Chapter 11

Derivation of Sediment Transport Models for Sand Bed Rivers from Data-Driven Techniques

Vasileios Kitsikoudis,
Epaminondas Sidiropoulos and Vlassios Hrissanthou

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/53432

1. Introduction

Hydraulic engineers and geologists have studied sediment transport in natural streams and rivers for centuries due to its importance in understanding river hydraulics. Erosion and deposition of sediment alters the hydraulic geometry of the channel and may cause increase of flood frequency as well as navigation problems from excessive deposition. Moreover, discharge of industrial and agricultural residuals sets the sediment particles to be the primary transporters of toxic substances that contaminate aquatic systems. High sediment discharge peaks may be destructive for fish habitats and ecosystems, and long-term sediment yield affects the design and function of constructions such as dams and reservoirs, as well as the coastal erosion at the basin outlet.

Sediment transport in sand bed rivers and natural streams is a complex process. For its quantification, numerous sediment transport functions have been introduced in the past years based on different concepts. There are four basic approaches used in the derivation of sediment transport formulae (Yang, 1977): 1) The deterministic approach, which obeys the laws of physics and usually is based on an independent variable like slope, shear stress, stream power, unit stream power etc. 2) The regression approach, which has emerged from the thought that sediment transport is such a complex phenomenon that cannot be described by a single dominant variable. 3) The pioneering probabilistic approach of Einstein (1942), which highlighted the complexity and the stochastic nature of the sediment transport in a rather laborious way for common usage in engineering, and 4) The regime approach, which was developed as a result of long-term measurements in equilibrium conditions.

The emerging results from all these concepts usually differ drastically from each other and from the measured data. Consequently, none of the published sediment transport equations has gained universal acceptance in confidently predicting sediment transport rates, especially in rivers. An alternative approach may be the usage of data-driven modeling, which is especially attractive for modeling processes, in which knowledge of the physics of the problem is inadequate. The scope of this chapter is the utilization of some widely used data-driven techniques, namely artificial neural networks (ANNs) and symbolic regression based on genetic programming (GP) in order to determine the dominant dimensionless variables that can be used as inputs in such schemes and generate sediment transport models for natural streams and rivers that are based solely on the data without presuming anything about their structure and their degree of nonlinearity.

For the proper training of a data-driven scheme, data of good quality are needed. Since field measurements accommodate the peculiarities of the considered streams and the inclusion of noise in the measurement process is inevitable, the training data comprise solely laboratory flume measurements. The testing data, however, comprise exclusively field measurements in order to implement the models in actual applications. Based on this concept, the approach of the basic trend of the function is feasible and the derived model will be applicable to the data range for which it will be trained. Regarding the efficiency of scaling in the sediment transport context, model-prototype comparisons have shown that correspondence of behavior is often well beyond expectations, as has been attested by the successful operation of many structures designed from model tests (Pugh, 2008). This study exhibits the potential of machine learning in capturing functions with physical meaning since the training and testing sets have significant differences in their statistical distributions. The determination of the input variables that best define the problem is accomplished by the assessment of some common independent dimensionless variables based on their correlation with the sediment concentration and the aid of ANNs on the basis of a tentative trial-and-error procedure. Subsequently, ANNs and symbolic regression are utilized in order to derive equations from the selected input combinations.

2. Data mining and data-driven techniques in the context of sediment transport

The recorded observations of a system can be further analyzed in the search for the information they encode. Such automated search for models accurately describing data constitutes a direction that can be identified as that of data mining. Data mining and knowledge discovery aim at providing tools to facilitate the conversion of data into a number of forms, such as equations. The latter provide a better understanding of the process generating or producing these data. These models combined with the already available understanding of the physical processes result in an improved understanding and novel formulations of physical laws and improved predictive capability (Babovic, 2000).

Data-driven modeling (DDM) and machine learning techniques used for predictions are essentially modernized regression schemes with the significant advantage over the classical regression schemes that they do not have to presume the structure of the nonlinear model, which they attempt to fit. They are based on simple ideas, usually inspired from the way nature works, and their only prerequisite is a good, although usually large, data set. The data are usually divided into three sets, namely the training, validation and testing set. The training set trains the scheme on the basis of a minimization criterion and the validation set is used as a stopping criterion for training to avoid overfitting to the data used for training. The test set is used to evaluate the generated model. The minimization criterion, on the basis of which the training process takes place, is usually a sum of errors between the computed outputs and the actual measured data. The optimization model that is used for the minimization depends on the data-driven scheme and may be deterministic as well as stochastic.

Inferring models from data is an activity of deducing a closed-form explanation based solely on observations. These observations, however, represent a limited source of information. The question emerges as to how this, a limited flow of information from a physical system to the observer, can result in the formation of a model that is complete in the sense that it can account for the entire range of phenomena encountered within the physical system in question and describe even the data outside the range of previously encountered observations. The present efforts are characterized by the search for a model that is capable of acquiring semantics from syntax. Clearly, every model has its own syntax. Artificial neural networks have the syntax of a network of interconnected neurons, whereas genetic programming has the syntax of treelike networks of symbolic expressions in reverse Polish notation. The question is whether such a syntax can capture the semantics of the system it attempts to model (Babovic, 2000). Witten et al. (2011) argued that the universal learner is an idealistic fantasy since experience has shown that no single machine learning scheme is appropriate to all data mining problems. Certain classes of model syntax may be inappropriate as a representation of a physical system. One may choose the model whose representation is complete, in the sense that a sufficiently large model can capture the data's properties to a degree of error that decreases with an increase in the model size. For example, one may decide to expand Taylor or Fourier series and decrease the error by adding terms in a series. However, in these cases, semantics almost certainly would not be caught (Babovic, 2000).

2.1. Artificial neural networks

ANN is the most widely used data-driven method. Since abundant information on ANNs is available in the literature [e.g. Haykin (2009)], only a brief description of ANNs is provided, with regard only to the methodology applied herein. ANN is a broad term covering a large variety of network architectures and structures. The most common of them, and the one utilized herein, is the multilayer feedforward network. This type of network is a parallel distributed information processing system that consists of the input layer, the hidden layer(s), and the output layer, and the information goes only in a forward direction. Each layer comprises a number of neurons, each one of which is connected with those in the successive layer with synaptic weights that determine the strength of the connections. The hidden and output layer neurons have an inherent activation function, which accommodates the nonlinear transformation of the input data to the targets. In this study, the neurons of the hidden layer(s) will have the hyperbolic tangent activation function, which squashes the data between (-1, 1), and the single neuron of the output layer will have the linear activation function, which simply returns the value that is passed to it. The input data are scaled to the range (-0.9, 0.9) because, if the values are scaled to the extreme limits of the transfer function, the size of the weight updates is extremely small and flat-spots in training are likely to occur (Maier and Dandy, 2000).

The training process of an ANN may be viewed as a "curve fitting" problem and the network itself may be considered simply as a nonlinear input-output mapping (Haykin, 2009). Supposing that a deterministic relation between sediment load concentration and some specific independent variables exists, a multilayer feedforward ANN is able to approximate this function, if it includes at least one hidden layer with a sufficient number of neurons (Hornik et al., 1989). However, this universal approximation theorem does not specify if a single hidden layer is optimal in the sense of learning time, ease of implementation, or (more importantly) generalization (Haykin, 2009). As a result, several network architectures are tested in order to determine the optimal one.

Although the implementation of ANNs is extensive and successful in water resources applications [e.g. Maier and Dandy (2000)] and in the prediction of daily suspended sediment data [e.g Cigizoglu (2004)], it is quite sparse in the prediction of sediment concentration from other independent hydraulic variables. Nagy et al. (2002) reviewed some widely used sediment discharge equations and selected some of the dominant dimensionless variables of the problem as input neurons for an ANN that was trained and tested with field data. Bhattacharya et al. (2005) used dimensionless parameters obtained from the Engelund and Hansen (1967) formula in order to train and test an ANN with a mixture of flume and field data, whilst in similar studies Bhattacharya et al. (2004, 2007) scrutinized further the possible input parameters based on the same data. Yang et al. (2009) chose as input variables combinations of dimensional quantities and applied them to field data. All of these works used the back-propagation algorithm (Rumelhart et al., 1986) for training the ANNs and compared the results with some of the most popular sediment transport formulae. For all the cases, the ANNs generated superior results.

2.2. Symbolic regression based on genetic programming

Many seemingly different problems in artificial intelligence, symbolic processing and machine learning can be viewed as requiring discovery of a computer program that produces some desired outputs for particular inputs. The process of solving these problems can be reformulated as a search for a highly fit individual computer program in the space of possible ones. GP extends the concept of genetic algorithms and provides a way to search for this fittest individual computer program (Koza, 1992).

GP works by randomly generating a population of computer programs (represented by tree structures) and each individual program in the population is measured in terms of how well it performs in the particular problem environment. This measure is called the fitness measure (Koza, 1992) and usually is a sum of errors between the outputs predicted by the program and the actual ones. Initially, the generated computer programs will have exceedingly poor fitness. Nonetheless, some individuals in the population will turn out to be somewhat fitter than others. These differences in performance are subsequently exploited. The Darwinian principle of reproduction and survival of the fittest and the genetic operations of sexual recombination (crossover) and mutation are used to create a new offspring population of individual programs from the current population. The reproduction principle involves the selection, in proportion to fitness, of a computer program from the current population that survives from the generation by being copied into the new population. The genetic process of sexual recombination is used to create new offspring programs from two parental programs selected in proportion to fitness. The parental programs are typically of different sizes and shapes. The offspring programs are composed of subexpressions from their parents and are, typically, of different sizes and shapes as well. Intuitively, if two programs are somewhat effective in solving a problem, then some of their parts probably have some merit. By recombining randomly chosen parts of somewhat effective programs, the result may be the production of new programs that are even fitter in solving the problem (Koza, 1992). Mutation serves the potentially important role of restoring lost diversity in a population by replacing random subtrees of variable length with other random ones. Its purpose is to prevent premature convergence to unsatisfactory solutions. After the operations of reproduction, crossover and mutation are performed on the current population, the offspring population replaces the old one. Each individual in the new population of programs is then measured for fitness and the process is iterated for a predetermined number of generations. This algorithm will produce populations of programs, which over many generations tend to exhibit increasing average fitness in dealing with their environment. The individual computer program that performs best in the evolved generations is considered to be the fittest.

A multigene individual consists of multiple genes, each of which is a GP evolved tree. In multigene symbolic regression, each prediction $\hat{y}$ of the output variable y is formed linearly by the weighted output of each of the genes plus a bias term (Searson, 2009). Each tree is a function of the input variables. Mathematically, a multigene regression model can be written as:

$$\hat{y} = d_0 + d_1 \times tree1 + \ldots + d_M \times treeM \tag{1}$$

where d_0=bias (offset) term; d_1, …, d_M are the gene weights and M is the number of genes comprising the current individual. The gene weights are automatically determined by a least squares procedure for each multigene individual. The number and structure of the trees is evolved automatically during a run (subject to user defined constraints) using the training data. Hence, multigene symbolic regression combines the power of classical linear regression with the ability to capture nonlinear behavior without needing to pre-specify the structure of the nonlinear model. During a run, genes are acquired and deleted using a tree crossover operator called two-point high level crossover. This allows the exchange of genes between individuals and it is used in addition to the “standard” GP recombination operators (Searson et al., 2010).

GP has been implemented in hydraulic engineering in the last years with very good results. Babovic and Abbott (1997) applied GP to some representative problems, while Babovic and Keijzer (2000) highlighted the usage of GP as a data mining tool in which the human expert interprets models suggested by the computer, aiming at knowledge discovery. Minns (2000) suggests that the symbolic expressions obtained from GP may be less accurate than the ANN in mapping the experimental data. However, these expressions may be more easily examined in order to provide insights into the processes that created the data. In the context of sediment transport, Zakaria et al. (2010) applied gene-expression programming, which is similar to multigene symbolic regression, to predict the total bed material load for rivers using dimensional quantities from field data, and outperformed some of the traditional sediment load formulae. Azamathulla et al. (2010) utilized GP in order to predict the scour depth at bridge piers and obtained results superior to those of ANNs and regression equations.

3. Sediment transport

Sediment load is the material being transported, and it can be divided into wash load and bed material load. The wash load is the fine material of sizes, which are not found in appreciable quantities on the bed, and is not considered to be dependent on the local hydraulics of the flow, but instead is dependent on the upstream supply. As a practical definition, the wash load is considered to be the fraction of the sediment load finer than 0.062 mm. The bed material load is the material of sizes, which are found in appreciable quantities on the bed and it can be conceptually divided into the bed load (the portion of the load that moves near the bed) and the suspended load (the portion of the load that moves in suspension), although the division is not precise. The consequent difficulty, however, to separate bed load from turbulence dominated suspended load leads to a total load definition for the quantification of sediment transport in sand bed rivers. A dimensionless, commonly used measure for sediment quantification is concentration by weight in parts per million (ppm), which is the ratio of the sediment discharge to the discharge of the water-sediment mixture, both expressed in terms of mass per unit time, here called C_t. This can be given as

$$C_t = 10^6 \frac{\rho_s Q_{st}}{\rho Q + \rho_s Q_{st}} \tag{2}$$

For practical reasons, the density of the water-sediment mixture is taken to be approximately equivalent to the density of water. This approximation will cause errors of less than one percent for concentrations less than 16000 ppm (Brownlie, 1981a).

The parameters governing a sediment transport process can be described by (Yalin, 1977)

$$q_t = f\left(V, D, d, S, g, \rho_s, \rho, \nu\right) \tag{3}$$

Since the data-driven schemes are trained and validated with flume data but tested with field data and in order to ensure dimensional consistency in the derived models, the input and output variables should be dimensionless. Instead of applying dimensional analysis and Buckingham's π theorem, the independent variables of Eq. (3) will be introduced by some common and well-known dimensionless variables that have physical meaning and have been utilized for the creation of various sediment transport formulae. These variables are directly related to quantities the engineer can readily visualize and measure; they are listed as follows and summarized in Table 1.

Froude number, which gives a measure of the ratio of inertial forces to gravitational forces of the flow. For the flume data, the depth will be the hydraulic radius of the bed which is equivalent to the mean depth of an infinitely wide channel with the same slope, velocity and bed friction as the flume, and is calculated according to the sidewall correction of Vanoni and Brooks (1957). This elaboration is due to the fact that in flume experiments the sand covered bed will generally be much rougher than the flume walls, and thus will be subjected to higher shear stresses.

$$Fr = \frac{V}{\sqrt{gD}} \tag{4}$$

Reynolds number, which gives a measure of the ratio of inertial forces to viscous forces of the flow

$$\mathrm{Re} = \frac{VD}{\nu} \tag{5}$$

Shear Reynolds number, the physical meaning of which, is the ratio of particle size to the thickness of the viscous sublayer δ, because δ is proportional to v/U_*.

$$\mathrm{Re}^* = \frac{U_* d_{50}}{\nu} \tag{6}$$

Dimensionless shear stress or Shields number

$$\tau^* = \frac{\tau}{(\gamma_s - \gamma) d_{50}} \tag{7}$$

Dimensionless grain diameter. It is a dimensionless expression for grain diameter that can be derived by eliminating shear stress from the two Shields parameters (Shields, 1936); or from the drag coefficient and Reynolds number of a settling particle, by eliminating the settling velocity; or dimensionally, with immersed weight of an individual grain, fluid density, and viscosity as the variables (Ackers and White, 1973). The dimensionless grain diameter is, therefore, generally applicable to coarse, transitional, and fine sediments and is the cube root of the ratio of immersed weight to viscous forces. Thus

$$d_{gr} = d_{50}\left[\frac{g\left(\gamma_s/\gamma - 1\right)}{\nu^2}\right]^{1/3} \tag{8}$$

Dimensionless stream power. The power equation appears first to have been applied to sediment transport by Rubey (1933) and later by Velikanov (1955). It was again suggested by Knapp (1938), and was later introduced by Bagnold (1956) in a paper wherein the flowing fluid was regarded as a transporting machine. The available power supply, or time rate of energy supply, to unit length of a stream is the time rate of liberation in kinetic form of the liquid's potential energy as it descends the gravity slope *S*. Denoting this power by *Ω*, Bagnold (1966) derived the formula

$$\Omega = \rho g Q S \tag{9}$$

The mean available power supply to the column of fluid over unit bed area, to be denoted by *ω*, is therefore

$$\omega = \frac{\Omega}{W} = \frac{\rho g Q S}{W} = \rho g D S V = \tau V \tag{10}$$

In order to define a dimensionless transport parameter that encapsulates Bagnold's view of sediment transport as a stream power related phenomenon, Eaton and Church (2011) developed the following formula

$$\omega^* = \frac{\omega}{\rho\left[g\left(\gamma_s/\gamma - 1\right)d\right]^{3/2}} \tag{11}$$

Dimensionless unit stream power. Yang (1972) reviewed the basic assumptions used in the derivation of conventional sediment transport equations. He concluded that the assumption that sediment transport rate could be determined from water discharge, average flow velocity, energy slope, or shear stress is questionable. Consequently, the generality and applicability of any equation derived from one of these assumptions is also questionable. The rate of energy per unit weight of water, available for transporting water and sediment in an open channel with reach length *x* and total drop of *Y*, is

$$\frac{dY}{dt} = \frac{dx}{dt}\frac{dY}{dx} = VS \tag{12}$$

Yang (1972) defines the unit stream power as the velocity-slope product and argues that the rate of work being done by a unit weight of water in transporting sediment must be directly related to the rate of work available to a unit weight of water. Thus, total sediment concentration or total bed material load must be directly related to unit stream power. While Bagnold (1966) emphasized the power that applies to a unit bed area, Yang (1972, 1973) emphasized the power available per unit weight of fluid to transport sediments. The fact that sediment discharge or concentration is dominated by the unit stream power has been confirmed by Vanoni (1978) as well. While Yang divided unit stream power *VS* by fall velocity ω_s to obtain a dimensionless variable, Vanoni (1978) divided the product *VS* by $(gd_{50})^{1/2}$. Both d_{50} and ω_s are commonly used for describing the size of sediment particles. However, d_{50} can only reflect the physical size of sediment particles, while ω_s can also reflect the interaction between sediment particles and water, which is affected by particle shape, water viscosity and temperature. On the other hand, the computation of fall velocity is problematic and a common source of errors. The emerging variables expressing dimensionless unit stream power according to Yang and Vanoni are, respectively, the following

$$\frac{VS}{\omega_s} \tag{13}$$

and

$$\frac{VS}{\sqrt{gd_{50}}} \tag{14}$$

No	Dimensionless variables
1	Froude number, *Fr*
2	Reynolds number, *Re*
3	Shear Reynolds number, Re^*
4	Dimensionless shear stress, τ^*
5	Dimensionless grain diameter, d_{gr}
6	Dimensionless stream power, ω^*
7	Yang's dimensionless unit stream power, VS/ω_s
8	Vanoni's dimensionless unit stream power, $VS/(gd_{50})^{1/2}$

Table 1. Dimensionless variables assessed for the determination of the dominant ones

Yang (1977, 2003) argued that total sediment discharge correlates best with unit stream power based on the plots of Figure 1. Nonetheless, equations based on the other hydraulic variables have been used successfully as well.

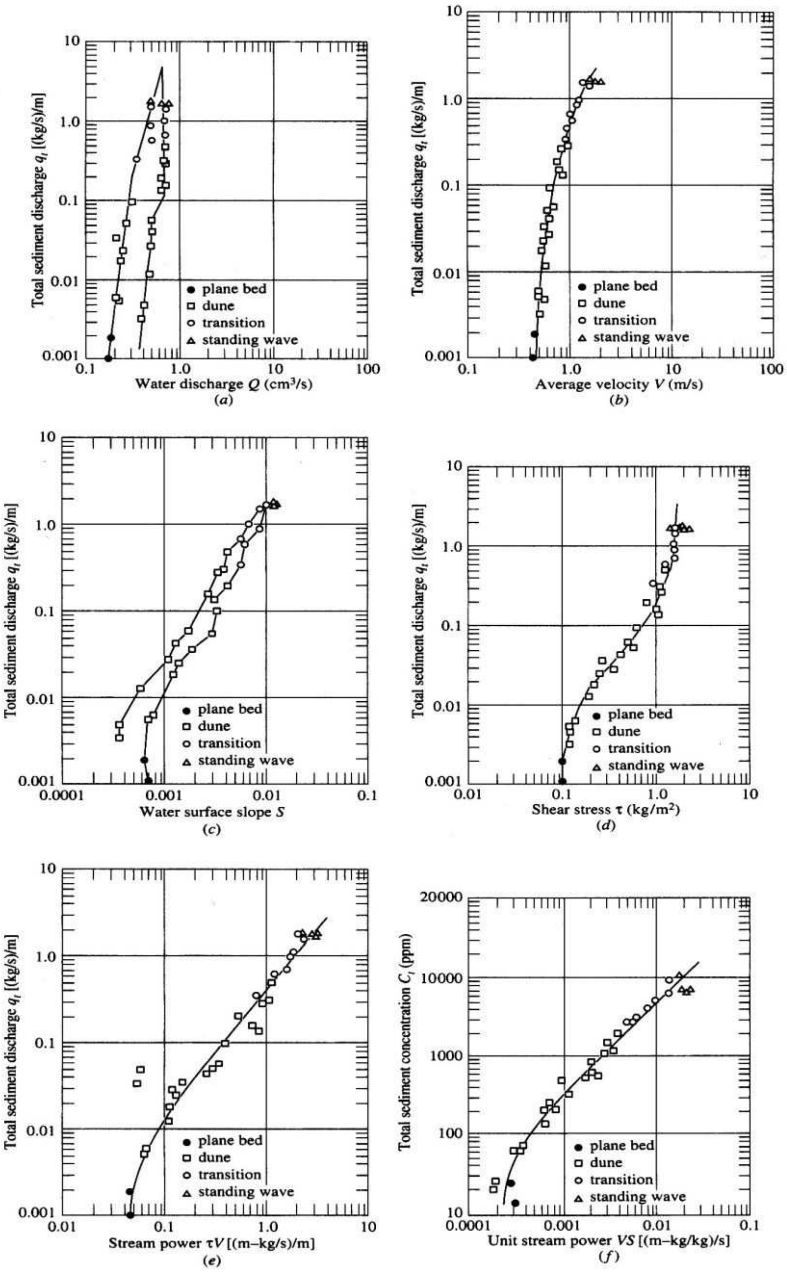

Figure 1. Relationships between total sediment discharge and (a) water discharge, (b) velocity, (c) slope, (d) shear stress, (e) stream power, and (f) unit stream power, for 0.93 mm sand in an 8 ft wide flume [obtained from Yang (2003)]

4. Data preparation and determination of the inputs

Since data-driven techniques require a large number of quality data that represent a wide spectrum of the considered problem in order to be trained efficiently, the database assembled by Brownlie (1981b) is utilized. Brownlie's (1981b) database contains 7027 records (5263 laboratory records and 1764 field records) in 77 data files. These data were subjected to a screening process similar to the one Brownlie (1981a) used for the derivation of his formula. Firstly, the measurements that were not verified by Brownlie, were incorrect or incomplete, were removed. Secondly, because only flows with sand beds were considered, median particle sizes were limited to values between 0.062 mm and 2.0 mm. To avoid samples with large amounts of gravel or fine, cohesive material, geometric standard deviations were restricted to values smaller than 5, and some other constraints were imposed in order to reduce sidewall effects, eliminate shallow water effects, and overcome accuracy problems associated with low sediment concentration. In addition to these, only flume measurements with uniform flows were considered and supercritical flows were removed due to the subcritical flows that usually prevail in nature, in sand bed rivers. Finally, the measurements with specific gravity outside the quartz density range were neglected as well as measurements that had extreme temperature values. Wherever the temperature was missing, a value of 15 °C was used for the calculation of kinematic viscosity. For the laboratory data, the sidewall correction of Vanoni and Brooks (1957) was utilized to adjust the hydraulic radius to eliminate the effects of the flume walls. If sediment concentration is correlated with velocity, however, the sidewall correction will be of little use. These restrictions are shown in Table 2.

Restriction	Reason
0.062 mm $\leq d_{50} \leq$ 2.0 mm	Sand only
$\sigma_g \leq 5$	Eliminate bimodal distributions
$W/D > 4$	Reduce sidewall effects (only for laboratory data)
$R/d_{50} > 100$	Eliminate shallow water effects
$C > 10$ ppm	Accuracy problems associated with low concentration
$Fr < 1$	Subcritical flows
2.57 ≤ Specific gravity ≤ 2.68	Natural sediments

Table 2. Restrictions imposed on data

Since measurements in natural streams and rivers are notoriously difficult, and sometimes inaccurate, and the inclusion of field data to the training set would result in a model applicable only to rivers similar to those the data were obtained from, field data are excluded from the training set. Consequently, the training set consists solely of laboratory flume data so that the noise embedded in the training set is minimized. The testing set, however, comprises exclusively field data in order to test the derived mathematical models in actual problems that occur in nature. With this technique, the generated models will have general applicability to the data range for which they are trained. The final database consists of 984 laboratory records and 600 field records that lie within the range of the laboratory records that constitute the training set, due to the data sensitive nature of DDM.

Further pruning of the outliers in the training dataset and the subsequent increase of data homogeneity would be beneficial for the training procedure, however, this would be at the expense of the amount of training data, which are already significantly reduced from the screening process. Since most DDM methods perform well when the data has a distribution that is close to normal (Bhattacharya et al., 2005), a log-transformation of the input and output variables of all datasets was applied so that the distributions of the transformed variables were closer to normal. Figure 2 depicts the distribution of the flume sediment concentrations for the original and the log-transformed values.

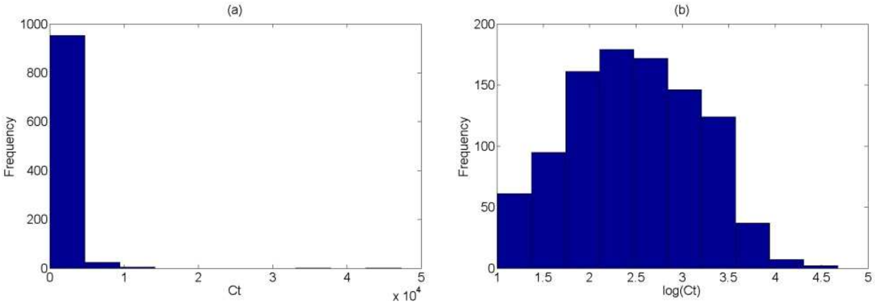

Figure 2. Distribution of flume sediment concentration in ppm (a) before log-transformation and (b) after log-transformation

For the creation of training and validation sets the available 984 laboratory measurements were placed in descending order with respect to sediment concentration and for every three successive measurements that were picked for the training set, the fourth one was selected for the validation set. This procedure was iterated for all the laboratory data and the emerged training and validation sets comprise 739 and 245 measurements, respectively. The 600 field measurements constitute the test set. Table 3 shows some statistical measures of the potential variables of these sets. Table 4 shows the datasets from which the data used in this study were obtained and some representative values of each set. The abbreviations used in Table 4 are the same with those Brownlie (1981b) used in his data compilation; consequently, all the references to the original datasets may be obtained from that study.

		Statistical measures	d_{gr}	Fr	Re^*	VS/ω_s	$VS/(gd_{50})^{0.5}$	ω^*	τ^*	C_t
Laboratory data	train set	Minimum value	2.405	0.166	2.468	0.0010	0.0014	0.093	0.0323	10.2
		Maximum value	38.800	0.999	158.292	0.5350	0.2818	134.6	5.828	47300
		Mean value	10.069	0.490	18.356	0.0297	0.0183	5.065	0.450	977.9
		Standard deviation	6.337	0.188	17.476	0.0425	0.0198	9.263	0.412	2575.8
		Skewness coefficient	1.456	0.943	3.219	5.127	5.633	6.598	4.618	11.758
	validation set	Minimum value	2.508	0.202	2.938	0.0017	0.0016	0.227	0.0578	11.3
		Maximum value	38.505	0.968	146.416	0.2984	0.1150	81.56	2.555	10630
		Mean value	9.708	0.482	17.220	0.0292	0.0173	5.397	0.459	861.4
		Standard deviation	6.760	0.173	17.467	0.0337	0.0153	9.128	0.374	1412.2
		Skewness coefficient	1.590	0.912	3.579	3.200	2.429	4.212	1.671	3.372
Field data	test set	Minimum value	2.531	0.166	6.165	0.0017	0.0015	0.365	0.094	11
		Maximum value	33.931	0.992	131.127	0.1528	0.0857	118.7	5.805	11400
		Mean value	10.654	0.375	27.910	0.0202	0.0156	14.13	0.943	1239.3
		Standard deviation	6.968	0.150	18.994	0.0194	0.0121	16.73	0.648	1472.2
		Skewness coefficient	1.705	1.027	2.105	2.676	1.893	2.350	1.726	2.554

Table 3. Statistical measures of the train, validation and test sets

Range of field variables

Code	No.	Velocity (m/s)		Depth (m)		Slope (‰)		C_t (ppm)	
		min	max	min	max	min	max	min	max
BAL	25	0.226	1.093	0.091	0.256	0.44	2.1	19	3776
BEN	1	0.205	0.205	0.038	0.038	0.5	0.5	10.2	10.2
BRO	6	0.372	0.616	0.047	0.060	2.4	3.5	1200	5300
CHY	7	0.423	0.586	0.066	0.101	1.11	2	99.4	345
COS	12	0.403	0.503	0.140	0.156	0.45	1.01	10.954	102.08
DAV	69	0.244	0.792	0.076	0.305	0.248	2.67	11.3	1760
EPA	16	0.440	0.706	0.088	0.300	0.6	3.68	32	1017
EPB	19	0.265	0.762	0.148	0.304	0.262	1.6	45	1810
FOL	6	0.388	0.599	0.036	0.047	3.74	4.02	845	1848
FRA	11	0.361	0.450	0.129	0.161	0.938	1.693	39.979	166.34
GKA	27	0.302	0.635	0.032	0.124	1.8	6.401	205	3160
GUY	145	0.225	1.321	0.058	0.405	0.37	9.5	12	47300
JOR	7	0.401	0.557	0.070	0.105	1.12	1.67	95.8	306.7
KEN	6	0.412	0.799	0.047	0.109	1.7	4.2	550	2070
KNB	9	0.277	0.674	0.070	0.168	0.56	2.5	14	1740
LAU	10	0.326	0.671	0.076	0.221	0.8	2.1	550	4240
MCD	11	0.480	0.660	0.082	0.146	1.11	1.67	151.2	615.8
MPR	15	0.426	0.835	0.112	0.490	0.42	4.066	14.357	1091.1
MUT	17	0.131	0.505	0.029	0.102	0.5	7.5	11	10630
NOR	27	0.524	1.802	0.256	0.585	0.47	5.77	33	8870
OBR	45	0.214	0.953	0.088	0.165	0.57	3.23	17	1332.5
OJK	14	0.338	0.586	0.075	0.135	1.09	2.67	66.791	3355.7
PRA	25	0.254	0.701	0.076	0.305	0.282	2.87	11.63	560
SAT	1	0.332	0.332	0.193	0.193	0.44	0.44	66.877	66.877
SIN	58	0.277	0.597	0.066	0.117	1	4	35.7	1105
SON	1	0.465	0.465	0.043	0.043	6.7	6.7	6300	6300
STE	27	0.514	1.364	0.091	0.302	2.01	4.03	640	4615
STR	15	0.345	0.835	0.047	0.223	0.950	4.62	417	6300
TAY	11	0.348	0.878	0.077	0.160	0.89	2.09	13.979	2269.7
VAB	12	0.234	0.772	0.071	0.169	0.7	2.8	37	2500
VAH	6	0.319	0.558	0.176	0.238	0.642	1.303	31	1490
WLM	5	0.538	0.669	0.204	0.223	0.912	2.14	31.125	196.1
WLS	61	0.358	1.360	0.110	0.302	0.269	1.98	102	11700
WSA	195	0.165	0.555	0.034	0.170	1	2	11.3	587.19
WSB	36	0.444	0.578	0.108	0.176	1	2	55.8	379
WSS	13	0.377	0.388	0.073	0.075	1	1	53.8	94.6
ZNA	13	0.224	0.783	0.05	0.783	1.66	4.7	150	1975
Total	984	0.131	1.802	0.029	0.585	0.248	9.5	10.2	47300

(a)

Range of field variables									
Code	No.	Velocity (m/s)		Depth (m)		Slope (‰)		C_t (ppm)	
		min	max	min	max	min	max	min	max
AMC	5	0.473	0.739	0.796	1.009	0.237	0.33	52	448
ATC	6	1.739	2.028	10.881	14.112	0.038	0.0513	102.374	567.343
CHO	26	0.846	1.597	2.103	3.414	0.115	0.254	149.826	1316.9
COL	58	0.617	1.266	1.134	3.371	0.107	0.407	35.6	768.7
HII	34	0.186	0.930	0.025	0.732	0.84	10.7	116.311	5638.6
MID	35	0.593	1.125	0.247	0.412	0.928	1.572	437.760	2269.2
MIS	5	1.756	2.423	11.400	17.282	0.082	0.134	178.001	511.707
MOU	91	0.366	1.350	0.040	0.438	1.36	3.15	26.763	2600.6
NIO	40	0.625	1.271	0.398	0.588	1.136	1.799	392	2750
RGC	8	0.805	1.518	0.923	1.512	0.53	0.8	674	2695
RGR	254	0.295	2.384	0.159	2.326	0.69	2.31	11	11400
RIO	38	0.624	2.384	0.332	1.463	0.74	0.89	463.65	4544.38
Total	600	0.186	2.423	0.025	17.282	0.038	10.7	11	11400
(b)									

Table 4. (a) Range of laboratory variables, (b) Range of field variables

Data-driven techniques can be used for data mining since the only prerequisite for their function is the determination of the input parameters without the need to predefine the structure of the model and the degree of nonlinearity. The determination of the input parameters for the data-driven schemes will be made with a tentative assessment through a trial-and-error procedure. The correlation coefficient *r* has been employed in order to reveal any existing linear dependence in log-log plots between sediment concentration and any of the variables listed in Table 1

$$r = \frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sqrt{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 \sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}} \tag{15}$$

where *Y* denotes sediment concentration and *X* denotes the independent variable. Table 5 shows the correlation coefficient for log-log plots for the flume and field data of Tables 4a and 4b. From the techniques proposed, the trial-and-error process will be accomplished with the aid of ANNs, due to their speed, and after the determination of the most promising combinations that may serve as an input layer, the other data-driven techniques will be implemented as well.

	VS/ω	$VS/(gd_{50})^{1/2}$	ω^*	τ^*	Fr	Re	Re^*	d_{gr}
Flume data r	0.862	0.885	0.754	0.681	0.759	0.463	-0.014	-0.314
Field data r	0.730	0.687	0.601	0.587	0.492	0.148	-0.144	-0.405

Table 5. Correlation between sediment concentration and independent dimensionless variables of the flume and field data of Table 4 in log-log plots

The findings shown in Table 5 partially agree with the diagrams depicted in Figure 1, since sediment discharge is best correlated with unit stream power and stream power both for laboratory and for field data.

After the tentative assessment based on ANNs, of several input combinations, the most potent ones, which will be applied to the data-driven schemes, seem to be those listed in Table 6. These combinations include the independent variables of Eq. (3) and others that are relatively easily measured and commonly used in engineering. It is noteworthy that all combinations comprise dimensionless grain diameter and Froude number among others. Whilst Froude number gives a measure of the ratio of inertial forces to gravitational forces of the flow and is a commonly used variable in hydraulic engineering, the potential usage of dimensionless grain diameter is twofold. Firstly, it introduces kinematic viscosity and median grain diameter and secondly provides homogeneity in the input data. The necessity for the provided homogeneity can be seen from combination (a) where shear Reynolds number, which essentially includes dimensionless grain diameter, is included as well. The absence of any of these two terms in combination (a) has detrimental effects in the predictive capability of the generated model. The other variables for the combinations examined herein are those that most sediment transport formulae rely heavily on, namely dimensionless unit stream power, dimensionless stream power and dimensionless shear stress, and are best correlated with sediment concentration as shown in Table 5. For combination (a) Yang's dimensionless unit stream power was preferred to Vanoni's because, despite the fact that the calculation of fall velocity may be problematic, it reduced significantly the sum of errors between calculated and observed values. The other two combinations (b) and (c) comprise just three variables because shear is embedded in dimensionless stream power and dimensionless shear stress, respectively. Furthermore, it seems that there is no other potential input combination, besides those listed in Table 6, since any other combination tested gave results that declined by orders of magnitude.

	Input combinations
a	d_{gr}, Fr, Re^*, VS/ω_s
b	d_{gr}, Fr, ω^*
c	d_{gr}, Fr, τ^*

Table 6. Input combinations that will be applied to the data-driven schemes

5. Applications and results

The potential of training a DDM scheme solely with flume data and subsequently applying it to a test set comprising exclusively field data has been shown in Kitsikoudis et al. (2012a, 2012b) where ANNs and symbolic regression were utilized, respectively, for the prediction of sediment concentration in sand bed rivers. In these studies, however, the data were not subjected to elaboration and screening, in order to demonstrate the potential modeling ability of this technique with crude data. As a result, input data were kept in large numbers, and the generated models yielded very good results, better than those obtained from the common sediment transport formulae. However, it is known that the incorporation of knowledge can be proved beneficial to the predictive capability of DDM schemes as long as this is accomplished by transformation and elaboration of the fundamentals. Sediment transport and open channel hydraulics rely heavily on empirical equations and ideal flows; therefore, data transformation based on such assumptions does not guarantee the enhancement of the predictive capabilities of the DDM scheme. Nevertheless, the sidewall correction of Vanoni and Brooks (1957) was applied for the proper calculation of the shear stress in flume measurements and additionally the restrictions of Table 2 were imposed to the data for the removal of various biases resulting to a significantly reduced data amount. On the contrary, a criterion for the initiation of motion has been omitted, due to the stochastic character of turbulence, and was left up to the DDM scheme to define the effective portion of the flow that quantifies the transport rate.

Since every data-driven technique has its own syntax, the three possible input combinations of Table 6 are tested individually with the aid of both ANNs and symbolic regression. The evaluation of the modeled results P_i with respect to the observed ones O_i will be made on the basis of the root mean square error (*RMSE*),

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(O_i - P_i\right)^2}{N}} \tag{16}$$

coefficient of determination (R^2) or Nash-Sutcliffe model efficiency coefficient (*E*) (Nash and Sutcliffe, 1970),

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(O_i - P_i\right)^2}{\sum_{i=1}^{N}\left(O_i - \bar{O}\right)^2} \tag{17}$$

and discrepancy ratio (*DR*). The latter is the percentage of calculated concentrations that lie between one half and two times the respective measured concentrations.

5.1. ANNs application

This study was implemented in MATLAB with the aid of the neural network toolbox (Demuth et al., 2009). Since the usage of Levenberg-Marquardt training function gave the best results in a similar study in Kitsikoudis et al. (2012a), it was utilized for training in this application as well. Due to the importance of the initial values of the synaptic weights in the search for local minima of the error function, which is the mean square error between calculated and observed values, a MATLAB code was written, which determines the most efficient ANN within 5000 training executions, for each network architecture, with random initial weights for every repetition. The most efficient ANN is taken to be the one that yields only positive sediment concentrations, in order for the results to have physical meaning, and after the training provides the highest *DR* in the test set. For this evaluation, *DR* is preferred over *RMSE*, because the latter emphasizes on large concentrations. Models that derived slightly worse results than others, but had much simpler structure were preferred due to the principle of parsimony. Figures 3-5 depict the scatter plots of the best derived models, for each input combination of Table 6, for the field data of the test set. These models that perform best are described in Table 7. Table 8 shows the best models and their performance measures for the training, validation and test sets. Finally, Table 10 shows a comparison between the ANN induced models and some of the commonly used sediment transport functions for the rivers data constituting the test set. It should be mentioned that several of these formulae are calibrated with part of the data (especially the Brownlie formula) that are used for the comparison and despite that significant advantage they still generate inferior results to those of the ANNs.

	Input combination from Table 6	Network architecture (neurons in input-hidden-output layers)
ANN (a)	a	4-5-1
ANN (b)	b	3-6-1
ANN (c)	c	3-11-2-1

Table 7. Best performing models for each possible input combination

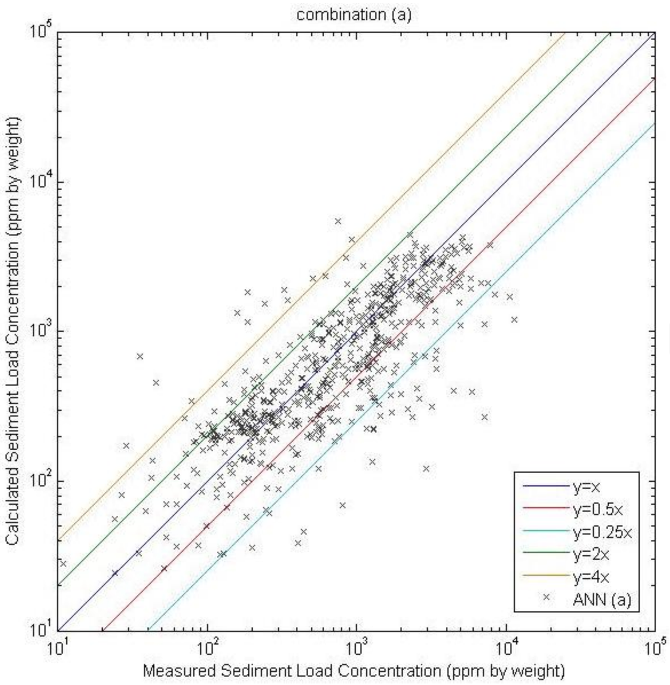

Figure 3. Scatter plot for the field data of the test set, of measured sediment concentration and computed from ANN, based on input combination (a)

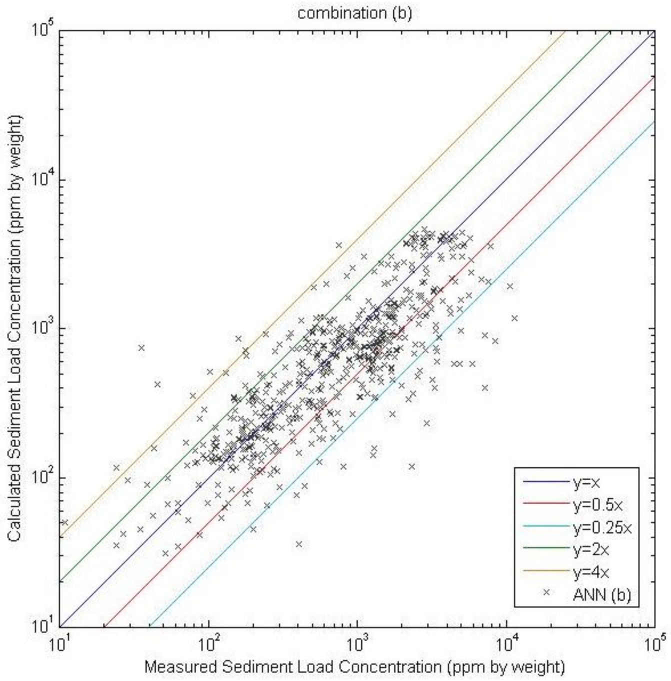

Figure 4. Scatter plot for the field data of the test set, of measured sediment concentration and computed from ANN, based on input combination (b)

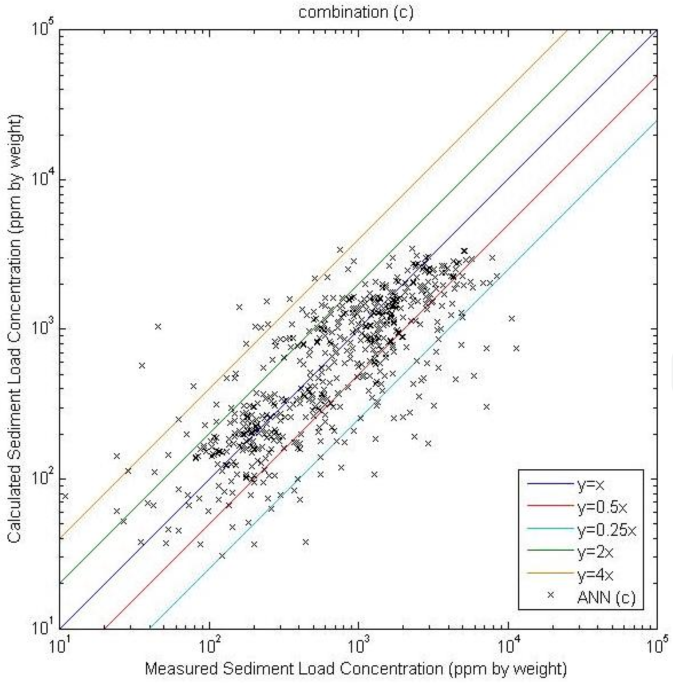

Figure 5. Scatter plot for the field data of the test set, of measured sediment concentration and computed from ANN, based on input combination (c)

		DR 0.5-2 (%)	*DR 0.25-4 (%)*	*RMSE*	R^2
ANN (a)	Training set	85.12	98.51	722.20	0.9213
	Validation set	82.86	97.55	936.72	0.5582
	Test set	72.50	92.33	1168.76	0.3687
ANN (b)	Training set	82.81	97.56	734.13	0.9187
	Validation set	85.71	95.92	884.41	0.6062
	Test set	71.67	93.17	1202.69	0.3315
ANN(c)	Training set	84.84	98.24	509.42	0.9608
	Validation set	84.08	98.37	872.36	0.6169
	Test set	70.67	91.33	1221.72	0.3102

Table 8. Performance measures of the optima ANNs

From Table 8 can be inferred that any of the three combinations listed in Table 6 has its own merit and that sediment transport can be quantified by physical quantities that can be either vectors or scalars.

5.2. Symbolic regression application

The basic computation tool for the implementation of symbolic regression is provided by GPTIPS (Searson, 2009), which is an open source MATLAB toolbox. Since every problem has its own peculiarities, proper adjustments must be made to the GPTIPS parameters in order to obtain good results. The most important parameters are the population size, the number of generations, the using functions, the maximum number of genes and the maximum tree depth. Searson et al. (2010) have found that enforcing stringent tree depth restrictions often allows the evolution of relatively compact models that are linear combinations of low order nonlinear transformations of the input variables. After several runs, only input combination (b) gave results superior to those of the classical formulae. The GPTIPS derived formula for this combination is the following

$$C_t = 1542\left(\omega^* Fr^3\right)^{0.4} - 0.4794\frac{\sqrt{d_{gr}Fr}\left(d_{gr} + Fr - \omega^*\right) + \dfrac{\omega^*}{Fr}}{\left(\dfrac{6.218}{\omega^*} - Fr + 1\right)^3} - 313.6 \tag{18}$$

Figure 6 depicts the scatter plot of measured and calculated from Eq. (18) sediment concentrations for the field data of the test set, whilst Table 9 and Table 10 show its performance for the training, validation and testing set, and the comparison with other formulae, respectively.

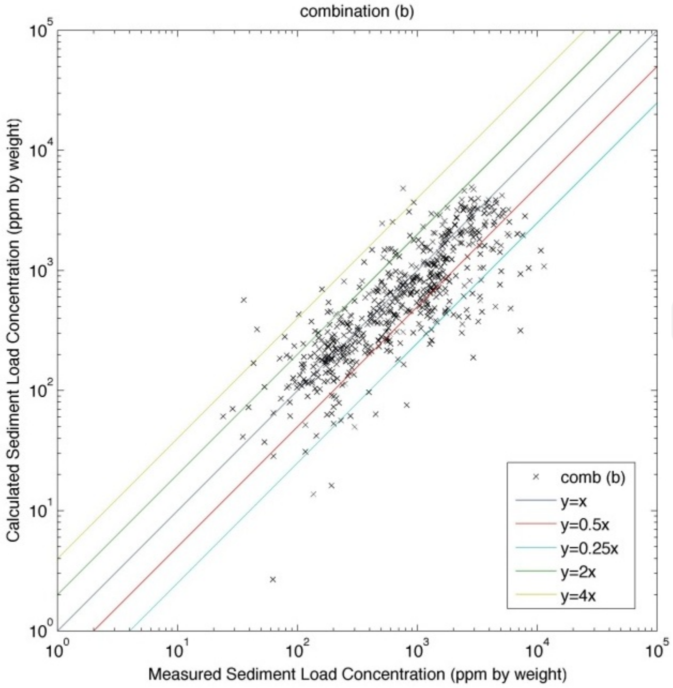

Figure 6. Scatter plot for the field data of the test set, of measured sediment concentration and computed from symbolic regression, based on input combination (b)

	DR 0.5-2 (%)	*DR 0.25-4 (%)*	*RMSE*	*R^2*
Training set	68.20	84.84	632.21	0.9397
Validation set	66.53	85.31	964.70	0.5315
Test set	71.00	91.33	1218.12	0.3143

Table 9. Performance measures of symbolic regression, based on combination (b)

	DR 0.5-2 (%)	*DR 0.25-4 (%)*	*RMSE*	*R^2*
ANN (a)	72.50	92.33	1168.76	0.3687
ANN (b)	71.67	93.17	1202.69	0.3315
ANN (c)	70.67	91.33	1221.72	0.3102
Symb. regression (b)	71.00	91.33	1218.12	0.3143
Ackers & White	58.33	88.67	1405.38	0.0872
Brownlie	68.33	91.00	1274.44	0.2494
Engelund & Hansen	69.67	92.33	1244.83	0.2838
Karim & Kennedy	64.83	91.50	1341.34	0.1685
Molinas & Wu	52.83	84.83	1423.06	0.0641
Yang	49.50	83.33	1403.07	0.0902

Table 10. Comparison of ANNs of the input combinations (a), (b) and (c) and Eq. (18), derived from symbolic regression for the combination (b), with sediment transport formulae based on the river data of the test set

The results obtained from ANNs for all the combinations are superior to those of the classical sediment transport formulae in terms of *DR*, *RMSE* and R^2. Combination (a) performed best in all evaluation measures, besides the second *DR* criterion in the range 0.25-4 where combination (b) gave better results. The third combination came up third with respect to all evaluation measures. However, these results by no means can be considered conclusive, since it is essentially unknown whether they are the best results derived from the ANN or just results obtained from the trapping in a local minimum of the minimization process in the network's training algorithm. From the results generated from symbolic regression, only combination (b) managed to surpass the classical sediment transport functions. The other two combinations gave results inferior to those of Engelund and Hansen and Brownlie formulae, but superior to those of the others. In addition, symbolic regression derived its best results without utilizing the log-transformation of the input data. Regarding the other sediment transport functions, the formula of Engelund and Hansen performed best. The small values of the coefficient of determination R^2 in Table 10 reflect the difficulty of predicting sediment transport rates in natural streams and rivers, due to random turbulent bursts that accentuate the stochastic nature and exacerbate the complexity of the problem.

Although these results cannot be considered conclusive, it seems that the ANNs yield better results. GPTIPS sometimes (usually when only a few input variables are involved) lags behind a neural network model in terms of raw predictive performance, but the equivalent GP models are often simpler, shorter and may be open to physical interpretation (Searson, 2009). This is partially due to the fact that ANNs are much faster than the time consuming GP and for given time they can run multiple times comparing to GP. Moreover, since the testing set comes from a database with different statistical distributions than the one from which the training set originates, the exploration of as many as possible local minima of the training function may prove beneficial to the training process. ANNs have this property, whilst GP is based on a stochastic concept seeking the global minimum. This may be one reason for the superiority of ANNs in this study, where the training data comprise flume measurements, whilst the testing data consists of field measurements.

6. Conclusions

This study utilized two widely used data-driven techniques, namely ANNs and symbolic regression, in a novel way since the data used for training and those used for testing came from datasets with different statistical distributions. This difference is owned to the fact that the training and validation set comprises exclusively laboratory flume data, while the testing set consists solely of field data. Based on this concept, the inclusion of noise emanated from the field measurements will not be embedded in the training data and additionally the generated models will have general applicability since the inclusion of field data in the training set would confine them to the specific streams from which the data were obtained. The determination of the input parameters was accomplished by a tentative assessment of some of the widely used dimensionless parameters in sediment transport and open channel hydraulics. This assessment showed that three combinations had the potential to serve as inputs and were involved in this application, in which they all yielded very good results, better than those obtained from the commonly used formulae on the basis of root mean square error and the ratio of computed to measured transport rates. Unit stream power, stream power, and shear stress were the dominant independent variables of the three combinations, respectively, and the results have shown that each one, of these widely used variables in the context of sediment transport, has its own merit. The results generated from the ANNs were better from those obtained from symbolic regression; however, the explicit equation that was derived from the latter can be more easily interpreted. Finally, the results obtained in this study may enhance the confidence in using data-driven techniques, despite their black-box nature, because, in order to perform well in a dataset from a different system from the one they were trained, the induced equations must have physical meaning.

Notation

The following symbols are used in this chapter:

C_t = sediment concentration by weight in parts per million (ppm)

D = mean flow depth (m)

Q = water discharge (m^3/s)

Q_{st} = sediment discharge (m^3/s)

R = hydraulic radius (m)

S = energy slope

T = water temperature (^{o}C)

u_* = shear velocity (m/s)

u_{*c} = critical shear velocity (m/s)

V = mean flow velocity (m/s)

W = channel width (m)

d = grain diameter (m)

d_{50} = median grain diameter (m)

f = friction factor

g = gravitational acceleration (m/s^2)

γ = specific weight of water (N/m^3)

γ_s = specific weight of sediment (N/m^3)

ν = kinematic viscosity of water (m^2/s)

ϱ = density of water (kg/m^3)

ϱ_s = density of sediment (kg/m^3)

σ_g = geometric standard deviation of bed particles [$(d_{84}/d_{50} + d_{50}/d_{16})/2$]

τ = shear stress [kg/(m.s^2)]

τ^* = dimensionless shear stress

ω = stream power (kg/s^3)

ω^* = dimensionless stream power

ω_s = settling or fall velocity of sediment (m/s)

Appendix A

In flume experiments, the sand covered bed will generally be much rougher than the flume walls, and thus will be subjected to higher shear stress. Separation of the shear force exerted on the bed from that on the lateral boundaries was first proposed by Einstein (1950). The line of analysis pursued as follows is that proposed by Johnson (1942) and modified by Vanoni and Brooks (1957). The principal assumption is that the cross-sectional area can be divided into two parts, A_b and A_w, in which the streamwise component of the gravity force is resisted by the shear force exerted in the bed and walls, respectively. It is further assumed that the mean velocity and energy gradient are the same for A_b and A_w, and that the Darcy-Weisbach relation can be applied to each part of the cross section as well as to the whole, i.e.

$$\frac{V^2}{S} = \frac{8gA}{fp} = \frac{8gA_b}{f_b p_b} = \frac{8gA_w}{f_w p_w} \tag{19}$$

in which, p = the wetted perimeter; and the subscripts b and w refer to the bed and wall sections, respectively. For a rectangular channel p=2D+W; p_w=2D; p_b=W. Introducing the geometrical requirement $A=A_b+A_w$ into Eq. (19) results in

$$f_b = f + \frac{2D}{W}\left(f - f_w\right) \tag{20}$$

The wall friction factor f_w is further related to the ratio of Re/f, where Re=4VR/ν and f can be calculated from the experimental data. This relationship, which was originally given as a graph of f_w against Re/f by Vanoni and Brooks (1957), can also be described by the function

$$f_w = \left[20\left(\mathrm{Re}/f\right)^{0.1} - 39\right]^{-1} \tag{21}$$

which is obtained by curve fitting (Cheng and Chua, 2005). Finally, f_b is calculated from Eq. (20) and $R_b = A_b/p_b$ from Eq. (19). R_b is consequently used for the calculation of the bed shear velocity and bed shear stress.

Despite its several obvious deficiencies (division of the cross section into two noninteracting parts, determination of friction factors for section components on the basis of a pipe friction diagram, use of the same mean velocity for each subsection, etc.), the side-wall correction procedure appears to yield fairly reliable estimates of the friction factors for flow over sand beds with no flume walls present (Vanoni, 2006).

Appendix B

For the calculation of particle fall velocity in a clear, still fluid, van Rijn (1984) suggested the use of the Stokes law for sediment particles smaller than 0.1 mm

$$\omega_s = \frac{1}{18}\frac{\rho_s - \rho}{\rho} g \frac{d^2}{\nu} \tag{22}$$

For suspended sand particles in the range 0.1 to 1 mm, the following type of equation, as proposed by Zanke (1977), can be used

$$\omega_s = 10\frac{\nu}{d}\left\{\left[1 + 0.01\left(\frac{\rho_s}{\rho} - 1\right)\frac{gd^3}{\nu^2}\right]^{1/2} - 1\right\} \tag{23}$$

For particles larger than about 1 mm, the following simple equation can be used (van Rijn, 1982)

$$\omega_s = 1.1\left[\left(\frac{\rho_s}{\rho} - 1\right)gd\right]^{1/2} \tag{24}$$

Appendix C

Ackers and White formula: Ackers and White (1973) applied dimensional analysis to express the mobility and transport rate of sediment in terms of some dimensionless parameters. It has been shown that the transport of fine materials is best related to gross shear, shear velocity being the representative variable, and that the transport of coarse materials is best related to the net grain shear, mean velocity being the representative variable. The following equations do not necessarily apply in an upper phase of transport. However, it was shown that the following relationships are not sensitive to bed form; they apply to plain, rippled, and duned configurations. Their mobility number for sediment is

$$F_{gr} = \frac{U_*^n}{\sqrt{gd\left(\gamma_s/\gamma - 1\right)}}\left[\frac{V}{\sqrt{32}\log\left(10D/d\right)}\right]^{1-n} \tag{25}$$

Coefficients C, A, m and n are related to the dimensionless grain diameter d_{gr} based on best-fit curves of laboratory data with sediment sizes greater than 0.04 mm and Froude numbers less than 0.8. They are shown in Table 11.

$$d_{gr} = d\left[\frac{g\left(\gamma_s/\gamma - 1\right)}{\nu^2}\right]^{1/3} \tag{26}$$

Finally, they related the bed material load to the mobility number as follows

$$G_{gr} = \frac{XD}{d\gamma_s/\gamma}\left(\frac{U_*}{V}\right)^n = C\left(\frac{F_{gr}}{A} - 1\right)^m \tag{27}$$

where X = rate of sediment transport in terms of mass flux per unit mass flow rate

$d_{gr} \geq 60$	$1 < d_{gr} < 60$
$n = 0.00$	$n = 1.00\text{-}0.56 \log d_{gr}$
$A = 0.17$	$A = 0.23 d_{gr}^{-0.5} + 0.14$
$m = 1.50$	$m = 9.66 d_{gr}^{-1} + 1.34$
$C = 0.025$	$\log C = -3.53 + 2.86 \log d_{gr} - (\log d_{gr})^2$

Table 11. Coefficients of the Ackers and White formula

Brownlie formula: The Brownlie (1981a) relations are based on regressions of over 1000 experimental and field data points. For normal or quasi-normal flow, the transport relation takes the form

$$C_t = 7115 c_f \left(F_g - F_{go}\right)^{1.978} S^{0.6601}\left(\frac{R}{d_{50}}\right)^{-0.3301} \tag{28}$$

where

$$F_g = \frac{V}{\sqrt{\left(\rho_s/\rho - 1\right) g d_{50}}} \tag{29}$$

$$F_{go} = 4.596\left(\tau_c^*\right)^{0.5293} S^{-0.1405} \sigma_g{}^{-0.1606} \tag{30}$$

$$\tau_c^* = 0.22Y + 0.06 \cdot 10^{-7.7Y} \tag{31}$$

$$Y = \left(\frac{\sqrt{\left(\rho_s/\rho - 1\right) g d_{50}}\, d_{50}}{v} \right)^{-0.6} \tag{32}$$

c_f = 1 for laboratory flumes and 1.268 for field channels.

Engelund and Hansen formula: Using Bagnold's stream power concept and the similarity principle, Engelund and Hansen (1967) established the following sediment transport formula

$$f'\phi = 0.1\theta^{5/2} \tag{33}$$

where

$$f' = \frac{2gDS}{V^2} \tag{34}$$

$$\phi = \frac{q_t}{\rho_s \sqrt{\left(\rho_s/\rho - 1\right) g d_{50}{}^3}} \tag{35}$$

$$\theta = \frac{\tau}{\left(\gamma_s - \gamma\right) d_{50}} \tag{36}$$

where q_t = total sediment discharge by weight per unit width. Strictly speaking, the Engelund and Hansen formula should be applied to those flows with dune beds in accordance with the similarity principle. However, many tests have shown that it can be applied to the upper flow regime with particle size greater than 0.15 mm without serious deviation from the theory.

Karim and Kennedy formula: Karim and Kennedy (1990) applied nonlinear multiple regression analysis to derive relations among flow velocity, sediment discharge, bed form geometry, and friction factor of alluvial rivers. A database comprising 339 river flows and 608 flume flows was used in their analysis. The obtained sediment load predictor is given by

$$\begin{aligned} \log \frac{q_s}{\sqrt{\left(\gamma_s/\gamma - 1\right) g d_{50}^3}} &= -2.279 + 2.972 \log \frac{V}{\sqrt{\left(\gamma_s/\gamma - 1\right) g d_{50}}} \\ &+ 1.060 \log \frac{V}{\sqrt{\left(\gamma_s/\gamma - 1\right) g d_{50}}} \log \frac{U_* - U_{*c}}{\sqrt{\left(\gamma_s/\gamma - 1\right) g d_{50}}} + 0.299 \log \frac{D}{d_{50}} \log \frac{U_* - U_{*c}}{\sqrt{\left(\gamma_s/\gamma - 1\right) g d_{50}}} \end{aligned} \tag{37}$$

where q_s =volumetric total sediment discharge per unit width.

Molinas and Wu formula: This empirical relation is based on Velikanov's gravitational power theory, which assumes that the power available in flowing water is equal to the sum of the power required to overcome flow resistance and the power required to keep sediment in suspension against gravitational forces. Molinas and Wu (2001) argued that the predictors of Ackers and White, Engelund and Hansen, and Yang have been developed with flume experiments representative of shallow flows and cannot be applied to large rivers having deep flow conditions. Motivated by the need for having a total bed material load predictor for application to large sand bed rivers, they used stream power and energy considerations together with data from large rivers (e.g., Amazon, Atchafalaya, Mississippi, Red River), to obtain an empirical fit for the total bed material load concentration in ppm

$$C_t = \frac{1430\left(0.86 + \sqrt{\Psi}\right)\Psi^{1.5}}{0.016 + \Psi} \tag{38}$$

where Ψ = universal stream power, which is defined as

$$\Psi = \frac{V^3}{g\left(\rho_s/\rho - 1\right)D\omega_s\left[\log\left(D/d_{50}\right)\right]^2} \tag{39}$$

One advantage of this approximation is that the energy slope does not have to be measured directly, which is always a challenge in large alluvial rivers. On the other hand, since Molinas and Wu (2001) do not mention how the wash load was separated from the bed material load and the same large river data were used both to develop and to test their formulation, Eq. (38) might overestimate bed material load concentrations when applied to other large rivers not included in the calibration (Garcia, 2008).

Yang formula: To determine total sediment concentration, Yang (1973) used Buckingham's π theorem and the concept of unit stream power, which is given by the product of mean flow velocity and energy slope. The coefficients in Yang's equation were determined by running a multiple regression analysis for 463 sets of laboratory data. The equation obtained is

$$\begin{aligned}\log C_t &= 5.435 - 0.286\log\frac{\omega_s d_{50}}{\nu} - 0.457\log\frac{U_*}{\omega_s} \\ &+\left(1.799 - 0.409\log\frac{\omega_s d_{50}}{\nu} - 0.314\log\frac{U_*}{\omega_s}\right)\log\left(\frac{VS}{\omega_s} - \frac{V_{cr}S}{\omega_s}\right)\end{aligned} \tag{40}$$

The critical dimensionless unit stream power $V_{cr}S/\omega$ is the product of the dimensionless critical velocity V_{cr}/ω and the energy slope S, where

$$\frac{V_{cr}}{\omega_s}=\begin{cases}\dfrac{2.5}{\log\left(\dfrac{U_* d_{50}}{\nu}\right)-0.06}+0.66 & for \quad 1.2<\dfrac{U_* d_{50}}{\nu}<70 \\ 2.05 & for \quad 70\leq\dfrac{U_* d_{50}}{\nu}\end{cases} \tag{41}$$

Author details

Vasileios Kitsikoudis[1], Epaminondas Sidiropoulos[2] and Vlassios Hrissanthou[1]

1 Department of Civil Engineering, Democritus University of Thrace, Xanthi, Greece

2 Department of Rural and Surveying Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

References

[1] Ackers P., White W. R. (1973). Sediment Transport: New Approach and Analysis. Journal of the Hydraulics Division, ASCE; 99(HY11) 2041-2060.

[2] Azamathulla H. Md, Ab Ghani A., Zakaria N. A., Guven A. (2010). Genetic Programming to Predict Bridge Pier Scour. Journal of Hydraulic Engineering, ASCE; 136(3) 165-169.

[3] Babovic V. (2000). Data Mining and Knowledge Discovery in Sediment Transport. Computer Aided Civil and Infrastructure Engineering; 15(5) 383-389.

[4] Babovic V., Abbott M. B. (1997). The Evolution of Equations from Hydraulic Data. Part II: Applications. Journal of Hydraulic Research, IAHR; 35(3) 411-430.

[5] Babovic V., Keijzer M. (2000). Genetic Programming as a Model Induction Engine. Journal of Hydroinformatics; 2(1) 35-60.

[6] Bagnold R. A. (1956). Flow of Cohesionless Grains in Fluids. Philosophical Transactions of the Royal Society [London], Series A; 249 235-297.

[7] Bagnold R. A. (1966). An Approach to the Sediment Transport Problem from General Physics. Prof. Paper 422-I. U.S. Geological Survey.

[8] Bhattacharya B., Price R. K., Solomatine D. P. (2004). A Data Mining Approach to Modelling Sediment Transport. In: Liong S., Phoon K., Babovic V. Proceedings of the 6th International Conference on Hydroinformatics, 21-24 June 2004, Singapore. World Scientific.

[9] Bhattacharya B., Price R. K., Solomatine D. P. (2005). Data Driven Modeling in the Context of Sediment Transport. Physics and Chemistry of the Earth; 30(4-5) 297-302.

[10] Bhattacharya B., Price R. K., Solomatine D. P. (2007). Machine Learning Approach to Modeling Sediment Transport. Journal of Hydraulic Engineering, ASCE; 133(4) 440-450.

[11] Brownlie W. R. (1981a). Prediction of Flow Depth and Sediment Discharge in Open Channels. Report No. KH-R-43A. Pasadena, California: W. M. Keck Laboratory of Hydraulics and Water Resources, California Institute of Technology.

[12] Brownlie W. R. (1981b). Compilation of Alluvial Channel Data: Laboratory and Field. Report No. KH-R-43B. Pasadena, California: W. M. Keck Laboratory of Hydraulics and Water Resources, California Institute of Technology.

[13] Cheng N. S., Chua L. H. C. (2005). Comparisons of Sidewall Correction of Bed Shear Stress in Open-Channel Flows. Journal of Hydraulic Engineering, ASCE; 131(7) 605-609.

[14] Cigizoglu H. K. (2004). Estimation and Forecasting of Daily Suspended Sediment Data by Multi-Layer Perceptrons. Advances in Water Resources; 27(2) 185-195.

[15] Demuth H. B., Beale M. H., Hagan M. T. (2009). Neural Network Toolbox: For Use With MATLAB. The Mathworks Inc.; 2009.

[16] Eaton B. C., Church M. (2011). A Rational Sediment Transport Scaling Relation Based on Dimensionless Stream Power. Earth Surface Processes and Landforms; 36(7) 901-910.

[17] Einstein H. A. (1942). Formulas for the transportation of Bed Load. Transactions, ASCE; 107 (Paper No. 2140) 561-573.

[18] Einstein H. A. (1950). The Bedload Function for Sediment Transportation in Open Channel Flows. Technical Bulletin No. 1026. Washington D.C.: U.S. Department of Agriculture, Soil Conservation Service.

[19] Engelund F., Hansen E. (1967). A Monograph on Sediment Transport in Alluvial Streams. Copenhagen: Teknisk Vorlag.

[20] Garcia M. H. (2008). Sediment Transport and Morphodynamics. In: Garcia M. H. (ed.) ASCE Manuals and Reports on Engineering Practice No. 110, Sedimentation Engineering: Processes, Measurements, Modeling, and Practice. Virginia, U.S.A.: ASCE, p21-163.

[21] Haykin S. (2009). Neural Networks and Learning Machines, 3rd Edition. New Jersey: Prentice Hall.

[22] Hornik K., Stinchcombe M., White H. (1989). Multilayer Feedforward Networks are Universal Approximators. Neural Networks; 2(5) 359-366.

[23] Johnson J. W. (1942). The Importance of Considering Side-Wall Friction in Bed-Load Investigations. Civil Engineering, ASCE; 12(6) 329-331.

[24] Karim M. F., Kennedy J. F. (1990). Menu of Coupled Velocity and Sediment-Discharge Relations for Rivers. Journal of Hydraulic Engineering, ASCE; 116(8) 978-996.

[25] Kitsikoudis V., Sidiropoulos E., Hrissanthou V. (2012a). A New Approach for ANN Modeling of Sediment Transport in Sand Bed Rivers. Proceedings of the 10th International Conference on Hydroinformatics, CD-ROM format, 14-18 July 2012, Hamburg, Germany.

[26] Kitsikoudis V., Sidiropoulos E., Hrissanthou V. (2012b). Implementation of Multigene Symbolic Regression in the Sediment Transport Quantification Problem for Sand Bed Rivers. Proceedings of the 9th International Symposium on Ecohydraulics, CD-ROM format, 17-21 September 2012, Vienna, Austria.

[27] Knapp R. T. (1938). Energy Balance in Stream Flows Carrying Suspended Load. Transactions, American Geophysical Union; 501-505.

[28] Koza J. (1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge MA: MIT Press.

[29] Maier H. R., Dandy G. C. (2000). Neural Networks for the Prediction and Forecasting of Water Resources Variables: A Review of Modeling Issues and Applications. Environmental Modeling and Software; 15(1) 101-124.

[30] Minns A. W. (2000). Subsymbolic Methods for Data Mining in Hydraulic Engineering. Journal of Hydroinformatics; 2(1) 3-13.

[31] Molinas A., Wu B. (2001). Transport of Sediment in Large Sand-Bed Rivers. Journal of Hydraulic Research, IAHR; 39(2) 135-146.

[32] Nagy H. M., Watanabe K., Hirano M. (2002). Prediction of Sediment Load Concentration in Rivers Using Artificial Neural Network Model. Journal of Hydraulic Engineering, ASCE; 128(6) 588-595.

[33] Nash J. E., Sutcliffe J. V. (1970). River Flow Forecasting Through Conceptual Models, Part I – A Discussion of Principles. Journal of Hydrology: 10(3) 282-290.

[34] Pugh C. A. (2008). Sediment Transport Scaling for Physical Models. In: Garcia M. H. (ed.) ASCE Manuals and Reports on Engineering Practice No. 110, Sedimentation Engineering: Processes, Measurements, Modeling, and Practice. Virginia, U.S.A.: ASCE, p1057-1066.

[35] Rubey W. W. (1933). Equilibrium Conditions in Debris-Laden Streams. Transactions, American Geophysical Union 14th Ann. Mtg; 497-505.

[36] Rumelhart D. E., Hinton G. E., Williams R. J. (1986). Learning Representations of Back-Propagation Errors. Nature (London); 323(9) 533-536.

[37] Searson D. (2009). GPTIPS: Genetic Programming & Symbolic Regression for MATLAB User Guide.

[38] Searson D. P., Leahy D. E., Willis M. J. (2010). GPTIPS: An Open Source Genetic Programming Toolbox for Multigene Symbolic Regression. In Ao S. I., Castillo O., Douglas C., Feng D. D., Lee J. (eds.) Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. I, IMECS 2010, 17-19 March 2010, Hong Kong.

[39] Shields A. (1936). Application of Similarity Principles and Turbulence Research to Bedload Movement. Transl. into English by Ott W. P. and Van Uchelen J. C. Pasadena, California: California Institute of Technology; 1936.

[40] van Rijn L. C. (1982). Computation of Bed-Load and Suspended Load. Report S487-II, Delft Hydraulics Laboratory, Delft, The Netherlands.

[41] van Rijn L. C. (1984). Sediment Transport, Part II: Suspended Load Transport. Journal of Hydraulic Engineering, ASCE; 110(11) 1613-1641.

[42] Vanoni V. A. (1978). Predicting Sediment Discharge in Alluvial Channels. In: Water Supply and Management. Oxford: Pergamon Press. p399-417.

[43] Vanoni V. A. (2006). ASCE Manuals and Reports on Engineering No. 54, Sedimentation Engineering. Virginia, U.S.A.: ASCE.

[44] Vanoni V. A., Brooks N. H. (1957). Laboratory Studies of the Roughness and Suspended Load of Alluvial Streams. Sedimentation Laboratory Report No. E68. Pasadena, California: California Institute of Technology.

[45] Velikanov M. A. (1955). Dynamics of Channel Flow – v.2. In: Sediments and the Channel, 3rd Edition. Moscow: State Publishing House for Tech.– Theoretical Lit.; p107-120 [in Russian].

[46] Witten I. H., Frank E., Hall M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition. Burlington, MA: Morgan Kaufmann.

[47] Yalin M. S. (1977). Mechanics of Sediment Transport. Oxford: Pergamon Press.

[48] Yang C. T. (1972). Unit Stream Power and Sediment Transport. Journal of the Hydraulics Division, ASCE; 98(HY10) 1805-1836.

[49] Yang C. T. (1973). Incipient Motion and Sediment Transport. Journal of the Hydraulics Division, ASCE; 99(HY10) 1679-1704.

[50] Yang C. T. (1977). The Movement of Sediment in Rivers. Surveys in Geophysics; 3(1) 39-68.

[51] Yang C. T. (2003). Sediment Transport: Theory and Practice. Original edition McGraw-Hill; 1996. Reprint edition by Krieger Publication Company.

[52] Yang C. T., Marsooli R., Aalami M. T. (2009). Evaluation of Total Load Sediment Transport Formulas Using ANN. International Journal of Sediment Research; 24(3) 274-286.

[53] Zakaria N. A., Azamathulla H. Md, Chang C. K., Ab Ghani A. (2010). Gene Expression Programming for Total Bed Material Load Estimation – A Case Study. Science of the Total Environment; 408(21) 5078-5085.

[54] Zanke U. (1977). Berechnung der Sinkgeschwindigkeiten von Sedimenten. Mitt. des Franzius – Instituts für Wasserbau, Heft 46, Seite 243. Technical University Hannover.