Predicting Longitudinal Dispersion Coefficient in Natural Streams Using M5′ Model Tree

Amir Etemad-Shahidi, Ph.D.[1]; and Milad Taghipour[2]

Abstract: The longitudinal dispersion coefficient is a key parameter in determining the distribution of pollution concentration, especially in temporally time-varying source cases after full cross-sectional mixing has occurred. Several studies have been performed to present simple formulas to predict it. However, they may not always result in an accurate prediction because of the complexity of the phenomenon. In this study, a M5′ model tree was used to develop a new model for predicting the longitudinal dispersion coefficient. The main advantages of the model trees are that (1) they provide transparent formulas and offer more insight into the obtained formulas and (2) they are more convenient to develop and employ compared with other soft computing methods. To develop the model tree, extensive field data sets consisting of hydraulic and geometrical characteristics of different rivers were used. By using error measures, the performance of the model was also compared with the performance of other existing equations. Overall, the results showed that the developed model outperforms the existing formulas and can serve as a valuable tool for predicting of the longitudinal dispersion coefficient. **DOI: 10.1061/(ASCE)HY.1943-7900.0000550.** © *2012 American Society of Civil Engineers.*

CE Database subject headings: Rivers and streams; Water pollution; Predictions; Coefficients.

Author keywords: Longitudinal dispersion; M5′ model tree; Spill modeling; River; Sinuosity.

Introduction

In rivers, longitudinal dispersion is the predominant mechanism in mixing of the tracer by several orders of magnitude when cross-sectional mixing is complete. This leads to the elimination of any further concentration gradient (Fischer et al. 1979). The dispersion coefficient plays an important role in spill modeling, in the design of water intakes, and in the outfall and treatment plants. It is representative of intensity of the mixing in rivers (Deng et al. 2002). Hence, accurately estimating the longitudinal dispersion coefficient is of great importance for engineers and scientists. A direct estimation of the dispersion coefficient by experimental means requires expensive and time-consuming tracer studies. As a result, the demand for a coefficient prediction tool still exists. Estimation of the longitudinal dispersion coefficient has been received considerable attention for a long time (e.g., Fischer et al. 1979; Liu 1977; Seo and Cheong 1998; Guymer 1998; Kashefipour and Falconer 2002; Shucksmith et al. 2010). Quantifying this coefficient remains a challenging task since various governing parameters cause complexity in the mixing process. Introducing mathematical expressions for the dispersion coefficient is consequently problematic. Considering that river reaches may vary in conditions, one formula may not produce accurate dispersion coefficients. However, this approach is a quite common practice in hydraulic engineering (Rowiński et al. 2005).

When a tracer is introduced to a channel, the shape of tracer cloud is largely affected by velocity variations across the channel. Taylor (1954) suggested that the transverse shear velocity and transverse mixing become in equilibrium after a certain timescale at some point downstream. Beyond this point, a Fickian diffusion equation can be used to model the tracer cloud concentration. The following simplified one-dimensional (1-D) advection-dispersion equation was derived by using Fickian's law for a uniform channel:

$$\left(\frac{\partial C}{\partial t}\right) + U\left(\frac{\partial C}{\partial x}\right) = K_x\left(\frac{\partial^2 C}{\partial x^2}\right) \quad (1)$$

where C = cross-sectional average concentration (kg/m^3); U = cross-sectional average velocity (m/s); x = direction of the mean flow; t = time in seconds (s); and K_x = longitudinal dispersion coefficient (m^2/s). Equilibrium is not guaranteed in natural streams. However, Eq. (1) adequately illustrates important features of tracer profiles in laboratory and river channels (Rutherford 1994).

Various experimental studies have explored different aspects of the longitudinal dispersion (Fukuoka and Sayre 1973; Guymer 1998; Murphy et al. 2007). Moreover, regression and dimensional based analysis and data-driven methods have been employed for predicting the dispersion coefficients, which have a wide range of variations (Seo and Cheong 1998; Kashefipour and Falconer 2002; Sahay 2011). More details are provided in the "Previous Works" section.

The main purpose of this study is to employ a M5′ algorithm (Wang and Witten 1997) to develop a transparent model to predict the longitudinal dispersion coefficient. The M5′ model tree is a new soft computing method that provides understandable formulas that allow users to have more insight in the physics of the phenomenon (Etemad-Shahidi and Bonakdar 2009). Rainfall-runoff modeling (Solomatine 2003), flood forecasting (Solomatine and Xue 2004), sediment transport (Bhattacharya and Solomatine 2005), and wave prediction (Etemad-Shahidi and Mahjoobi 2009) are

[1]Griffith School of Engineering, Gold Coast Campus, Griffith Univ., QLD 4222, Australia; and School of Civil Engineering, Iran Univ. of Science and Technology, Narmak, Tehran, Iran (corresponding author). E-mail: a.etemadshahidi@griffith.edu.au; etemad@iust.ac.ir

[2]M.Sc., School of Civil Engineering, Iran Univ. of Science and Technology, Narmak, Tehran, Iran. E-mail: miladtaghipour@civileng.iust.ac.ir

Note. This manuscript was submitted on February 16, 2011; approved on December 14, 2011; published online on May 15, 2012. Discussion period open until November 1, 2012; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydraulic Engineering***, Vol. 138, No. 6, June 1, 2012. ©ASCE, ISSN 0733-9429/2012/6-542–554/$25.00.

examples of successful model tree applications. This method has not been used for predicting the dispersion coefficient. The authors of this paper used a comprehensive field data set consisting of 149 field measurements that were extracted from the technical literature to develop the model. By using ststistical error measures, the authors then compared the performance of the developed model with the performance of previous models.

Previous Work

In rivers, a range of variables affect the longitudinal dispersion coefficient. The most important variables are density, viscosity, channel width, flow depth, mean velocity, shear velocity, bed slope, bed roughness, horizontal stream curvature (i.e., sinuosity), and bed shape factor (Seo and Cheong 1998; Guymer 1998). Previous efforts have mostly been devoted to developing a formula for estimating K_x by using easily measurable parameters such as mean velocity and depth. The authors give an overview of these investigations and then provide a brief report of other affecting parameters (e.g., sinuosity, vegetation) and soft computing methods that are used for the predicting K_x.

Elder (1959) expanded Taylor's method for an open channel of infinite width. By using laboratory measurements and assuming a logarithmic distribution for the velocity profile in the vertical direction, Elder suggested

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

where H = depth of flow; and U_* = bed shear velocity. The transverse variation in the velocity profile was not considered in deriving Eq. (2). This may lead to underestimated predictions because in most natural channels the transverse shear is more important than the vertical shear.

Fischer (1967) used the lateral velocity profile instead of the vertical velocity profile, and developed the following integral equation:

$$K_x = -\frac{1}{A}\int_0^W hu' \int_0^y \frac{1}{\varepsilon_t h}\int_0^y hu' dydydy \tag{3}$$

in which A = cross-sectional area; W = channel width; $h = h(y)$ = local flow depth; u' = deviation of the velocity from the cross-sectional mean velocity; and ε_t = transverse turbulent diffusion coefficient. This equation shows that K_x is inversely related to ε_t. In narrow and deep rivers, ε_t is high and therefore K_x is low. By contrast, in relatively wide rivers, the transverse variation of velocity is large and K_x will be greater (Rutherford 1994).

Because of difficulties in using the integral form and unavailability of detailed transverse velocity profile, Fischer (1975) simplified Eq. (3) into the following nonintegral form:

$$K_x = 0.011\left(\frac{W^2}{H}\right)\left(\frac{U^2}{U_*}\right) \tag{4}$$

Liu (1977) [Eq. (5)], Iwasa and Aya (1991) [Eq. (6)], and Koussis and Rodrigues-Mirasol (1998) [Eq. (7)] have considered the effect of the lateral velocity gradient on dispersion as well as Fischer's (1975) expression by using laboratory and field data. Their formulas were

$$\frac{K_x}{HU_*} = \beta\left(\frac{W}{H}\right)^2\left(\frac{U}{U_*}\right)^2; \qquad \beta = 0.18\left(\frac{U_*}{U}\right)^{1.5} \tag{5}$$

$$\frac{K_x}{HU_*} = 2\left(\frac{W}{H}\right)^2 \tag{6}$$

$$\frac{K_x}{HU_*} = \phi\left(\frac{W}{H}\right)^2; \qquad \phi = 0.6 \tag{7}$$

Koussis and Rodrigues-Mirasol (1998) compared their model with Fischer's (Fisher 1975) and stated that their results were much closer to the measurements.

Seo and Cheong (1998) used 59 data sets from rivers in the United States. To obtain the following equation, they implemented dimensional analysis to select appropriate variables for model construction and applied the one-step Hubor method, which is a nonlinear multi regression method:

$$\frac{K_x}{HU_*} = 5.915\left(\frac{W}{H}\right)^{0.62}\left(\frac{U}{U_*}\right)^{1.428} \tag{8}$$

They stated that Liu's equation (Liu 1977) is generally in good agreement with the measured data, whereas Iwasa and Aya's equation (Iwasa and Aya 1991) underestimates K_x in many cases.

Deng et al. (2001) developed a mathematical expression for the terms h, u', and ε_t from Eq. (3) and predicted the dispersion coefficient as

$$\frac{K_x}{HU_*} = 5.915\frac{0.15}{8\varepsilon_{t0}}\left(\frac{W}{H}\right)^{5/3}\left(\frac{U}{U_*}\right)^2 \quad \text{for } \frac{W}{H} > 10;$$
$$\varepsilon_{t0} = 0.145 + \frac{1}{3520}\left(\frac{U}{U_*}\right)\left(\frac{W}{H}\right)^{1.38} \tag{9}$$

where ε_{t0} = dimensionless transverse mixing coefficient. Their model is limited to straight-uniform streams with W/H greater than 10; however, they showed that it is superior to the model of Seo and Cheong (1998) in predicting the K_x. The model of Deng et al. (2001) has the disadvantage of the complexity caused by the approximation methods for triple numerical integration with a set of regression equations (Rowiński et al. 2005).

By using 81 sets of field data in the United States, Kashefipour and Falconer (2002) developed an equation on the basis of dimensional and regression analysis:

$$K_x = 10.612\ HU\left(\frac{U}{U_*}\right) \tag{10}$$

They also found out that the average computed ratio of (K_x/HU_*), obtained from the formula of Seo and Cheong (1998) and from their formula were 1508 and 887, respectively, whereas the corresponding average measured ratios was 1045. Therefore, they combined Eqs. (8) and (10) to obtain a more accurate model by using trial and error. Their final equation was

$$K_x = \left[7.428 + 1.775\left(\frac{W}{H}\right)^{0.62}\left(\frac{U_*}{U}\right)^{0.572}\right]HU\left(\frac{U}{U_*}\right) \tag{11}$$

According to their analysis, the models of Fischer (1975) and of Koussis and Rodrigues-Mirasol (1998) overestimate the longitudinal dispersion coefficient. Kashefipour and Falconer (2002) proposed that for open channel flows with W/H greater and less than 50, Eqs. (10) and (11), respectively, can be used for practical applications.

In a more fundamental study, Papadimitrakis and Orphanos (2004) stated that the dispersion processes depend on transverse and vertical velocity profiles, and their relative importance depends on the W/H ratio. They divided W/H values into three regions and studied each region individually. Various combinations of parameters derived from river geometry and velocity data were tested, and an empirical expression was proposed for different ranges of the W/H ratios.

J. Hydraul. Eng. 2012.138:542-554.

Seo and Baek (2004) developed a theoretical method to predict longitudinal dispersion coefficient on the basis of the distributions of transverse velocity profile in natural streams. They first tested different velocity profile equations for irregular cross sections. They then developed a new equation for the longitudinal dispersion coefficient on the basis of the velocity profile. The comparison showed that the predictions of the developed equation have better agreement with the observed values.

Sahay and Dutta (2009) applied a genetic algorithm (GA) to 65 field measurements and proposed

$$\frac{K_x}{HU_*} = 2\left(\frac{W}{H}\right)^{0.96}\left(\frac{U}{U_*}\right)^{1.25} \tag{12}$$

They mentioned that expressions given by Seo and Cheong (1998), Deng et al. (2001), and Kashefipour and Falconer (2002) perform well, especially when K_x values greater than 100 m^2/s are excluded from the analysis. They also found that the most effective parameter for accurately predicting the longitudinal dispersion coefficient is the term U/U_*.

Tayfur (2009) also used the GA approach on the basis of 85 field data and proposed the following empirical equation:

$$K_x = 0.91Q + 9.94 \tag{13}$$

in which Q = flow discharge. According to this study, Eq. (13) may have limited predictive capacity for fast-flowing mountainous streams or for streams with a very low flow discharge rate.

In addition to these studies, some investigations have focused on other influential parameters. For instance, Fukouka and Syre (1973) experimentally investigated the effect of sinuosity in a laboratory flume with various bending conditions. They found that in these cases the dispersion coefficient is larger and the initial convective period is shorter than those of an equivalent straight channel. Other researchers (Rutherford 1994; Guymer 1998, Boxall et al. 2003; Boxall and Guymer 2007; Bashitialshaaer et al. 2011) have also investigated the effect of this parameter. The effects of other factors such as vegetation, dead zones, and hydraulic structures have been studied as well. Nepf et al. (1997) found that the longitudinal dispersion coefficient was decreased in the presence of vegetation, whereas Shucksmith et al. (2010) noticed an increase in longitudinal mixing in submerged conditions. Valentine and Wood (1977) conducted numerical modeling to study two-dimensional flow with regular dead zones. They observed that dead zones increase the rate of dispersion and delay the occurrence of Fickian-type dispersion. Considerable research efforts have been devoted to the modeling of dead/storage zones in the last decade. More details in this regard can be found in Seo and Cheong (2001), Singh (2003), Smith et al. (2006), Cheong et al. (2007), and Marion et al. (2008). Caplow et al. (2004) suggested that dams (as a hydraulic structure) reduce the longitudinal dispersion coefficient below the expected value in a natural channel with the same discharge. However, quantifying the effects of such parameters requires detailed information of the river hydraulics and experimental investigations.

Soft computing methods have been also applied by several investigators to estimate K_x. Fuzzy logic (Tayfur 2006; Toprak and Savci 2007), adaptive neuro-fuzzy inference system techniques (Riahi-Madvar et al. 2009; Noori et al. 2009), support vector machine (Noori et al. 2009; Azamathulla and Ghani 2010), and genetic programming (Azamathulla and Wu 2011) are the examples of these approaches. Artificial neural network (ANN) models have been also employed to predict K_x (Rowiński et al. 2005; Tayfur and Singh 2005; Toprak and Cigizoglu 2008; Sahay 2011).

Material and Method

Data Set

This study used a collection of different data sets that were measured in different rivers (Fischer 1968; Yotsukura et al. 1970; McQuivey and Keffer 1974; Nordin and Sabol 1974; Rutherford 1994; Graf 1995). By considering the published data sets, 149 distinctive data records were selected and are presented in the appendix. The data sets contain geometric and hydraulic characteristics, including channel width, channel depth, average velocity, shear velocity, and longitudinal dispersion coefficient. Fig. 1 illustrates the histograms of K_x, W/H, and U/U_*. Approximately 80% of K_x values are less than 100 m^2/s, which is the expected maximum value of K_x in natural rivers (Chapra 1997). The histogram of W/H implies that the studied cases varied from narrow rivers ($W/H < 10$) to very wide rivers ($W/H > 100$). The friction term, defined as U/U_* (Seo and Cheong 1998), can be interpreted as the hydrodynamic characteristics of the river bed. In other words, the wide range of U/U_* in Fig. 1 covers different bed roughnesses. The reported coefficients and hydraulic characteristics such as water depth, width, and shear velocity may have some uncertainties in their values. Poor estimation procedures, tracer loss, or measurements made in the advective zone are the examples of such uncertainties in K_x values. Software and hardware errors are also inevitable in measuring the hydraulic characteristics of a river (Rutherford 1994).

Model Tree

The main concept of the model tree approach is the process of dividing complex problems into smaller problems (Bhattacharya et al. 2007). Therefore, the model tree (MT) can be regarded as a robust method for classification and prediction and is more understandable than ANN (Jung et al. 2010). In fact, the MT combines the conventional decision tree with linear regression equation at the leaves (Wang and Witten 1997). The M5 algorithm, initially introduced by Quinlan (1992), is one of the most commonly used approaches of MTs. Two main processes are considered in the algorithm: building the tree and deriving the knowledge from it. The first process involves dividing the input parameter space into a smaller subspace for which a multiple regression model is assigned. The scheme resembles an inverted tree in which the root is on top while the leaves are at the bottom. In the second process, a data record is introduced into the root of the tree. Fig. 2 illustrates splitting the space for building a tree and eliciting knowledge from the structure.

The record finds its way down by passing through the nodes. Nodes in the tree represent the testing of a particular parameter. This testing process involves comparing the given parameter with a constant value. These nodes are arranged on the basis of dividing condition of the first process (i.e., the process of building the tree). The related prediction of the introduced record is obtained when a leaf is reached, and it is recognized as an output. That record is indeed classified on the basis of the class appointed to that leaf.

The M5 algorithm was later improved as the M5′ algorithm by Wang and Witten (1997). The new version is more robust, produces simpler trees, and can deal with enumerated and missing values. The M5′ algoritm generally consists of three steps: building, pruning, and smoothing the tree. M5′ is a recursive algorithm that constructs the regression tree by using standard deviation reduction (SDR) factor to split the space:

$$\text{SDR} = sd(T) - \sum_i \frac{T_i}{|T|} \times sd(T_i) \tag{14}$$

J. Hydraul. Eng. 2012.138:542-554.

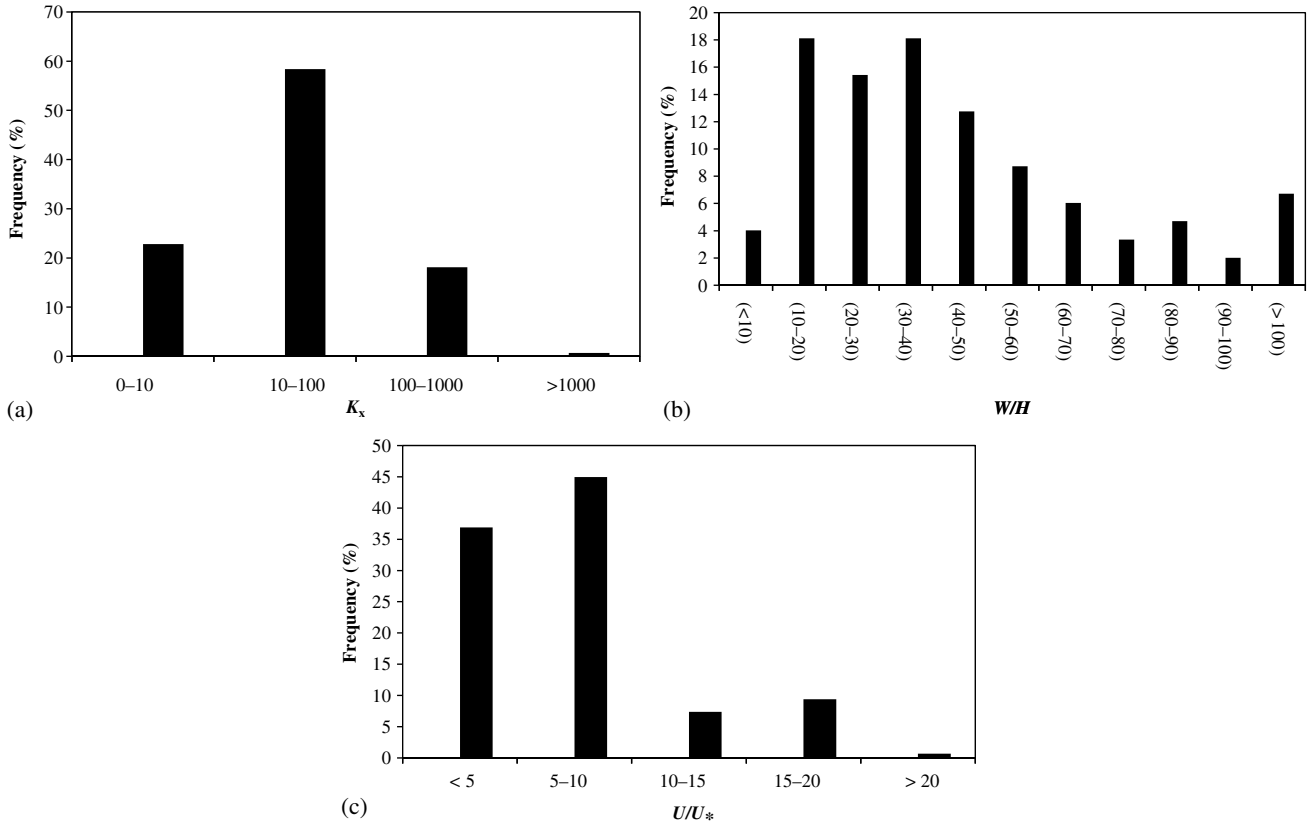

Fig. 1. Histograms of: (a) K_x; (b) W/H; (c) U/U_*

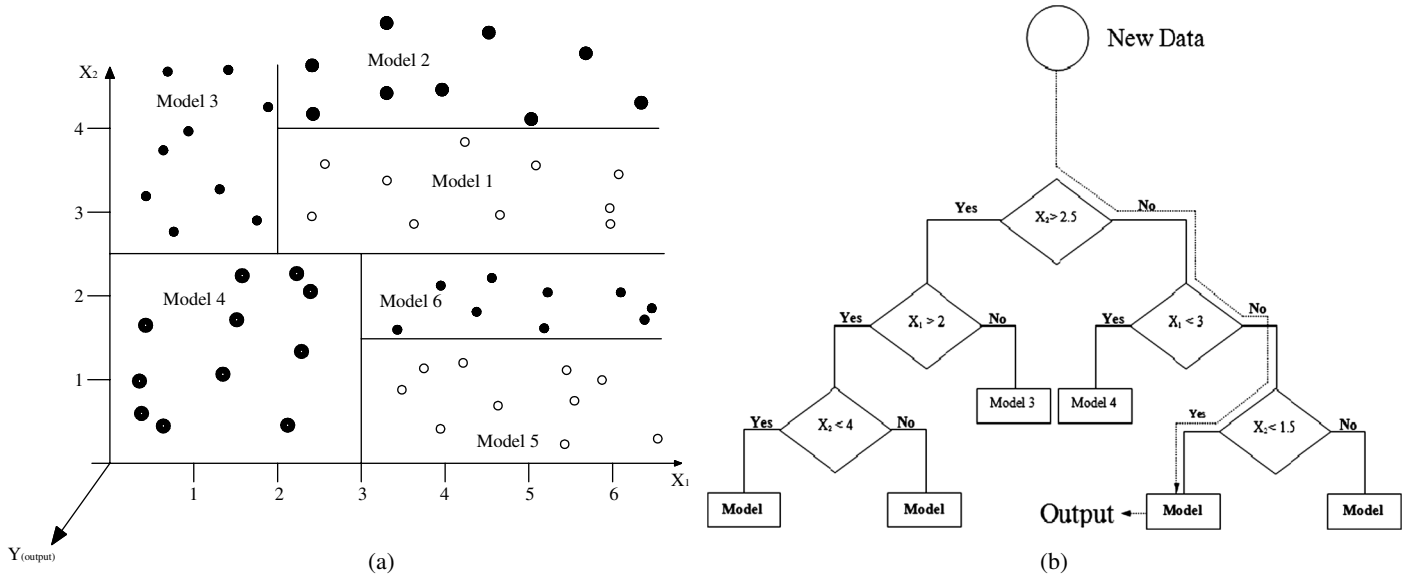

Fig. 2. Splitting the input space and prediction by the model tree for a new data record: (a) splitting of the input space ($X_1 \times X_2$) by the M5 model tree algorithm; (b) predicting a new data record by the model tree

in which T = set of the data points before splitting; T_i = data point that results from splitting the space and falls into one subspace according to the chosen splitting parameter; and sd = standard deviation (Wang and Witten 1997). Standard deviation is considered as an error measure for the data points that fall into a subspace. The M5′ model tree tests different splitting points for all input parameters. For each subspace, standard deviation is calculated and then compared with the standard deviation of data records before dividing the space into smaller ones. When a value of the input parameters maximizes the expected error reduction, it is selected as the splitting point (i.e., node). This process (i.e., splitting) is repeated for every subspace. The splitting process ends when a standard deviation reduction is less than 5% or when a few data points remain in a subdomain. After building the tree, it is used

J. Hydraul. Eng. 2012.138:542-554.

to calculate a linear multiple regression model for each subspace by using the input parameters.

As the tree grows, the accuracy of the model increases uniformly for the training set. Overfitting may consequently be inevitable while the tree is built; therefore, pruning plays an important role in this step. Pruning is the process of merging some of the lower subtrees into one node to avoid generating too accurate and overfitted trees. Predicting the expected error at each node for the test data is used in pruning. The average absolute difference between the predicted value and the actual output is calculated for each of the training sets that reach the node. To prevent underestimating the expected error for new data, the output value is multiplied by $(n+\nu)/(n-\nu)$, where n is the number of training data points that reach to the node and ν is the number of input parameters that represent the output value at that node. The leaf (i.e., subspace) can be pruned if the predicted error is less than the expected error (Witten and Frank 2005).

The last step is the regularization process to compensate sharp discontinuities. This may happen between adjacent linear models in the leaves after the tree is pruned. In this step, models built in each subspace are used to calculate the predicted value. That value is then modified along the route back to the root of the tree on top (i.e., the first splitting point) by smoothing it at each node. The predicted value established by the leaf model is combined with the value of linear model for each node (Quinlan 1992).

Modeling and Result

As discussed in the "Previous Works" section, different parameters can affect the longitudinal dispersion coefficient. With the available data in this study, the effects of some parameters such as vegetation, dead zones, and hydraulic structures cannot be investigated. However, the authors assume that the studied cases in this paper can represent average conditions that may occur in natural environments. Therefore, the following term correlates the remaining parameters that affect K_x:

$$K_x = f_1(\rho, \mu, W, H, U, U_*, S_f, \sigma, \text{ slope, roughness}) \tag{15}$$

where ρ = fluid density; μ = viscosity; S_f = bed shape factor; and σ = sinuosity. According to Seo and Cheong (1998), bed shape factor and sinuosity represent the vertical and lateral irregularities, respectively.

As mentioned previously, by using dimensional analysis, Eq. (15) can be written in a dimensionless form (Seo and Cheong 1998; Kashefipour and Falconer 2002), as follows:

$$\frac{K_x}{HU_*} = f_2\left(\rho\frac{HU}{\mu}, \frac{W}{H}, \frac{U}{U_*}, S_f, \sigma, \text{ slope, roughness}\right) \tag{16}$$

in which K_x/HU_* = dimensionless dispersion coefficient and $\rho HU/\mu$ = Reynolds number. Because the flow in natural rivers is usually turbulent, the effect of the Reynolds number is negligible and can be ignored. The effects of channel slope and roughness can be reflected by U_* and U/U_*, respectively, and can be excluded.

Because of the complexity of obtaining σ and the limited number of available data for this parameter, most previous studies have omitted it. However, some investigators have commented on the effect of σ. Sahay (2011), Tayfur and Singh (2005), and Rowiński et al. (2005) stated including σ in the input vector of the ANN models improves the accuracy of prediction. By contrast, Tayfur (2006) stated that no strong dependence exists between K_x and σ. The authors of the present study have in fact confirmed the former finding. In this study, σ was reported in approximately 40% of the whole measurements. Therefore, σ was excluded from the input parameters of the model tree at the first step to simplify the problem.

S_f values are not reported in the data sets, and therefore it was impossible to use them. This parameter is not easily collected from natural streams, and its corresponding effect can be included by U/U_* (Seo and Cheong 1998). However, Deng et al. (2001) introduced the expression $\beta = \ln(W/H)$; it is called the channel shape parameter. The parameter β may be able to reflect the vertical irregularities as the bed shape factor. As will subsequently be seen, an input of the authors' model is $\log(W/H)$, which corresponds to β. Therefore, Eq. (16) can be written as

$$\frac{K_x}{HU_*} = f_3\left(\frac{W}{H}, \frac{U}{U_*}\right) \tag{17}$$

By assuming f_3 is a power function, the general expression of the longitudinal dispersion coefficient can be

$$\frac{K_x}{HU_*} = a\left(\frac{W}{H}\right)^b\left(\frac{U}{U_*}\right)^c \tag{18}$$

in which a, b, and c = constants of the equation and possess different values in different expressions.

Because model trees ordinarily can only produce linear relationships, the model was developed log(*inputs*) and log(*output*) to obtain a nonlinear relationship. Furthermore, most of the data-driven approaches perform well while dealing with data having nearly uniform or normal distributions (Pyle 1992). It is easy to infer from Fig. 3 that the distributions of the used variables are nearly log normal.

By considering possible combinations of dimensionless forms for the longitudinal dispersion coefficient, plots of (K_x/HU_*), (K_x/HU), (K_x/WU_*), (K_x/WU) versus W/H and U/U_* were plotted and their correlation coefficients were calculated. The authors found that (K_x/HU_*) is the best dimensionless form of K_x and that it has the highest correlation with W/H and U/U_*. These three terms were consequently used as the inputs and the output for developing the model.

Taking the logarithms of Eq. (18) derives the following linear formula:

$$\log\left(\frac{K_x}{HU_*}\right) = \log a + b\log\left(\frac{W}{H}\right) + c\log\left(\frac{U}{U_*}\right) \tag{19}$$

The test-and-train technique was used to develop the model. This is a common technique in learning algorithms on a data set (Mahjoobi et al. 2008). In this method, a data set is randomly divided into two subsets (i.e., the train and the test). The train data set is used to train the model; the model is then tested (i.e., verified) by using the test data set. In this study, 119 data records were used for training and the remaining data sets were used for testing the model. The statistics of the parameters used for training the model are listed in Table 1. The developed MT generated the following formulas:

$$\text{If } \log(W/H) \le 1.486,$$
$$\text{then } \log(K_x/HU_*) = 1.90 + 0.78\log(W/H) + 0.11\log(U/U_*) \tag{20a}$$

$$\text{If } \log(W/H) > 1.486,$$
$$\text{then } \log(K_x/HU_*) = 1.15 + 0.61\log(W/H) + 0.85\log(U/U_*) \tag{20b}$$

After transformation, Eqs. (20*a*) and (20*b*) can be written

J. Hydraul. Eng. 2012.138:542-554.

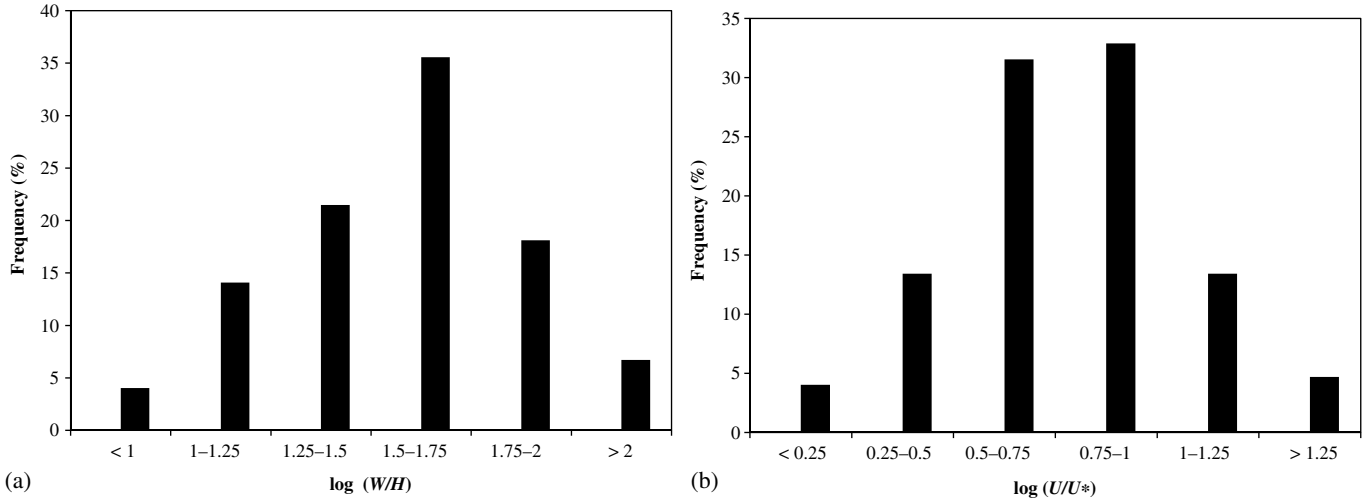

Fig. 3. Histograms of: (a) $\log(W/H)$; (b) $\log(U/U_*)$

Table 1. Statistics of parameters used for training the model

Statistical indicator	W (m)	H (m)	U (ms^{-1})	U_* (ms^{-1})	W/H	U/U_*	K_x (m^2 s^{-1})	K_x/U_*H
Maximum	253.6	8.2	1.73	0.55	403.75	20.25	1,486.5	37,140
Minimum	1.4	0.14	0.03	0.002	2.20	0.77	0.2	3.08
Average	48.6	1.36	0.48	0.087	47.72	6.96	79.4	1,172
Standard deviation	47.2	1.39	0.33	0.078	49.64	4.75	174.9	3,570

$$\text{If } W/H \le 30.6, \quad \text{then } \left(\frac{K_x}{HU_*}\right) = 15.49\left(\frac{W}{H}\right)^{0.78}\left(\frac{U}{U_*}\right)^{0.11} \tag{21a}$$

$$\text{If } W/H > 30.6, \quad \text{then } \left(\frac{K_x}{HU_*}\right) = 14.12\left(\frac{W}{H}\right)^{0.61}\left(\frac{U}{U_*}\right)^{0.85} \tag{21b}$$

The splitting parameter is W/H and the splitting value is approximately 30. This is close to the value that Papadimitrakis and Orphanos (2004) obtained. This splitting value is obtained by minimizing the prediction error and does not necessarily have a physical interpretation (Bhattacharya et al. 2007; Bonakdar and Etemad-Shahidi 2011). However, the importance of W/H in determining K_x has been mentioned by others (Asay and Fujisaki 1991; Kashefipour and Falconer 2002; Papadimitrakis and Orphanos 2004; Tayfur and Singh 2005). Transverse shear is less important with relatively small values of W/H, whereas it dominates the dispersion characteristic when the aspect ratio is large. Therefore, different regimes may exist for low and high W/H ratios.

The exponents of W/H and U/U_* are different in these formulas. In rivers for which $W/H \le 30.6$, it is the width-to-depth ratio that outweighs the dispersion coefficient. In wider rivers for which $W/H > 30.6$, the influence of U/U_* increases and the effect of W/H decreases (see also Papadimitrakis and Orphanos 2004). A possible interpretation is K_x may be less influenced by the W/H ratio in very wide reivers than in narrow rivers. As Rutherford (1994) discusses, the role of velocity is more pronounced in determining the K_x in relatively wide rivers than in narrow rivers. In this regard, the power of U in Eq. (21*b*) is nearly eight times greater than its value in Eq. (21*a*). The obtained exponents of W/H and U/U_* are in the range reported in previous works. The average exponents of W/H and U/U_* are 0.7 and 0.48, which are approximatley the values obtianed by Seo and Cheong (1998) and Liu (1977), respectively. In brief, the authors concluded that the obtained formulas are in good agreement with engineering sense and previous findings.

The performance of the developed model was evaluated against the performance of other existing models by using error measures such as the discrepancy ratio (DR) (White et al. 1973); the mean of the absolute error (ME); and the root mean square (RMS). These parameters are defined as

$$\text{DR} = \log\frac{K_{x_p}}{K_{x_m}} \tag{22}$$

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N}|\text{DR}_i| \tag{23}$$

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{DR}_i)^2} \tag{24}$$

in which K_{x_p} and K_{x_m} = predicted and measured dispersion coefficients, respectively; and N = total number of data points.

If DR is equal to zero, there will be an exact match between the measured and predicted values. An overestimation (DR > 0) or underestimation (DR < 0) otherwise occurs. Accuracy is defined as the percentage of DR values that fall between −0.3 and 0.3 (Seo and Cheong 1998; Kashefipour and Falconer 2002). Comparing the calculated values of ME and RMS with zero can also determine the performance of each model. The closer the values are to zero, the more accurate the model will be.

Table 2 presents error measures of previous models and the developed model. The results in the last two rows of Table 2 show that

J. Hydraul. Eng. 2012.138:542-554.

Table 2. Comparison of the Performance of Various Models

Model	DR < -0.3	$-0.3 <$ DR < 0	$0 <$ DR < 0.3	DR > 0.3	Accuracy	ME	RMS
Elder (1959), all data	98.0	1.3	0.7	0.0	2.0	1.85	1.95
Fischer (1975), all data	30.2	18.1	16.8	34.9	34.9	0.56	0.71
Liu (1977), all data	17.4	22.1	28.9	31.6	51.0	0.42	0.57
Seo and Cheong (1998), all data	18.8	16.1	30.2	34.9	46.3	0.43	0.59
Deng et al. (2001), all data	20.1	19.5	27.5	32.9	47.0	0.42	0.56
Kashefipour and Falconer (2002) I, all data, Eq. (10)	36.9	30.2	10.7	22.2	40.9	0.54	0.74
Kashefipour and Falconer (2002) II, all data, Eq. (11)	26.1	29.5	19.4	25	48.9	0.46	0.66
Sahay and Dutta (2009), all data	20.1	22.8	22.8	34.3	45.6	0.40	0.53
MT, all data	17.4	28.9	34.2	19.5	63.1	0.32	0.44
MT, testing data	13.3	36.7	26.7	23.3	63.3	0.35	0.48

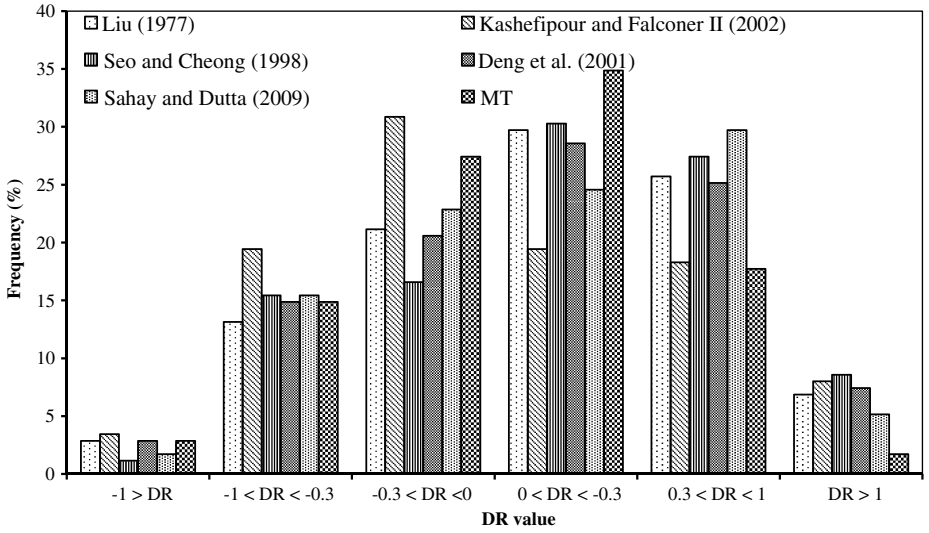

Fig. 4. Comparison of the DR values of different models

the errors of the developed model for testing data and all data are approximately the same. Elder's equation (Elder 1959) is more suitable for the rivers with no transverse shear; however, the comparison of this equation with others can merely illustrate the importance of transverse variation. In Table 2, the performance of Elder's model is least satisfactory followed by Fischer's model (Fischer 1975). All the error measures of the developed model show improved prediction of the longitudinal dispersion coefficient. The MT has an accuracy of 63%, the highest among the models. The nearest value of the accuracy to that of MT is that of Liu (1977) with approximately 51%. The difference between these two accuracy values well shows superiority of MT over other models. The ME and RMS are other performance indicators. The developed model outperforms other models because it has the lowest values for these two error measures. In addtion, the percentage of DR values greater than 0.3 and less than -0.3 of MT are 17.4% and 19.5%, respectively. This indicates that DR values outside of this range are almost equally distributed between overestimated and underestimated values. For other models, nonsymmetric distributions for values out of the range of -0.3 to 0.3 are somewhat considerable. Liu (1977), Seo and Cheong (1998), Deng et al. (2001), and Sahay and Dutta (2009) overpredict the dispersion coefficient by 1.7 times more than the underpredicted cases. In other words, the models generally overestimate the measured values of the longitudinal dispersion coefficient. Overestimating of the longitudinal dispersion coefficient obtains a lower maximum concentration. This is an important issue, especially in practical applications regarding the estimation of the maximum concentration. In such cases, it may be unsafe to use overpredicted K_x values.

Fig. 4 shows a comparison of the histograms of DR values for six models. The DR distribution of the MT shows a nearly symmetrical distribution between -1 and 1, which indicates that relatively no skewness exists toward positive or negative values. However, other models overestimate. For example, the model of Sahay and Dutta (2009) is skewed to positive values and does not have a symmetrical distribution. This can be understood from Table 3, which gives the mean, standard deviation, and skewness of

Table 3. Mean ($\overline{\text{DR}}$), Standard Deviation (σ_{DR}), Maximum DR ($|\overline{\text{DR}_{\max}}|$), and Skewness of DR ($SK_{\text{DR}}$) of different models

Model	$\overline{\text{DR}}$	σ_{DR}	$\lvert\overline{\text{DR}_{\max}}\rvert$	SK_{DR}
Liu (1977)	0.11	0.56	2.05	-0.52
Seo and Cheong (1998)	0.18	0.56	2.15	0.8
Deng et al. (2001)	0.11	0.56	1.89	0.33
Kashefipour and Falconer (2002) II	0.03	0.66	2.74	1.1
Sahay and Dutta (2009)	0.11	0.52	1.79	0.38
MT	0.00	0.43	1.54	-0.32

DR of different models. The skewness values of the model of the authors and of Deng et al. (2002) are the lowest. Moreover, the mean DR of MT is zero, which indicates a symmetrical distribution. Approximately 68% of the DR values of the model by Sahay and Dutta (2009) fall between −0.41 and 0.63, with 95% of the values ranging from −0.93 to 1.15; this implies skewness toward positive values. Another example addresses the model of Kashefipour and Falconer (2002) II. The corresponding mean value of DR is close to zero, although it has a relatively high standard deviation. Approximately 95% the DR values are between −0.94 and 1.15, whereas the DR values of the MT are in the range of −0.86 to 0.86. This implies that the DR values of MT prediction are close to zero. Table 3 also shows that the MT has the lowest maximum error, whereas the model of Kashefipour and Falconer (2002) II has the largest error.

The correlation coefficient (CC) and the slope of the regression line are also other tools for evaluating the performance of a model. If the slope of the regression line for prediction versus the measured data is close to 1 and the value of CC is high, then the model is accurate. The developed model greatly outperforms other models in predicting K_x when extreme measured values of the dispersion coefficient ($K_x > 100$ m^2/s) are excluded from the analysis. As Table 4 presents, the slope of the regression line of MT is close to 1 and it has the highest CC.

Introducing model tree equations gained some more information. Information such as the splitting point and its value and the values of the exponent of the input parameters helped the authors include σ parameter for the cases in which it was reported. By using Eqs. (21*a*) and (21*b*) and the splitting point, the data set was divided into two subsets. For each subset, the effect of σ was considered a power function (i.e., $\lambda = A\sigma^B$). Nonlinear regression relation obtained the constants by using the reported data (including σ). Eqs. (21*a*) and (21*b*) were finally modified as

Table 4. Slope of the Regression Line and CC for the Predicted versus the Measured K_x ($K_x < 100$ m^2/s) of different models

Model	Slope of regression line	CC
Liu (1977)	3.86	0.29
Seo and Cheong (1998)	1.31	0.46
Deng et al. (2001)	0.61	0.14
Kashefipour and Falconer (2002) II	0.51	0.18
Sahay and Dutta (2009)	1.4	0.49
MT	0.96	0.6

Table 5. Comparison of Error Measures of MT Equations With and Without σ

Error measure	$W/H < 30.6$		$W/H > 30.6$		All data points with reported σ value	
	Without	With	Without	With	Without	With
MAE	0.53	0.33	0.25	0.2	−0.07	0
RMSE	0.71	0.42	0.31	0.27	0.44	0.32
$\overline{DR}$	−0.23	0	0	0	0.32	0.24
σ_{DR}	0.7	0.44	0.32	0.27	0.44	0.31

$$\text{If } W/H \le 30.6, \quad \text{then } \left(\frac{K_x}{HU_*}\right) = 2.75\left(\frac{W}{H}\right)^{0.78}\left(\frac{U}{U_*}\right)^{0.11}(\sigma)^{4.04} \tag{25a}$$

$$\text{If } W/H > 30.6, \quad \text{then } \left(\frac{K_x}{HU_*}\right) = 8.36\left(\frac{W}{H}\right)^{0.61}\left(\frac{U}{U_*}\right)^{0.85}(\sigma)^{1.70} \tag{25b}$$

These equations show that K_x is directly related to the sinuosity. This is in line with the previous findings of Fukouka and Syre (1973) and Bashitialshaaer et al. (2011). The performance of the modified equations interestingly improved when including sinuosity. Table 5 summarizes the error measures of Eqs. (21) and (25) for both ranges and for all data with reported σ values. The accuracy of new equations accounting for σ was enhanced, especially for the lower range of W/H. The power of σ depends on the W/H ratios. As Eq. (25) indicates, sinuosity has a greater effect on the lower W/H ratios. This is in good agreement with engineering sense because mixing is more influenced by the river curvatures in narrow rivers.

Applying piecewise regression may provide better understanding of the physics of the phenomenon in comparison with a simple equation that may not be appropriate for all cases. However, a comparison of two equations with existing equations is inevitable for illustrating the performance of the new model. The model tree approach used in this study requires minimum effort in comparison with other soft computing methods. The model tree provides simple regression formulas with low computational cost (Jafari and Etemad-Shahidi 2012). In contrast to other soft computing methods such as ANN, the MT does not require much trial and error to obtain the best model. It is more transparent and can provide understandable formulas. The latter advantage can benefit users by giving greater insight into the physics of the phenomenon and by quantifying the role of each input parameter. Other soft computing such as ANN have limited applicability because they are more like a black box model and do not reveal any direct mathematical expressions (Tayfur 2006). Model trees also have some limitations. As mentioned previously, they can only produce linear relationships. In addition, the transformation of input parameters may not be that simple in more complex cases and may not necessarily lead to a few simple linear formulas.

Conclusion

In this study, a M5′ model tree was used to predict the longitudinal dispersion coefficient in natural streams. The model was developed by using 149 field data records consisting of hydraulic and geometrical characteristics. Because of the limited number of reported values of σ, the authors of this paper decided to develop the equations without σ in the first step. On the basis of previous studies and trial and error, W/H and U/U_*, and (K_x/HU_*) were used as the model tree's inputs and output, respectively. Two formulas were generated and the splitting parameter was W/H, which is an important parameter in dispersion mechanism. The performance of the new model was evaluated. By using different error measures, the results were compared with the results of existing formulas. The developed model outperformed other models in accuracy. Effect of σ was then considered and the results showed improved prediction of the dispersion coefficient. The suggested models seem to be safely applicable in hydraulic and environmental studies such as design of outfalls or in evaluating risks from spills of hazardous contaminants.

J. Hydraul. Eng. 2012.138:542-554.

Appendix. Data Sets Used in This Study

No	Stream	W (m)	H(m)	U (m/s)	U_*(m/s)	K_x(m^2/s)	σ
1	Copper Creek, VA (below gauge)	15.9	0.49	0.21	0.079	19.52	
2	Copper Creek, VA (below gauge)	18.3	0.84	0.52	0.1	21.4	
3	Copper Creek, VA (below gauge)	16.2	0.49	0.25	0.079	9.5	
4	Clinch River, TN (below gauge)	46.9	0.86	0.28	0.067	13.93	
5	Clinch River, TN (below gauge)	59.4	2.13	0.86	0.104	53.88	
6	Clinch River, TN (below gauge)	53.3	2.09	0.79	0.107	46.45	
7	Copper Creek, VA (above gauge)	18.6	0.39	0.14	0.116	9.85	
8	Power River, TN	33.8	0.85	0.16	0.055	9.5	
9	Clinch River, VA	36	0.58	0.3	0.049	8.08	
10	Green and Duwamish	21.77	1.58	0.31	0.058395	6.5	
11	Green and Duwamish	29.61	1.08	0.36	0.048279	0.5	1.41
12	Bayou Anacoco	19.8	0.41	0.29	0.044	13.94	1.30
13	Nooksack River	86	2.94	1.2	0.514	153.29	
14	Antietam Creek	15.8	0.39	0.32	0.06	9.29	
15	Antietam Creek	19.8	0.52	0.43	0.069	16.26	
16	Antietam Creek	24.4	0.71	0.52	0.081	25.55	
17	Monocacy River	35.1	0.32	0.21	0.04	4.65	
18	Monocacy River	36.6	0.45	0.32	0.05	13.94	
19	Monocacy River	47.5	0.87	0.44	0.07	37.16	
20	Missouri River	182.9	2.23	0.93	0.065	464.52	1.35
21	Missouri River	201.2	3.56	1.27	0.082	836.13	1.35
22	Missouri River	196.6	3.11	1.53	0.077	891.87	1.35
23	Wind/Bighom Rivers	67.1	0.98	0.88	0.11	41.81	
24	Elkhom River	32.6	0.3	0.43	0.046	9.29	
25	Elkhom River	50.9	0.42	0.46	0.046	20.9	
26	John day River	25	0.56	1.01	0.137	13.94	1.08
27	Comite River	12.5	0.26	0.31	0.043	6.97	1.31
28	Comite River	15.8	0.41	0.37	0.055	13.94	1.31
29	Amite River	36.6	0.81	0.29	0.068	23.23	
30	Amite River	42.4	0.8	0.42	0.068	30.19	
31	Sabine River	103.6	2.04	0.56	0.054	315.87	
32	Sabine River	127.4	4.75	0.64	0.081	668.9	
33	Muddy Creek	13.4	0.81	0.37	0.077	13.94	
34	Muddy Creek	19.5	1.2	0.45	0.093	32.52	
35	Sabine River, Texas	35.1	0.98	0.21	0.041	39.48	
36	White River	67.1	0.55	0.35	0.044	30.19	
37	Chattahoochee River	65.5	1.13	0.39	0.075	32.52	
38	Susquehanna River	202.7	1.35	0.39	0.065	92.9	1.13
39	Antietam Creek	10.97	0.52	0.21	0.074909	17.5	
40	Antietam Creek	23.47	0.7	0.52	0.101491	101.5	
41	Antietam Creek	24.99	0.45	0.41	0.081374	25.9	
42	Antietam Creek	12.8	0.3	0.42	0.057	17.5	1.40
43	Antietam Creek	24.08	0.98	0.59	0.098	101.5	2.25
44	Antietam Creek	11.89	0.66	0.43	0.085	20.9	2.25
45	Antietam Creek	21.03	0.48	0.52	0.069	25.9	1.26
46	Monocacy River	48.7	0.55	0.26	0.05	37.8	1.28
47	Monocacy River	92.96	0.71	0.16	0.05	41.4	1.28
48	Monocacy River	51.21	0.65	0.62	0.04	29.6	1.28
49	Monocacy River	97.54	1.15	0.32	0.058	119.8	1.61
50	Monocacy River	49.99	0.95	0.32	0.074778	29.6	
51	Monocacy River	33.53	0.58	0.16	0.041315	66.5	
52	Monocacy River	40.54	0.41	0.23	0.04	66.5	1.61
53	Conococheague Creek	42.21	0.69	0.23	0.064	40.8	2.25
54	Conococheague Creek	49.68	0.41	0.15	0.081	29.3	2.25
55	Conococheague Creek	42.98	1.13	0.63	0.081	53.3	1.31

J. Hydraul. Eng. 2012.138:542-554.

No	Stream	W (m)	H(m)	U (m/s)	U_*(m/s)	K_x(m^2/s)	σ
56	Conococheague Creek	43.28	0.69	0.22	0.063729	40.8	
57	Conococheague Creek	63.7	0.46	0.1	0.056203	29.3	
58	Conococheague Creek	59.44	0.76	0.68	0.072242	53.3	
59	Chattahoochee River	75.6	1.95	0.74	0.138	88.9	1.27
60	Chattahoochee River	91.9	2.44	0.52	0.094	166.9	1.57
61	Chattahoochee River	99.97	2.5	0.3	0.105054	166.9	
62	Salt Creek	32	0.5	0.24	0.038	52.2	1.38
63	Difficult Run	14.5	0.31	0.25	0.062	1.9	1.09
64	Difficult Run	11.58	0.4	0.22	0.087475	1.9	
65	Bear Creek	13.7	0.85	1.29	0.553	2.9	1.08
66	Little Pincy Creek	15.9	0.2	0.39	0.053	7.1	1.13
67	Bayou Anacoco	17.5	0.45	0.32	0.024	5.8	1.41
68	Bayou Anacoco	25.9	0.94	0.34	0.067	27.6	1.41
69	Bayou Anacoco	36.6	0.91	0.4	0.067	40.2	1.41
70	Comite River	15.7	0.2	0.36	0.04	69	1.31
71	Comite River	6.1	0.49	0.25	0.057591	69	
72	Bayou Bartholomew	33.4	1.4	0.2	0.03	54.7	2.46
73	Bayou Bartholomew	37.49	2.07	0.1	0.040306	54.7	
74	Amite River	21.3	0.5	0.54	0.027	501.4	
75	Amite River	46.02	0.53	0.41	0.042659	501.4	
76	Tickfau River	14.9	0.59	0.27	0.08	10.3	1.75
77	Tickfau River	41.45	1.04	0.07	0.090343	10.3	
78	Tangipahoa River	31.4	0.81	0.48	0.072	45.1	1.46
79	Tangipahoa River	29.9	0.4	0.34	0.02	44	1.46
80	Tangipahoa River	42.98	1.28	0.26	0.068162	45.1	
81	Tangipahoa River	31.7	0.76	0.36	0.053227	44	
82	Red River	253.6	0.81	0.48	0.072	45.1	1.20
83	Red River	161.5	0.4	0.34	0.02	44	1.44
84	Red River	152.4	1.62	0.61	0.032	143.8	1.44
85	Red River	155.1	3.96	0.29	0.06	130.5	1.24
86	Red River	248.11	4.82	0.31	0.065235	143.8	
87	Sabine River, LA	116.4	3.66	0.45	0.057	227.6	1.17
88	Sabine River, LA	160.3	1.74	0.47	0.036	177.7	1.17
89	Sabine River, TX	14.2	1.65	0.58	0.054	131.3	2.53
90	Sabine River, TX	12.2	2.32	1.06	0.054	308.9	2.05
91	Sabine River, TX	21.3	0.5	0.13	0.037	12.8	1.47
92	Sabine River, TX	21.64	0.61	0.08	0.04237	12.8	
93	Sabine River, TX	17.37	1.23	0.04	0.050338	14.7	
94	Sabine River, TX	31.39	1.43	0.13	0.041029	24.2	
95	Wind/Bighom Rivers	44.2	1.4	0.99	0.14	184.6	1.56
96	Wind/Bighom Rivers	85.3	2.4	1.73	0.15	464.6	1.56
97	Copper Creek	16.7	0.5	0.2	0.08	16.8	2.54
98	Clinch River	48.5	1.2	0.21	0.07	14.8	1.25
99	Copper Creek	18.3	0.4	0.15	0.12	20.7	2.54
100	Powell River	36.8	0.9	0.13	0.05	15.5	
101	Clinch River	28.7	0.6	0.35	0.07	10.7	2.20
102	Copper Creek	19.6	0.8	0.49	0.1	20.8	1.14
103	Clinch River	57.9	2.5	0.75	0.1	40.5	
104	Conchelaa Canal	24.7	1.6	0.66	0.04	5.9	1.14
105	Clinch River	33.53	0.78	0.19	0.049483	10.7	
106	Clinch River	55.78	2.26	0.69	0.098768	36.93	
107	Clinch River	53.2	2.4	0.66	0.11	36.9	
108	Coachell Canal, CA	23.77	1.6	0.67	0.04	5.96	1.14
109	Coachell Canal, CA	24.99	1.54	0.66	0.037	5.92	
110	Copper Creek	16.8	0.5	0.24	0.08	24.6	
111	Missouri River	180.6	3.3	1.62	0.08	1,486.5	
112	Bayou Anacoco	25.9	0.9	0.34	0.07	32.5	

J. Hydraul. Eng. 2012.138:542-554.

No	Stream	W (m)	H(m)	U (m/s)	U_*(m/s)	K_x(m^2/s)	σ
113	Bayou Anacoco	36.6	0.9	0.4	0.07	39.5	
114	Nooksack River	64	0.8	0.67	0.27	34.8	1.30
115	Wind/Bighom Rivers	59.4	1.1	0.88	0.12	41.8	1.18
116	Wind/Bighom Rivers	68.6	2.2	1.55	0.17	162.6	1.18
117	John Day River	34.1	2.5	0.82	0.18	65	1.89
118	Yadkin River	70.1	2.4	0.43	0.1	111.5	2.17
119	Yadkin River	71.6	3.8	0.76	0.13	260.1	2.17
120	Colorado River	106.1	6.1	0.79	0.088201	181	
121	Colorado River	71.6	8.2	1.2	0.336784	243	
122	Albert	100	4.4	0.029	0.0016	0.2	
123	Dessel-Herentals	35	2.5	0.037	0.0022	0.2	
124	Yuma Mesa A	7.6	3.45	0.68	0.047	0.5	
125	Bocholt-Dessel	35	2.5	0.107	0.0063	1.4	
126	Villemsvaart	34	2.5	0.13	0.0079	1.7	
127	Chicago Ship Canal	49	8.07	0.27	0.019	3	
128	Irrigation	1.4	0.19	0.38	0.11	9.6	
129	Irrigation	1.5	0.14	0.33	0.1	1.9	
130	Puneha	5	0.28	0.26	0.21	7.2	
131	Kapuni	9	0.3	0.37	0.15	8.4	
132	Kapuni	10	0.35	0.53	0.17	12.4	
133	Manganui	20	0.4	0.19	0.18	6.5	
134	Waiongana	13	0.6	0.48	0.24	6.8	
135	Stony	10	0.63	0.55	0.3	13.5	
136	Waiotapu	11.4	0.75	0.41	0.061	8	
137	Manawatu	59	0.72	0.37	0.07	32	
138	Manawatu	63	1	0.32	0.094	22	
139	Manawatu	60	0.95	0.46	0.092	47	
140	Tarawera	25	1.21	0.73	0.084	27	
141	Tarawera	20	1.92	0.62	0.123	11.5	
142	Tarawera	25	1.38	0.77	0.091	20.5	
143	Tarawera	25	1.4	0.78	0.091	15.5	
144	Tarawera	25	1.57	0.83	0.096	18	
145	Tarawera	85	2.6	0.69	0.06	52	
146	Waikato	120	2	0.64	0.05	67	
147	Miljacka	11	0.29	0.35	0.058	2.7	
148	Upper Tame	9.9	0.83	0.46	0.09	5.5	
149	Upper Tame	9.9	0.92	0.52	0.1	5.1	

Acknowledgments

The authors acknowledge the comments of professor Jorg Imberger on the previous version of the manuscript. The authors also thank Meysam Bali, Ebrahim Jafari, Ali Behnood and Amin Asgarian for their help in improving the manuscript.

Notation

The following symbols are used in this paper:

A = cross-sectional area;
C = cross-sectional average concentration;
H = depth of flow;
h = local flow depth;
K_x = longitudinal dispersion coefficient;
K_{x_m} = predicted dispersion coefficient;
K_{x_p} = measured dispersion coefficient;
N = number of data points;
n = number of train data points;
Q = flow discharge;
S_f = bed shape factor;
sd = standard deviation;
T = set of the examples that reach the node;
T_i = set of the results of the node splitting in accordance with the selected parameter;
t = time;
U = cross-sectional average velocity;
U_* = bed shear velocity;
u' = deviation of the velocity from the cross-sectional mean velocity;
x = direction of the mean flow;
β = channel shape parameter;
ε_t = transverse turbulent diffusion coefficient;
ε_{t0} = dimensionless transverse mixing coefficient;
μ = fluid viscosity;
ρ = fluid density;
σ = sinuosity; and
ν = number of inputs.

References

Asai, K., and Fujisaki, K. (1991). "Effect of aspect ratio on longitudinal dispersion coefficient." *Proc., Int. Symp. on Envir. Hydr.*, A. A. Balkema, Rotterdam, Netherlands, 493–498.

Azamathulla, H. M., and Ghani, A. A. (2011). "Genetic programming for predicting longitudinal dispersion coefficients in streams." *Water Resour. Manage.*, 25(6), 1537–1544.

Azamathulla, H. M., and Wu, F.-C. (2011). "Support vector machine approach for longitudinal dispersion coefficients in natural streams." *Appl. Soft Comput.*, 11(2), 2902–2905.

Bashitialshaaer, R., Bengtsson, L., Larson, M., Persson, K. M., Aljaradin, M., and Hossam, A.-I. (2011). "Sinuosity effects on longitudinal dispersion coefficient." *Int. J. Sust. Water Environ. Sys.*, 2(2), 77–84.

Bhattacharya, B., Price, R. K., and Solomatine, D. P. (2007). "Machine learning approach to modeling sediment transport." *J. Hydraul. Eng.*, 133(4), 440–450.

Bhattacharya, B., and Solomatine, D. P. (2005). "Neural networks and M5 model trees in modelling water level-discharge relationship." *Neurocomp.*, 63, 381–396.

Bonakdar, L., and Etemad-Shahidi, A. (2011). "Predicting wave run-up on rubble-mound structures using M5′ machine learning method." *Ocean Eng.*, 38(1), 111–118.

Boxall, J. B., and Guymer, I. (2007). "Longitudinal mixing in meandering channels: New experimental data set and verification of a predictive technique." *Water Res.*, 41(2), 341–354.

Boxall, J. B., Guymer, I., and Marion, A. (2003). "Transverse mixing in sinuous natural open channel flows." *J. Hydraul Res.*, 41(2), 153–165.

Caplow, T., Schlosser, P., and Ho, D. T. (2004). "Tracer study of mixing and transport in the upper Hudson River with multiple dams." *J. Environ. Eng.*, 130(12), 1498–1506.

Chapra, S. C. (1977). *Surface water-quality modeling*, McGraw-Hill, New York.

Cheong, T. S., Younis, B. A., and Seo, I. W. (2007). "Estimation of key parameters in model for solute transport in rivers and streams." *Water Resour. Manage.*, 21(7), 1165–1186.

Deng, Z. Q., Bengtsson, L., Singh, V. P., and Adrian, D. D. (2002). "Longitudinal dispersion coefficient in single-channel streams." *J. Hydraul. Eng.*, 128(10), 901–916.

Deng, Z. Q., Singh, V. P., and Bengtsson, L. (2001). "Longitudinal dispersion coefficient in straight rivers." *J. Hydraul. Eng.*, 127(11), 919–927.

Elder, J. W. (1959). "The dispersion of a marked fluid in turbulent shear flow." *J. Fluid Mech.*, 5(4), 544–560.

Etemad-Shahidi, A., and Bonakdar, L. (2009). "Design of rubble-mound breakwaters using M5′ machine learning method." *Appl. Ocean Res.*, 31(3), 197–201.

Etemad-shahidi, A., and Mahjoobi, J. (2009). "Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior." *Ocean Eng.*, 36(15–16), 1175–1181.

Fischer, H. B. (1967). "The mechanics of dispersion in natural streams." *J. Hydraul. Div.*, 93(HY6), 187–216.

Fischer, H. B. (1968). "Dispersion predictions in natural streams." *J. Sanit. Eng. Div.*, 94(5), 927–943.

Fischer, H. B. (1975). "Discussion of 'simple method for predicting dispersion in streams.' by R. S. McQuivey and T. N. Keefer." *J. Environ. Eng. Div.*, 101(3), 453–455.

Fischer, H. B., List, E. J., Koh, R. C. Y., Imberger, J., and Brooks, N. H. (1979). *Mixing in land and costal waters*, Academic, New York, 104–138.

Fukuoka, S., and Sayre, W. W. (1973). "Longitudinal dispersion in sinuous channels." *J. Hydraul. Div.*, 99(1), 195–217.

Graf, B. (1995). "Observed and predicted velocity and longitudinal dispersion at steady and unsteady flow, Colorado River, Glen Canyon Dam to Lake Mead." *J. Am. Water Resour. Assoc.*, 31(2), 265–281.

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

Iwasa, Y., and Aya, S. (1991). "Predicting longitudinal dispersion coefficient in open-channel flows." *Proc., Int. Symp on Envir. Hydr.*, Hong Kong University Press, Hong Kong, 505–510.

Jafari, E., and Etemad-Shahidi, A. (2012). "Derivation of a new model for prediction of wave overtopping at rubble-mound structures." *J. Waterw., Port, Coastal Eng. Div.*, 138(1), 42–52.

Jung, N.-C., Popescu, I., Kelderman, P., Solomatine, D. P., and Price, R. K. (2010). "Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea." *J. Hydroinf.*, 12(3), 262–274.

Kashefipour, S. M., and Falconer, R. A. (2002). "Longitudinal dispersion coefficients in natural channels." *Water Res.*, 36(6), 1596–1608.

Koussis, A. D., and Rodríguez-Mirasol, J. (1998). "Hydraulic estimation of dispersion coefficient for streams." *J. Hydraul. Eng.*, 124(3), 317–320.

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

Mahjoobi, J., Etemad-Shahidi, A., and Kazeminezhad, M. H. (2008). "Hindcasting of wave parameters using different soft computing methods." *Appl. Ocean Res.*, 30(1), 28–36.

Marion, A., Zaramella, M., and Bottacin-Busolin, A. (2008). "Solute transport in rivers with multiple storage zones: The STIR model." *Water Resour. Res.*, 44(10), W10406.

McQuivey, R. S., and Keffer, T. N. (1974). "Simple method for predicting dispersion in streams." *J. Environ. Eng. Div.*, 100(4), 997–1011.

Murphy, E., Ghisalberti, M., and Nepf, H. (2007). "Model and laboratory study of dispersion in flows with submerged vegetation." *Water Resour. Res.*, 43(5), W05438.

Nepf, H. M., Mugnier, C. G., and Zavistoski, R. A. (1997). "The effects of vegetation on longitudinal dispersion." *Estuary Coast Shelf Sci.*, 44(6), 675–684.

Noori, R., Karbassi, A. R., Farokhnia, A., and Dehghani, M. (2009). "Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques." *Environ. Eng. Sci.*, 26(10), 1503–1510.

Nordin, C. F., and Sabol, G. V. (1974). "Empirical data on longitudinal dispersion in rivers." *Water-Resources Investigations 20-74*, U.S. Geological Survey, Reston, VA.

Papadimitrakis, I., and Orphanos, I. (2004). "Longitudinal dispersion characteristics of rivers and natural streams in Greece." *Water, Air, Soil Pol.: Focus.*, 4(4–5), 289–305.

Pyle, D. (1992). *Data preparation for data mining*, Morgan Kaufmann, San Francisco.

Quinlan, J. R. (1992). "Learning with continuous classes." *Proc., 5th Australian Joint Conf. on Artificial Intelligence*, World Scientific, Singapore, 343–348.

Riahi-Madvar, H., Ayyoubzadeh, S. A., Khadangi, E., and Ebadzadeh, M. M. (2009). "An expert system for predicting longitudinal dispersion coefficient in natural streams by using ANFIS." *Expert Syst. Appl.*, 36(4), 8589–8596.

Rowiński, P. M., Piotrowski, A., and Napiórkowski, J. J. (2005). "Are artificial neural network techniques relevant for the estimation of longitudinal dispersion coefficient in rivers?" *Hydrol. Sci. J.*, 50(1), 175–187.

Rutherford, J. C. (1994). *River mixing*, Wiley, Chichester, UK.

Sahay, R. R. (2011). "Prediction of longitudinal dispersion coefficients in natural rivers using artificial neural network." *J. Fluid Mech.*, 11(3), 247–261.

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

Seo, I. W., and Baek, K. O. (2004). "Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams." *J. Hydraul. Eng.*, 130(3), 227–236.

Seo, I. W., and Cheong, T. S. (1998). "Predicting longitudinal dispersion coefficient in natural streams." *J. Hydraul. Eng.*, 124(1), 25–32.

Seo, I. W., and Cheong, T. S. (2001). "Moment-based calculation of parameters for the storage zone model for river dispersion." *J. Hydraul. Eng.*, 127(6), 453–465.

Shucksmith, J. D., Boxall, J. B., and Guymer, I. (2010). "Effects of emergent and submerged natural vegetation on longitudinal mixing in open channel flow." *Water Resour. Res.*, 46(4), W04504.

Singh, S. K. (2003). "Treatment of stagnant zones in riverine advection-dispersion." *J. Hydraul. Eng.*, 129(6), 470–473.

J. Hydraul. Eng. 2012.138:542-554.

Solomatine, D. P., Dulal K. N. (2003). "Model trees as an alternative to neural networks in rainfall-runoff modelling." *Hydrol. Sci. J.*, 48(3), 399–411.

Solomatine, D. P., and Xue, Y. P. (2004). "M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China." *J. Hydrol. Eng.*, 9(6), 491–501.

Smith, P., Beven, K., Tawn, J., Blazkova, S., and Merta, L. (2006). "Discharge-dependent pollutant dispersion in rivers: Estimation of aggregated dead zone parameters with surrogate data." *Water Resour. Res.*, 42(4), W04412.

Tayfur, G. (2006). "Fuzzy, ANN, and regression models to predict longitudinal dispersion coefficient in natural streams." *Nord. Hydrol.*, 37(2), 143–164.

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

Tayfur, G., and Singh, V. P. (2005). "Predicting longitudinal dispersion coefficient in natural streams by artificial neural network." *J. Hydraul. Eng.*, 131(11), 991–1000.

Taylor, G. I. (1954). "The dispersion of matter in turbulent flow through a pipe." *Proc. R. Soc. Lond. A*, The Royal Society, London, 223(1155), 446–468.

Toprak, Z. F., and Cigizoglu, H. K. (2008). "Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods." *Hydrol. Processes*, 22(20), 4106–4129.

Toprak, Z. F., and Savci, M. E. (2007). "Longitudinal dispersion coefficient modeling in natural channels using fuzzy logic." *Clean-Soil Air Water*, 35(6), 626–637.

Valentine, E. M., and Wood, I. R. (1977). "Longitudinal dispersion with dead zones." *J. Hydraul. Div.*, 103(9), 975–990.

Wang, Y., and Witten, I. H. (1997). "Induction of model trees for predicting continuous classes." *Proc. of the Poster Papers of the European Conf. on Machine Learning, 1997*, Univ. of Economics, Faculty of Informatics and Statistics, Prague, Czech Republic.

White, W. R., Milli, H., and Crabbe, A. D. (1973). "Sediment transport an appraisal methods, Vol 2: Performance of theoretical methods when applied to flume and field data." *Hydraulic Research Station Rep., No. 1T119*, Wallingford, UK.

Witten, I. H., and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco.

Yotsukura, N., Fischer, H. B., and Sayre, W. W. (1970). "Measurement of mixing characteristics of the Missouri River between Sioux City, Iowa and Plattsmouth, Nebraska." *U.S. Geological Survey Water-Supply Paper 1899-G*, U.S. Government Printing Office, Washington, DC.

J. Hydraul. Eng. 2012.138:542-554.