were applied for training while 20% of the data were used for testing.

Performance evaluation criteria

The performance of different models may be evaluated using the coefficient of determination (R^2), the root mean square error (RMSE), the discrepancy ratio (DR) and the coefficient of efficiency (E) (Johnson and Omland 2004; Barati 2011; Omole et al. 2013; Barati 2013; Barati et al. 2014a,b; Hosseini et al. 2016). Generally, the models' predictions are optimum if R^2, $RMSE$, DR and E are found to be close to 1, 0, 0 and 1, respectively. Accuracy is defined as the percentage of DR values that fall between −0.3 and 0.3 (Kashefipour and Falconer 2002; Seo and Cheong 1998). DR values greater than 0.3 ($DR > 0.3$) and smaller than −0.3 ($DR < -0.3$) indicate overestimation and underestimation for the dispersion coefficient, respectively. The above-mentioned indices (i.e., R^2, $RMSE$, DR *and* E) are described as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(k_{i(\text{measured})} - \bar{k}_{(\text{measured})}\right)\left(k_{i(\text{predicted})} - \bar{k}_{(\text{predicted})}\right)\right]^2}{\sum_{i=1}^{n}\left(K_{xi(\text{measured})} - \bar{k}_{x(\text{measured})}\right)^2 \sum_{i=1}^{n}\left(K_{xi(\text{predicted})} - \bar{k}_{x(\text{predicted})}\right)^2} \tag{19}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(K_{i(\text{measured})} - K_{i(\text{predicted})}\right)^2}{n}} \tag{20}$$

$$DR = \log\frac{K_{(\text{predicted})}}{K_{(\text{measured})}} \tag{21}$$

$$E = 1 - \frac{\sum_{i=1}^{n}\left(K_{i(\text{measured})} - K_{i(\text{predicted})}\right)^2}{\sum_{i=1}^{n}\left(K_{i(\text{measured})} - \bar{k}_{(\text{measured})}\right)^2} \tag{22}$$

where K and n, respectively, are the longitudinal dispersion and the number of data. In Eq. (17), $\bar{k}_{(\text{measured})}$ and $\bar{k}_{(predicted)}$ represent mean values of the measured and predicted longitudinal dispersion coefficient.

Results and discussion

In the present study, a model was introduced to increase the efficiency of the existing estimating models for the longitudinal dispersion coefficient (K) in natural rivers. In this regard, different types of BN models were developed and organized to assess the efficiency of the BN approach in dimensional and dimensionless frameworks. The best structure for the BN, an ANN model based on Levenberg–Marquardt algorithm, was developed, and the performance of BN models and the ANN model were assessed by R^2, RMSE and DR indices. The results obtained through testing step are presented in Table 2. According to this table, the best model performances were achieved by the cluster-based BN model. It shows that the Bayesian network models give a more accurate estimation of dispersion

Table 2 Results of the different developed models during testing step

Model	RMSE	R^2	Accuracy (%)
BN-dimensional	1183.2	0.32	53
BN-dimensionless	836.6	0.555	70
ANN	893	0.428	47
BN-cluster	791.95	0.764	83

coefficient compared with the ANN model. For example, the accuracy of BN models with and without clustering data (i.e., the BN-cluster and BN-dimensionless), respectively, is 83 and 70%, while the accuracy of ANN model is less than 50%, which is far less than BN models.

The results of the dimensional and dimensionless BN models were compared in terms of *RMSE*, R^2 and accuracy. This evaluation shows that appropriate selection of the input variables plays an important role in model performance. Based on the results, input dimensional variables (W, H, U, U_*) decrease the model performances, whereas input dimensionless variables of $\frac{W}{H}$ and $\frac{U}{U_*}$ significantly improve the model accuracy. Table 2 also confirms that the BN-clustering method improves the performance of the main model about 13% and 0.22 in terms of accuracy and R^2, respectively. The model also decreases *RMSE* of the main model by 5%. For the BN-cluster model, the observed and predicted values of dispersion coefficient for testing data set are presented in Fig. 8.

The BN-cluster model provides predictions closer to the measured values of the longitudinal dispersion coefficient in both low and average values of *K*. Moreover, its predictions of the extreme values are acceptable. The proposed model predicts the extreme values of 177.7 and 227.6 about 107.42 and 224.67, respectively.

For further discussion about the accuracy of the developed model, the results are represented against $\frac{W}{H}$ and $\frac{U}{U_*}$ in Fig. 9a, b, respectively. It is concluded that the BN-cluster model is more sensitive to $\frac{U}{U_*}$ than $\frac{W}{H}$ (i.e., hydraulic parameters than geometric parameters). The applied model shows a good performance in Fig. 9a for $\frac{W}{H} < 80$ and also $\frac{W}{H} > 120$. Good agreement between observed and predicted values of the dispersion coefficient is shown in Fig. 9b when $\frac{U}{U_*}$ is <10, whereas for $\frac{U}{U_*} > 10$ a reliable conclusion cannot be extracted.

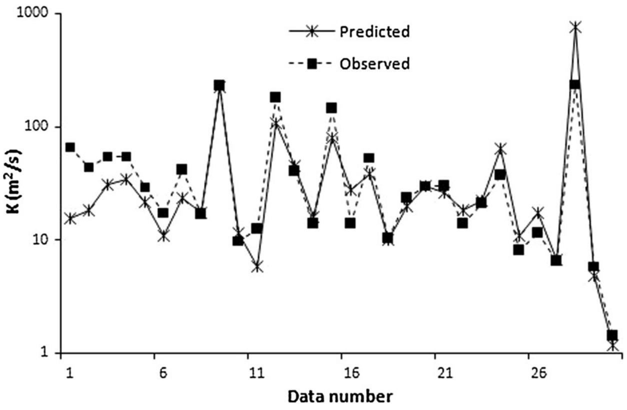


Fig. 8 Performance of BN-cluster model during testing step

The capability and efficiency of the proposed model (the BN-cluster) was also checked with the existing models from previous studies. In this regard, the results of the present study for the explained testing data were compared with those of existing equations in Table 3. Generally, the best accurate model is presented by lowest value of RMSE (close to 0) and highest values of accuracy, coefficient of efficiency (*E*) and R^2. Considering all these performance indicators, the superiority of the developed model (the BN-cluster) over the other models is approved. At first it can be seen in Table 3 that some existing models, such as Fischer (1975) and Seo and Cheong (1998), have low accuracy, and one can easily find out that these models are not

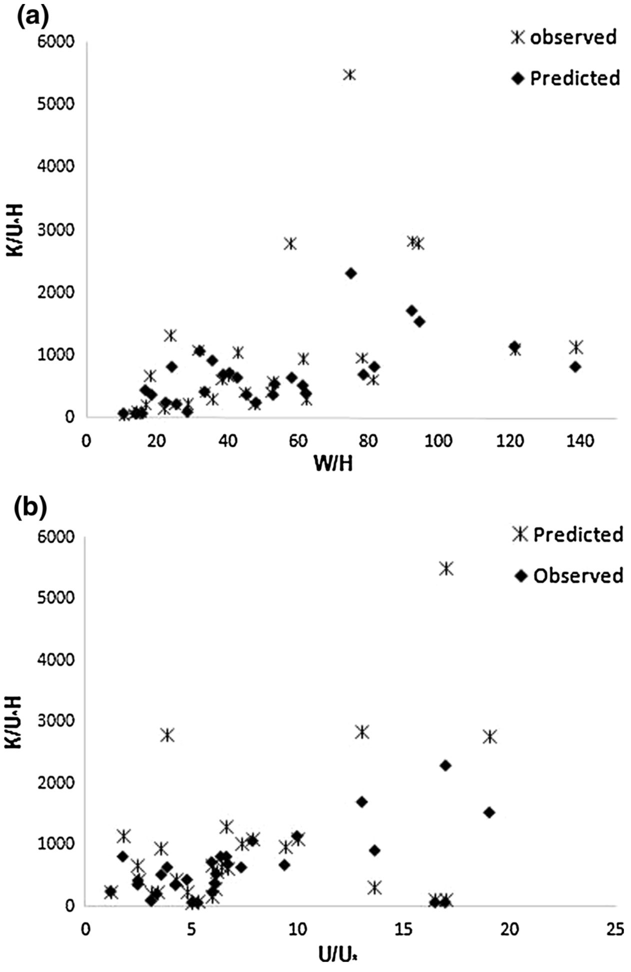


Fig. 9 Observed and predicted $\frac{K}{U_* H}$ against **a** $\frac{W}{H}$ **b** $\frac{U}{U_8}$

Table 3 Comparing the results of this study with some well-known equations

Model	RMSE	*E*	R^2	Accuracy (%)
Fischer (1975)	7432.36	−41.235	0.4458	36.6
Liu (1977)	1922.7	−1.826	0.3168	53.3
Seo and Cheong (1998)	1151.2	−0.013	0.48	43.3
Kashefipour and Falconer (2002)	1069.18	0.126	0.2619	56.6
Sahay and Dutta (2009)	1052.39	0.153	0.4988	43.3
Etemad-Shahidi and Taghipour (2012)	838.7	0.462	0.5118	70
Disley et al. (2015)	886.46	0.399	0.402	66.6
BN-cluster	791.95	0.521	0.764	83

capable to give acceptable results. However, other models have promising results. The proposed model outperforms the existing equations in terms of coefficient of efficiency remarkably. The highest values of E are obtained from the present model and Eqs. 9 and 10.

The ability of the predictive models is also compared with the present model in terms of *RMSE* and R^2 in Figs. 10 and 11, respectively. The scatter plots are obtained through the BN-cluster model and those of Kashefipour and Falconer (2002), Etemad-Shahidi and Taghipour (2012), Sahay and Dutta (2009) and Disley et al. (2015).

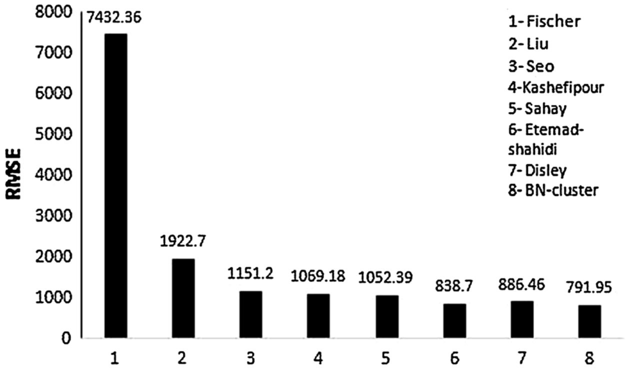
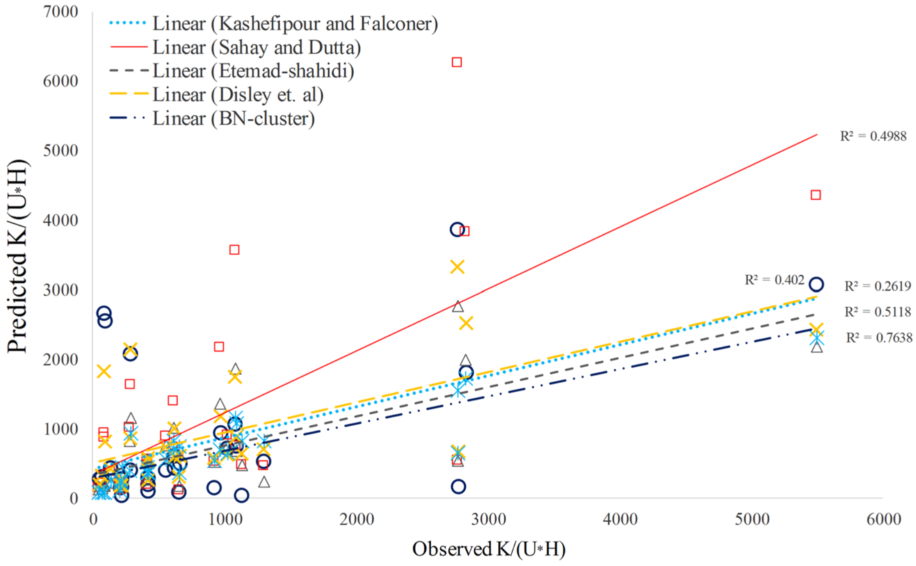


Fig. 10 RMSE of different models for testing data set

The present model has the highest values of the accuracy and R^2 and the lowest value of RMSE. Regarding R^2 values (see Fig. 11; Table 3), the BN-cluster is recognized as the most accurate model. The highest value of R^2 is about 0.764, which represents a good correlation between estimated and observed values. The model is then followed by Etemad-Shahidi and Taghipour (2012) with $R^2 = 0.5118$. Therefore, the present model provides better performance in terms of R^2. The lowest value of the RMSE is about 791.95 that is related to the BN-cluster model. It is also followed by Etemad-Shahidi and Taghipour (2012) with low value of $RMSE = 838.7$. The proposed models by Kashefipour and Falconer (2002) and also Fischer (1975) show relatively high values of RMSE for $\frac{K}{U_*H}$. The BN-cluster model improves the accuracy of the existing models from 16 to 56% in terms of the accuracy criterion.

To provide more details about the accuracy of the different models, DR value of the developed model and existing models is compared as shown in Fig. 12. A large

Fig. 11 R^2 of different models for testing data



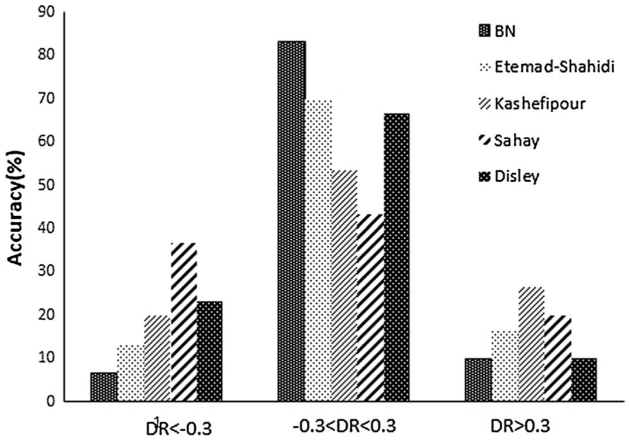


Fig. 12 Comparison of DR values of the developed model and existing models

number of estimated values of $\frac{K}{U_{*}H}$ by the BN-cluster model have DR values ranging from −0.3 to 0.3, which confirms the accuracy of the proposed model for predicting dispersion coefficient. Moreover, it has roughly symmetric distribution of DR values. Sahay and Dutta (2009) show the most frequency, out of the range −0.3 to 0.3, which introduces the least accuracy in this comparison. In Disley et al. (2015) and Sahay and Dutta (2009), positive skewness can be observed that implies its overestimation. Instead, Kashefipour and Falconer (2002) underestimate the $\frac{K}{U_{*}H}$ values overall.

Conclusion

Longitudinal dispersion coefficient is one of the most important factors for surface water quality modeling. This study presents a new approach of Bayesian network model with application of cluster data to predict the longitudinal dispersion coefficient in natural streams. To do this, a large number of field data were clustered by *k-means* approach and inserted as input and output variables for structural learning of the probabilistic model (i.e., the BN model). To assess the efficiency of the proposed model, separate ANN model with the same input variables was developed. Moreover, the efficiency of the cluster-based BN model was compared with some of the existing equations for dispersion coefficient. Results of this study revealed that application of dimensionless parameters $\frac{W}{H}$ and $\frac{U}{U_{*}}$ provides more accurate predictions of $\frac{K}{U_{*}H}$ by the BN model, compared to dimensional parameters W, H, U and U_{*}. The BN model including dimensionless input variables increases the performance of the dimensional BN model in terms of R^2 from 0.32 to 0.555 and accuracy from 53 to 70%. Also, RMSE of BN model using dimensional and dimensionless input variables is 1183.2 and 836.6, respectively. It can also be concluded that clustering data linked to the main BN model significantly improves the model efficiency. The results indicated that the most accurate model is the BN-cluster model with highest values of R^2 and accuracy (0.764 and 83%) and lowest value of *RMSE* (791.95). The model also greatly outperforms ANN model in terms of the accuracy, R^2 and *RMSE*.

A comparison between the BN-cluster model and the existing equations for predicting dispersion coefficient demonstrated that the present BN model is more accurate. For the testing data, the BN model improves the accuracy of the previous predictive models in the range of 13–46%. The model minimally improved R^2 of the previous models about 0.25 which is a remarkable improvement. Also, the minimum RMSE among previous tested models is related to the present BN model. Briefly, the BN-cluster model provides more accurate prediction of dispersion coefficient than the existing models. Considering the coefficient of efficiency (E), it can be derived that the BN model has the highest value ($E = 0.52$) which indicates its superiority over the predictive equations.

Generally, this work showed a conjunctive model of clustering approach and Bayesian network can be successfully applied for predicting dispersion coefficient in natural streams. The proposed model gives acceptable predictions of dispersion coefficient for rivers with a wide range of geometric and hydraulic characteristics. The cluster-based BN model showed a great ability for predicting low and medium values of dispersion coefficient with high correlation between measured and predicted values. In future research, it is hoped that the procedure of this study can be successfully applied for other problems in the field of river and environmental engineering.

Acknowledgements The authors would like to thank Majid Rahimzadegan and Hamid Kardan Moghadam for their comments to improve the manuscript.

References

Alizadeh MJ, Kavianpour MR (2015) Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean. Mar Pollut Bull 98(1):171–178

Altunkaynak A (2016) Prediction of longitudinal dispersion coefficient in natural streams by prediction map. J Hydro-Environ Res 12:105–116

Antal P, Fannes G, Timmerman D, Moreau Y, De Moor B (2004) Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. Artif Intell Med 30(3):257–281

Azamathulla HM, Ghani AA (2011) Genetic programming for predicting longitudinal dispersion coefficients in streams. Water Resour Manag 25(6):1537–1544

Azamathulla HM, Wu FC (2011) Support vector machine approach for longitudinal dispersion coefficients in natural streams. Appl Soft Comput 11(2):2902–2905

Azid A, Juahir H, Toriman ME, Kamarudin MKA, Saudi ASM, Hasnam CNC et al (2014) Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. Water Air Soil Pollut 225(8):1–14

Barati R (2011) Parameter estimation of nonlinear Muskingum models using Nelder-Mead simplex algorithm. J Hydrol Eng 16(11):946–954

Barati R (2013) Application of excel solver for parameter estimation of the nonlinear Muskingum models. KSCE J Civ Eng 17(5):1139–1148

Barati R, Neyshabouri SAAS, Ahmadi G (2014a) Development of empirical models with high accuracy for estimation of drag coefficient of flow around a smooth sphere: an evolutionary approach. Powder Technol 257:11–19

Barati R, Neyshabouri SS, Ahmadi G (2014b) Sphere drag revisited using shuffled complex evolution algorithm. In: River flow

Deng ZQ, Singh VP, Bengtsson L (2001) Longitudinal dispersion coefficient in straight rivers. J Hydraul Eng 127(11):919–927

Deng ZQ, Bengtsson L, Singh VP, Adrian DD (2002) Longitudinal dispersion coefficient in single-channel streams. J Hydraul Eng 128(10):901–916

Disley T, Gharabaghi B, Mahboubi A, McBean E (2015) Predictive equation for longitudinal dispersion coefficient. Hydrol Process 29(2):161–172

Elder J (1959) The dispersion of marked fluid in turbulent shear flow. J Fluid Mech 5(04):544–560

Etemad-Shahidi A, Taghipour M (2012) Predicting longitudinal dispersion coefficient in natural streams using M5′ model tree. J Hydraul Eng 138(8):542–554

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

Fischer HB (1975) Discussion of "simple method for predicting dispersion in streams". J Environ Eng Div 101(3):453–455

Fischer HB, List JE, Koh CR, Imberger J, Brooks NH (1979) Mixing in inland and coastal waters. Elsevier, Amsterdam

Guymer I (1998) Longitudinal dispersion in sinuous channel with changes in shape. J Hydraul Eng 124(1):33–40

Hagan MT, Menhaj MB (1994) Training feedforward networks with the Marquardt algorithm. IEEE Trans Neural Netw 5(6):989–993

Ham F, Kostanic I (2001) Fundamental neurocomputing concepts. Principles of Neurocomputing for Science and Engineering, Arnold Publishers, London, pp 24–91

Hosseini K, Nodoushan EJ, Barati R, Shahheydari H (2016) Optimal design of labyrinth spillways using meta-heuristic algorithms. KSCE J Civ Eng 20(1):468–477

Jensen FV (2001) Bayesian networks and decision graphs. Springer, Berlin

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19(2):101–108

Kashefipour SM, Falconer RA (2002) Longitudinal dispersion coefficients in natural channels. Water Res 36(6):1596–1608

Leu SS, Bui QN (2016) Leak Prediction Model for Water Distribution Networks Created Using a Bayesian Network Learning Approach. Water Resour Manag 30(8):2719–2733

Li X, Liu H, Yin M (2013) Differential evolution for prediction of longitudinal dispersion coefficients in natural streams. Water Resour Manag 27(15):5245–5260

Liu H (1977) Predicting dispersion coefficient of streams. J Environ Eng Div 103(1):59–69

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA, (vol 1, 14, 281–297)

Madadgar S, Moradkhani H (2014) Spatio-temporal drought forecasting within Bayesian networks. J Hydrol 512:134–146

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

Matheussen BV, Granmo O-C (2015) Modeling snow dynamics using a Bayesian network. In: Current approaches in applied artificial intelligence. Springer, Berlin, pp 382–393

Najafzadeh M, Tafarojnoruz A (2016) Evaluation of neuro-fuzzy GMDH-based particle swarm optimization to predict longitudinal dispersion coefficient in rivers. Environ Earth Sci 75(2):1–12

Noori R, Karbassi AR, Mehdizadeh H, Vesali-Naseh M, Sabahi MS (2011) A framework development for predicting the longitudinal dispersion coefficient in natural streams using an artificial neural network. Environ Prog Sustain Energy 30(3):439–449

Nourani V, Alami MT, Aminfar MH (2009) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Eng Appl Artif Intell 22(3):466–472

Omole DO, Longe EO, Musa AG (2013) An approach to reaeration coefficient modeling in local surface water quality monitoring. Environ Model Assess 18(1):85–94

Pearl J (1988) Probabilistic inference in intelligent systems. Morgan Kaufmann, San Mateo

Perucca E, Camporeale C, Ridolfi L (2009) Estimation of the dispersion coefficient in rivers with riparian vegetation. Adv Water Resour 32(1):78–87

Sahay RR (2011) Prediction of longitudinal dispersion coefficients in natural rivers using artificial neural network. Environ Fluid Mech 11(3):247–261

Sahay RR, Dutta S (2009) Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm. Hydrol Res 40(6):544–552

Sahin S (2014) An empirical approach for determining longitudinal dispersion coefficients in rivers. Environ Process 1(3):277–285

Seo IW, Baek KO (2004) Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams. J Hydraul Eng 130(3):227–236

Seo IW, Cheong TS (1998) Predicting longitudinal dispersion coefficient in natural streams. J Hydraul Eng 124(1):25–32

Swamee PK, Pathak SK, Sohrab M (2000) Empirical relations for longitudinal dispersion in streams. J Environ Eng 126(11):1056–1062

Tayfur G (2009) GA-optimized model predicts dispersion coefficient in natural channels. Hydrol Res 40(1):65–78

Tayfur G, Singh VP (2005) Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. J Hydraul Eng 131(11):991–1000

Taylor G (1953) Dispersion of soluble matter in solvent flowing slowly through a tube. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol 219, no 1137, The Royal Society, pp 186–203

Toprak ZF, Cigizoglu HK (2008) Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods. Hydrol Process 22(20):4106–4129

Tutmez B, Yuceer M (2013) Regression kriging analysis for longitudinal dispersion coefficient. Water Resour Manag 27(9):3307–3318

Wang Y, Huai W (2016) Estimating the longitudinal dispersion coefficient in straight natural rivers. J Hydraul Eng 142(11):04016048

Xu M, Zeng G, Xu X, Huang G, Jiang R, Sun W (2006) Application of Bayesian regularized BP neural network model for trend analysis, acidity and chemical composition of precipitation in North Carolina. Water Air Soil Pollut 172(1–4):167–184

Zeng Y, Huai W (2014) Estimation of longitudinal dispersion coefficient in rivers. J Hydro-Environ Res 8(1):2–8