GEOFIZIKA VOL. 32 2015

*DOI: 10.15233/gfz.2015.32.2*
*Original scientific paper*
*UDC 556.18*

# Prediction of bed load via suspended sediment load using soft computing methods

*Ali Osman Pektaş* [1] *and Emrah Doğan* [2]

[1] Bahcesehir University, Civil Engineering, Istanbul, Turkey

[2] Sakarya University, Civil Engineering, Sakarya, Turkey

*Received 9 September 2014, in final form 16 February 2015*

Appropriate and acceptable prediction of bed load being carried by streams is vitally important for water resources quantity and quality studies. Although measuring the rate of bed load in situ is the most consistent method, it is very expensive and cannot be conducted for as many streams as the measurement of suspended sediment load. Therefore, in this study the role of suspended load on bedload prediction was examined by using sensitivity analysis. On the other hand, conventional sediment rating curves and equations can not predict sediment load accurately so recently the usage of machine learning algorithms increase rapidly. Accordingly, soft computational methods are used in the study. These are; artificial neural network (ANN), support vector machine (SVM) models and a decision tree (CHAID) model that is not used before in sediment studies. Some particular parameters are frequently used in these soft computational methods to form input sets. Hence, well known and commonly used three input sets and a new generated set are used as inputs to predict bedload and then the suspended load variable is added in these input sets. The performances of models with respect to input sets are compared to each other. To generate the results and to push the limits of models a very skewed and heterogeneous data is collected from distributed locations. The results indicate that the performance of ANN and CHAID tree models are good when compared to SVM models. The usage of a suspended load as an additional input for the models boosts the model performances and the suspended load has significant contributions to all models.

*Keywords*: sediment prediction, bed load, suspended load, artificial neural networks, support vector machines, CHAID tree models

## 1. Introduction

Modeling and predicting the amount of bed load and suspended load is extremely important by the side of planning and handling the water resources projects. The sediment load transported by the streams may cause a decrease in a useful storage of a dam (Nakato, 1990; McBean and Al-Nassri, 1988). The transportation of sediment also changes ecologic and hydraulic equilibrium of the river bed. Furthermore, the design of steady channels, estimation of bedding and degradation at platform piers and abutments, estimation of sand and gravel mining effects, and the analysis of the ecological impact evaluation are also dependent to sediment load transport.

The sediment load of a stream is commonly determined from direct measurements or otherwise calculated indirectly by using sediment transport formulas. Even though direct measurement of sediment transport rate is more trustworthy, it is unfeasible and uneconomical to establish gauging stations at all desired locations and acquire data for a satisfactory long period of time. Furthermore, measurement of bed load is more expensive and complex than measurement of suspended load.

In the literature there are many sediment rate transportation models. These models have been proposed in different forms for many parameters as a function of the river and sediment characteristics. Some of them are obtained in a laboratory environment, while others are developed using in-situ data or theoretical methods. On the other hand, most of the sediment transport equations need comprehensive information on the channel, flow and sediment characteristics (Öztürk et al., 2001; Yang and Wan, 1991). With respect to the conditions under which the data are gathered, the same formulation could yield dissimilar scores of accuracy, and usually do not fit with the observed data. Therefore, none of such equations have achieved universal acceptance (Vanoni, 1971; Yang, 1996). Because of these facts it can be asserted that the assumptions stated in the derivation of these specific equations is only valid under certain situations and also is not to be regarded as a general rule (Yang, 1972). Because of the encountered difficulties, the researchers strive to search easy methods to estimate sediment load. Initially, such relationships have obtained by using regression analysis and usually these models are called as sediment rating curves (e.g. Jain, 2001; Ciğizoğlu, 2002a, b; Öztürk et al., 2001). But in this technique the interior uncertainties are not considered explicitly while determining the sediment yield with water discharge (Şen and Altunkaynak, 2003). Additionally, sediment rating curves does not contribute much on the insight of the physical meaning of used parameters and so, do not improve understanding of sediment transport processes (Yang, 1996). As known, the regression techniques can not determine the non linear relationships or it is only suitable to present simple non linearities after basic transformations. Recently, because of these problems, researchers are looking for simpler, cheaper and easier methods to predict sediment load, and they are beginning to use nonlinear models such as neural networks to solve nonlinear problems.

There are many implementations of artificial neural networks (ANN) at almost all branches of science. The method is famous for its capacity to model the nonlinear relationships and high predictive accuracy. Motivated by successful applications of ANNs have been applied in hydrological engineering problems. In hydrology the method has been emerged as a strong application for planning studies and management purposes. ANNs have been used for rainfall-runoff modeling, flow predictions, flow/pollution simulation, parameter identification, and modeling nonlinear/input-output time series (ASCE, 2000).

Jain (2001) used the ANN models to build up an integrated stage-discharge-sediment concentration relation for two watersheds of the Mississippi River. Ciğizoğlu (2002a, 2002b) used ANNs to analyze suspended sediment concentrations and made an assessment between ANNs and sediment rating curves for two catchments in the Northern England. He asserted that the results of ANN model are superior to classical sediment rating curve method. Nagy et al. (2002) made sediment discharge predictions that and concluded that the ANN model gives better results when compared to different widely used formulas of sediment discharge. They used Multi layer perceptrons (MLP) in their ANN model and indicated that MLP could capture the complex nonlinear behavior of the sedimentary series relatively better than the conventional models. Tayfur and Guldal (2005) computed the daily total suspended sediment in natural rivers by ANN and using a two dimensional unit sediment graph theory (2D-USGT) from precipitation data. The evaluation of results demonstrated that the ANN model has better performance than the 2D-USGT. Raghuwanshi et al. (2006) designed an ANN model to estimate both runoff and sediment yield in daily and weekly time frame, for a small watershed. When compared to ANN applications, the other machine learning algorithm implementations on sediment modeling is scarce and new. Bhattacharya et al. (2007) compared the ANN and decision tree (DT) model performances on the bed load transport dataset and concluded that both machine learning algorithms give sufficient results but the ANN model is superior to DT model. Oehler et al. (2012) used a boosted form of regression trees to to predict suspended-sediment reference (near-bed) concentration in six shelf areas of New Zeland. Kisi et al. (2012) compared the genetic programming results with ANN, adaptive neuro-fuzzy inference system (ANFIS) and support vector machine (SVM) results on the suspended load prediction. Misra et al. (2009) established SVM models to predict the daily, weekly and monthly discharge and sediment yield of an Indian watershed.

Since bed load observations are labor extensive and expensive, the sediment studies have been focused on total or suspended load models. A comprehensive study about bed load transportation is carried out by Sasal et al. (2009). The researchers used a large dataset and concluded that the developed ANN model gives satisfying predictive performance on bed load transport model studies. Yu et al. (2009) observed bed load sediment transportation rates of Diaoga River in China then they investigates the relations of bedload transport and some widely used non dimensional parameters such shear stress and stream power. Gao (2011) derived a power formula to predict maximum bed load transport rates by using nonlinear regression models.

There are three main purposes in this article; the first of them is evaluating the performances of widely-used machine learning algorithms while predicting the bed load sediment. For this purpose ANN, SVM and chi-squared automatic interaction detection (CHAID) Tree model is used and compared in the study. The second aim is to assess the performances of well known and commonly used input sets of bed load sediment models. The third aim is examining the role of suspended load observations while predicting the bed load so Suspended load parameter is added all built models and sensitivity analysis are performed. The results indicate that the performance of neural networks and CHAID models are good when compared to SVM models. The usage of a suspended load variable as an input for the models boosts the model performances and has a significant contribution on model accuracy. The input sets have a similar predictive performance but only the Pektaş (2015) input set has given sufficient results for all applied models.

## 2. Material and methods

Even though direct measurement of sediment transport rate is the most reliable method, it is impractical and expensive to set up gauging stations at the desired locations and collect data for a sufficiently long period of time. Especially the traditional bed-load data collection methods tend to be expensive, labor intensive, time-consuming, difficult, and under some conditions, hazardous (Gray et al., 2007). In most cases, the only available measured sediment data is suspended load data since it is very rare and very difficult to measure the bed load data (Tsai et al., 2010). For this reason, the observed suspended sediment load has included in predictive models as an input parameter within other input variables. In sediment modeling literature it is common to use either observable parameters of sediment system or non-dimensional parameters that are derived from these parameters. Therefore in this study, four input sets were used. The first input set consists of observable parameters of sediment system. Other input sets are depending upon dimensional analysis, the variables are non-dimensional parameters. The second and third input set is widely- used input sets of Bhattacharya (2007) and Sasal (2009). The last input set is depicted from the sensitivity analysis study of Pektaş (2015), in this study a large non-dimensional variable set is examined and the most relevant parameters for bed load prediction models are selected. These input sets are presented in Tab. 1.

After the compilation process of the overall dataset (observed parameters) the non-dimensional parameters are derived by using the formulations of the Tab. 1. Then these input sets are used to predict the non-dimensional bed load, two times. In the first step, models are performed without using the dimensionless suspended load and in second step dimensionless suspended load is added by preserving all specifications of models. The additive input sets are abbreviated by adding +1 such as *Input set x*+1 to the original *Input set x*. Furthermore, sensitivity analyses have been performed on models to evaluate the effect of additional parameter. The work flow of the study is shown in Fig. 1.

*Table 1. Investigated Input sets and parameters.*

| | Input set 1 | Input set 2 | Input set 3 | Input set 4 |
|---|---|---|---|---|
| | (Observable Parameters) | Bhattacharya (2007) | Sasal (2009) | Pektaş (2013) |
| Model inputs | River width<br>(B, L) | Dimensionless Flow depth<br>( $D_{non-dimen} = D / d_{50}$ ) | Dimensionless Flow depth<br>( $D_{non-dimen} = D / d_{50}$ ) | Dimensionless unit stream power<br>( $Sp_{non-\mathrm{dim}en} = \frac{u_m S}{\omega}$ ) |
| | River depth<br>(D, L) | Froud Number particle<br>( $Fr_p = \frac{u_m}{\sqrt{(G_s - 1) g d_{50}}}$ ) | Froud Number particle<br>( $Fr_p = \frac{u_m}{\sqrt{(G_s - 1) g d_{50}}}$ ) | Dimensionless velocity parameter<br>( $U_{non-\mathrm{dim}en} = \frac{u_m^3}{gD\omega}$ ) |
| | Slope<br>(S, L/L) | Dimensionless Particle diameter<br>( $d_* = \left[\frac{g(G_s - 1)}{\nu^2}\right]^{1/3} d_{50}$ ) | Dimensionless Particle diameter<br>( $d_* = \left[\frac{g(G_s - 1)}{\nu^2}\right]^{1/3} d_{50}$ ) | Dimensionless unit discharge<br>$q_{non-\mathrm{dim}en} = \frac{q}{u_* d_{50}}$ |
| | Median grain diameter<br>($d_{50}$, L) | Grain Size Reynolds Number<br>$\mathrm{Re}_* = \frac{u_* d_{50}}{\nu}$ | Depth scale ratio<br>(D/B) | Froude number<br>$Fr = \frac{q}{\sqrt{gDD}}$ |
| | Stream velocity<br>( $u_m, L.T^{-1}$ ) | Rouse Number<br>( $\frac{\omega}{k u_*}$ ) , $(k = 0.4)$ | | Particle parameter 2<br>( $P_{par} = \frac{\nu u_*}{g(G_s - 1) d_{50}^2}$ ) |
| Additional input | Dimensionless Suspended load<br>( $\phi_S = \frac{C_{vS} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) | Dimensionless Suspended load<br>( $\phi_S = \frac{C_{vS} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) | Dimensionless Suspended load<br>( $\phi_S = \frac{C_{vS} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) | Dimensionless Suspended load<br>( $\phi_S = \frac{C_{vS} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) |
| Model output | Dimensionless Bed load<br>( $\phi_B = \frac{C_{vB} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) | Dimensionless Bed load<br>( $\phi_B = \frac{C_{vB} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) | Dimensionless Bed load<br>( $\phi_B = \frac{C_{vB} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) | Dimensionless Bed load<br>( $\phi_B = \frac{C_{vB} u_m R}{\sqrt{g(G_s - 1) d_{50}^3}}$ ) |

where:

$D\,(m)$ = uniform flow depth,
$R\,(m)$ = hydraulic radius,
$d_{50}\,(m)$ = median diameter,
$B\,(m)$ = channel width,
$u_*\,(m/s)$ = shear velocity,
$v\,(m^2/s)$ = kinematic viscosity,
$S\,(m/m)$ = slope,
$G_s$ = specific gravity of sediment,
$u_m\,(m/s)$ = flow mean velocity,
$\omega\,(m/s)$ = fall velocity,
$q\,(m^3/s/m)$ = discharge per unit width,
$g\,(m/s^2)$ = the acceleration due to gravity,
$C_{vB}$ = volumetric concentration of bedload,
$C_{vS}$ = volumetric concentration of suspended load.

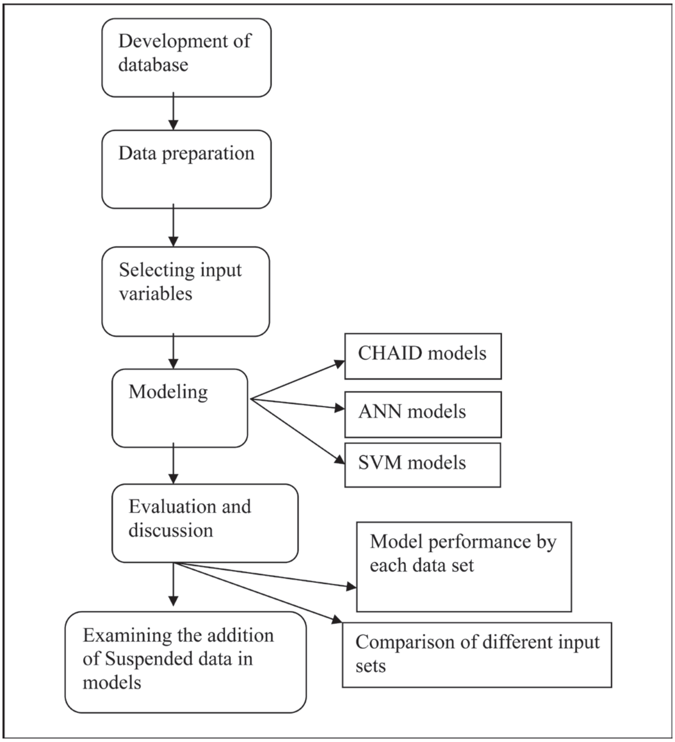
Development of database
Data preparation
Selecting input variables
Modeling
CHAID models
ANN models
SVM models
Evaluation and discussion
Model performance by each data set
Comparison of different input sets
Examining the addition of Suspended data in models

**Figure 1.** The workflow of the study.

## 3. Development of a data base

For this study a wide range of existing field data sets was compiled and analyzed. The field data used in this study were obtained from Emmet et al. (1978), Kircher (1983), Long and Liang (1994), Sinnakaudan et al. (2006) and Abdel-Fattah et al. (2004). The data were converted to consistent SI units. The units for all variables are shown in Tab. 2.

*Table 2. The units for variables.*

| No. | Variables | Unit |
|---|---|---|
| 1 | Water discharge | $m^3/s$ |
| 2 | Channel width | m |
| 3 | Flow depth | m |
| 4 | Water surface slope | m/m |
| 5 | Mean bed diameter $d_{50}$ | mm |
| 6 | Transported total sediment concentration | ppm |
| 7 | Transported bed load sediment concentration | ppm |
| 8 | Transported suspended load concentration | ppm |

*Table 3. List of investigations for field data.*

| Data code | Investigator(s) | Number of records |
|---|---|---|
| KAM | Kampar River, Sinnakaudan et al. (2006) | 21 |
| KER | Kerayong River, Sinnakaudan et al. (2006) | 27 |
| KIN | Kinta River, Sinnakaudan et al. (2006) | 20 |
| KUL | Kulim River, Sinnakaudan et al. (2006) | 16 |
| LAN | Langat River, Sinnakaudan et al. (2006) | 24 |
| LUI | Lui River, Sinnakaudan et al. (2006) | 92 |
| PAR | Pari River, Sinnakaudan et al. (2006) | 56 |
| RAI | Raia River, Sinnakaudan et al. (2006) | 41 |
| SEM | Semenyih River, Sinnakaudan et al. (2006) | 49 |
| PLT | Plate River, Kircher (1983) | 20 |
| TAN | Tanana River, Emmett et al.(1978) | 10 |
| YEL | Yellow River, Long and Liang (1994) | 19 |
| YLW | Lower Yellow River, Long and Liang (1994) | 1086 |
| ASW | Aswan River, Abdel-Fattah, S. et al. (2004) | 6 |
| QUE | Quena River, Abdel-Fattah, S. et al. (2004) | 6 |
| SOH | Sohag River, Abdel-Fattah, S. et al. (2004) | 6 |
| BSW | Bani-Sweif River, Abdel-Fattah, S. et al. (2004) | 6 |
| **TOTAL** | **17 Rivers** | **1505** |

A total of 1505 field data points from several kinds of river beds and river sizes were selected for analysis. In total database comprising of 17 river systems, a list of which is given in Tab. 3. The data also represent a wide variety of locations, including rivers in the USA, Europe, and Asia.

## 4. Application of the models

Data was divided into two parts as model development and model validation parts (also called testing). Since because the partition could be effective and has a bias in model performance evaluation, the Probability density functions (PDF) were taken into consideration. While partitioning the dataset by randomly, attention was paid to obtain similar PDFs of target variable (non-dimensional bed load) for two parts of the dataset. The model validation data were not used in any part of models and only used in the evaluation of model performances. Figure 1 shows the PDFs of model development and validation parts. As shown in figure the PDFs are fairly similar.

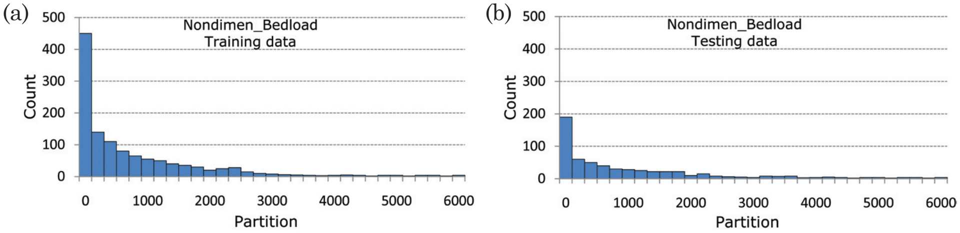

**Figure 2.** The PDF of partitioned data by non-dimensional bed load.

### *4.1. ANN models*

A neural network, commonly known as a multilayer perceptron, is essentially a simplistic model of the way the human brain handles the information. ANN performs by simulating many interconnected units that imitate the work style of neurons. The functioning units are positioned in layers. There are characteristically three regions in a neural network: an input layer, with units on behalf of the input fields; at least one hidden layer(s); and an output layer, affiliated with the output field(s). The units are linked with changeable connection strengths (or weights). Input data are displayed in the first layer, and consequently values are propagated from each neuron to every neuron within the next layer. Ultimately, a result is responded from the output layer to minimize an error function.

The network learns by analyzing single data points, attaining a prediction per each record, and making adjustments to the weights once it creates a wrong prediction. This is a repetitive process, therefore the network moves forward to improve its predictions until any of the stopping criteria have been met (SPSS neural network, 2010). Neural networks can be used to tackle nonlinear problems which are not adaptable to conventional statistical and mathematical methods. In the past few years there has been escalating interest in neural networks modeling in several fields of hydrology engineering (ASCE, 2000).

At the beginning, all weights are random, so the responses which come right from the network are more likely absurd. The network learns via training. Examples in which the output is known are repeatedly introduced to the network, and the answers the program gives are compared to the known outcomes. Data derived from this assessment is forwarded back through the network, progressively adjusting the weights. Along with the training progresses, the network gets to be significantly accurate in replicating the known outcomes. One of the problems that arise during the course of neural network training is called overtraining / over fitting (SPSS neural network, 2010). In this study, to prevent from overtraining the training part of the dataset was randomly split into two parts. The network was trained on the first part and the accuracy was estimated based on the second part. The test dataset was considered as a hold-out data set and only used in the evaluation of the model performance. In the study, the sigmoid transfer functions were used on a three layered network. The momentum term Alpha which is used in updating the weights during training is set to 0.9. The learning rate (Eta), which controls how much the weights were adjusted at each update, was set initially to 0.3 and the edge values were set at 0.1 and 0.01.

### *4.2. CHAID models*

Decision tree models enable to establish classification systems that predict upcoming responses depending upon certain decision rules. Typically, tree models have some advantages when compared with black box models like neural networks. Originally, the judging procedure behind the model is explicitly evident while exploring the tree. Furthermore, the process will automatically involve in its rule system solely the characteristic that actually makes a difference to generate a decision. Factors which do not promote the accuracy of the tree are ignored, therefore valuable information could be involved regarding the data. In this study, the chi-squared automatic interaction detection (CHAID) decision tree model was used, since because applications of CHAID tree models in sediment modeling does not exist. The CHAID algorithm creates decision trees using chi-square statistics to establish optimal splits. Dissimilar to the classification and regression tree models, CHAID is able to create nonbinary trees, which means that some splits get more than two branches. In CHAID procedure the initial step is forming categorical predictors right from any continuous predictors by isolating the respective continuous distributions into certain categories with an approximately equal number of observations. After that the cross tabulations are checked between every one of the predictor variables and the outcome. Finally, significance tests are performed; F tests are used for continuous variables.

If the corresponding test for the predictor categories is not statistically significant on the ground that specified by an alpha-to-merge value, then CHAID will integrate the respective predictor categories to perform this stage again (IBM SPSS Decision trees, 2011). In case that more than one of these relations is statistically significant, CHAID will choose the predictor stated as the most significant (smallest $p$ value). In this study, the significance levels (alpha levels) for splitting nodes and merging the categories were selected as 0.05. Maximum iterations for convergence was limited to 100 iterations and convergence epsilon was selected as 0.001, which determines how much change must occur for iterations to continue.

### *4.3. SVM models*

Support Vector Machine (SVM) is a powerful classification and regression technique that maximizes the predictive accuracy of a model without over fitting the training data. SVM operates by mapping the data into a high-dimensional characteristic space to ensure that the data points could be classified, even if the data are not linearly separable. A separator between the categories identified, and the data are transformed in such a manner that the separator might possibly be drawn as a hyper plane. After the transformation, the boundary between the two categories can be described by a hyper plane. The mathematical operators chosen for the transformation is regarded as the kernel function. In literature, there are many kernel types like linear, polynomial, radial basis function (RBF) or sigmoid. A linear kernel function is recommended when linear separation of the data is straightforward (IBM SPSS modeler, 2011). In other cases, one of the other functions should be used. In this study RBF kernel functions were used. The RBF gamma value was selected as 0.1. This value is lower than the calculated threshold value that should be between $3/k$ and $6/k$, where $k$ is the number of input fields, to avoid over fitting. Regression precision epsilon value was taken 0.1 as suggested by IBM SPSS modeler (2011). Regularization parameter ($C$) which controls the trade-off between maximizing the margin and minimizing the training error term was selected as 10. This is the default and suggested optimum value since increasing this value boost the accuracy but causes over fitting in data training (IBM SPSS modeler, 2011).

## 5. Results and discussion

The performance control of the model outputs was evaluated by widely used and well known correlation coefficient ($R$), coefficient of determination ($R^2$) and Nash-Sutcliffe model efficiency coefficient ($E_{Nash}$). As known determination and correlation coefficient value changes between 0 and 1. The values close to 1 are the indication of high accuracy. Nash-Sutcliffe efficiencies can range from $-\infty$ to 1. An efficiency of 1 corresponds to a perfect match of modeled discharge to the observed data. An efficiency of 0 indicates that the model predictions are as accurate as the mean of the observed data, whereas an efficiency less than zero occurs when the observed mean is a better predictor than the model or, in other words, when the residual variance, is larger than the data variance (Moriasi, 2007). The equation of the Nash-Sutcliffe efficiency coefficient is:

$$E_{Nash} = 1 - \left\{ \frac{\sum_{t=1}^{n} \left( Q_o^t - Q_m^t \right)^2}{\sum_{t=1}^{n} \left( Q_o^t - \bar{Q}_o \right)^2} \right\} \tag{1}$$

where $\bar{Q}_o$ is the mean of observed values, $Q_m^t$ and $Q_o^t$ are modeled and observed value at time *t* respectively and *n* is the total number of the testing patterns.

In Tab. 4, for each model, all the described model fit statistics are presented.

*Table 4. Goodness of fit statistics ($E_{Nash}$ – Nash-Sutcliffe model efficiency coefficient, R – correlation coefficient, $R^2$ – coefficient of determination).*

| | Input set 1 | | | Input set 1+1 | | |
|---|---|---|---|---|---|---|
| | ANN | CHAID | SVM | ANN | CHAID | SVM |
| $E_{Nash}$ | 0,470 | 0,696 | 0,126 | 0,655 | 0,704 | 0,133 |
| $R$ | 0,686 | 0,834 | 0,621 | 0,809 | 0,839 | 0,639 |
| $R^2$ | 0,470 | 0,696 | 0,386 | 0,655 | 0,704 | 0,409 |
| | Input set 2 | | | Input set 2+1 | | |
| | ANN | CHAID | SVM | ANN | CHAID | SVM |
| $E_{Nash}$ | 0,668 | 0,668 | 0,154 | 0,715 | 0,671 | 0,160 |
| $R$ | 0,818 | 0,818 | 0,646 | 0,847 | 0,819 | 0,659 |
| $R^2$ | 0,669 | 0,669 | 0,418 | 0,717 | 0,672 | 0,434 |
| | Input set 3 | | | Input set 3+1 | | |
| | ANN | CHAID | SVM | ANN | CHAID | SVM |
| $E_{Nash}$ | 0,687 | 0,671 | 0,158 | 0,720 | 0,676 | 0,165 |
| $R$ | 0,829 | 0,820 | 0,679 | 0,849 | 0,822 | 0,692 |
| $R^2$ | 0,687 | 0,672 | 0,461 | 0,720 | 0,676 | 0,479 |
| | Input set 4 | | | Input set 4+1 | | |
| | ANN | CHAID | SVM | ANN | CHAID | SVM |
| $E_{Nash}$ | 0,662 | 0,649 | 0,073 | 0,767 | 0,689 | 0,082 |
| $R$ | 0,820 | 0,807 | 0,743 | 0,877 | 0,830 | 0,777 |
| $R^2$ | 0,672 | 0,651 | 0,552 | 0,769 | 0,689 | 0,603 |

### *5.1. Evaluation of applied methods*

For all input sets the correlation coefficients are shown is Fig. 3. In most of the models, the prediction performances of ANN and CHAID models are good (mostly $R > 0.8$) on the other hand the performance of SVM is comparatively bad. This is probably because of the exility of the selected model fining parameters (RBF gamma and regularization parameter) in SVM. To avoid over training problem these parameters have been selected within minimum level as far as possible. In the scope of this study, only constant and default model fine settings were used (as explained earlier) for the sake of comparing the performances of different input sets and variables concurrently. For example, while tuning the SVM models RBF gamma option is fixed to 0.1. To obtain high predictive capacity this value should be between $3/k$ and $6/k$, where $k$ is the number of input fields. Increasing this value would improves the model accuracy (via reducing the regression error) for the training data, but this can also lead to over fitting. Similarly, for each type of model, some tuning parameters could be increased to increase the model accuracies but the main focus of the study is selecting the best input set and within this process also comparing the models. So minimum comparable model fining values are selected and fixed for all applied models, constantly.

For all eight input sets ANN models give the best performance except for the *input set* 1+1, for this input set the CHAID rule induction method give the highest $R$ value (Fig. 3).

In Fig. 4 the fit performance of more accurate models of each applied method and their scatter diagrams are presented for testing partitions of *input sets x*. The most accurate models of ANN, CHAID and SVM are obtained in the input

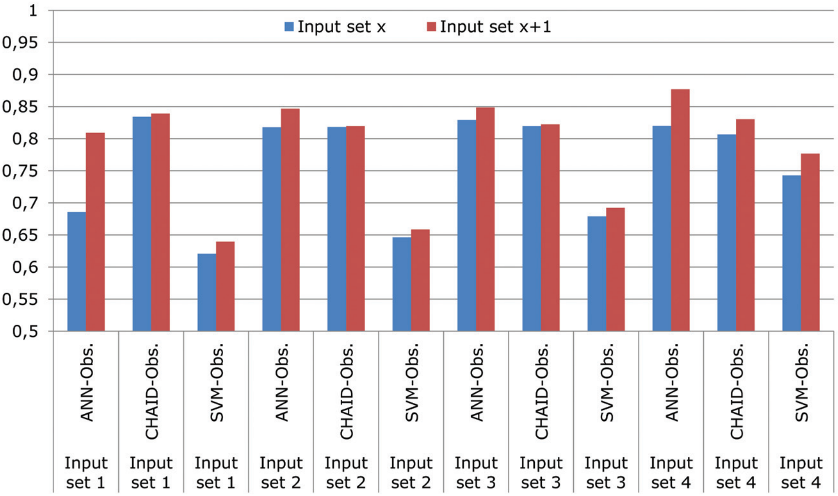

**Figure 3.** Comparison of model performances by depending input sets.

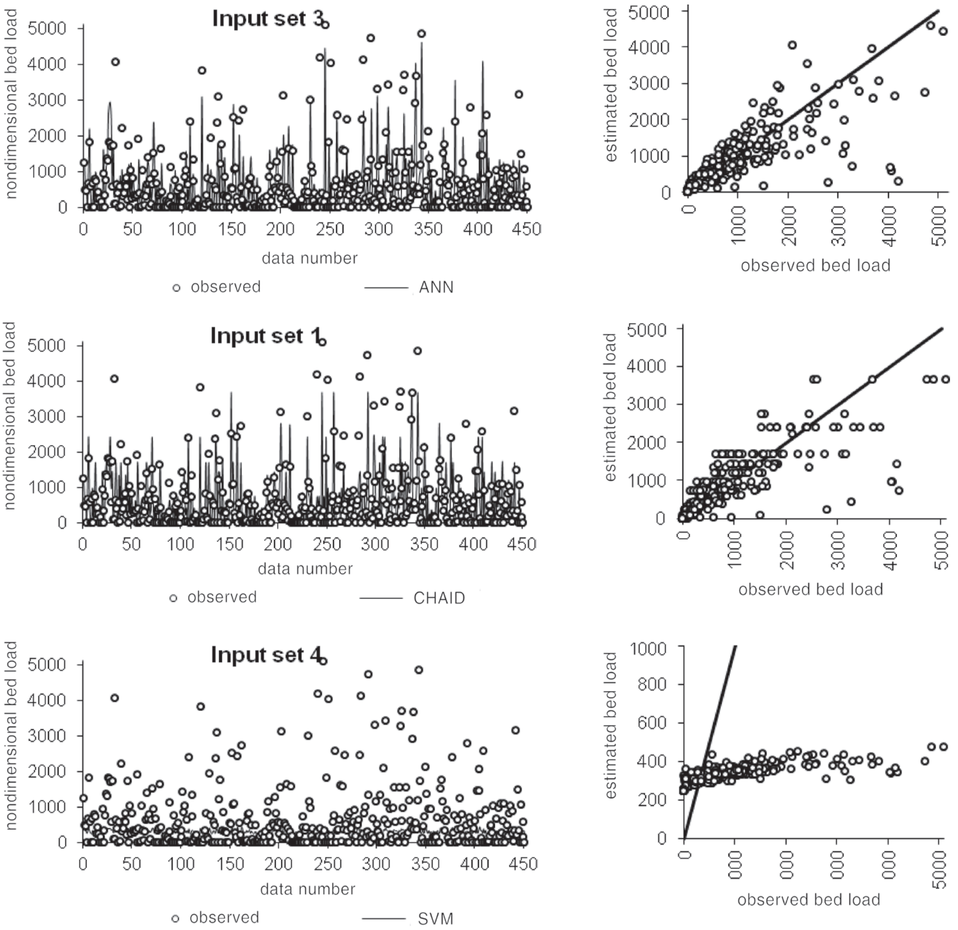

**Figure 4.** The best model of each method *input set x.*

sets of *Input set 3*, *Input set 1* and *Input set 4*, respectively. As shown, for the high values of *Dimensionless bed load* (>2000), all models give poor predictions by deviating the $y=x$ line on scatter diagrams. Especially, the VM model has a smooth horizontal line in the scatter diagram mostly underestimating the observed values. The results of SVM is converging a constant range value of 0–500. In Fig. 5 the three best models are presented for the *input sets* $x+1$. As shown in scatter diagrams for higher values of Bed load ( > 3000 ) ANN and CHAID make significant deviations from $y=x$ line. General fit performance of these two models is fairly good with the correlation coefficient values of $R=0.88$ and $R=0.85$, respectively. On the other hand the SVM model has a poor accuracy ($R=0.75$) with large deviations from the $y=x$ line. By comparing the scatter dia-

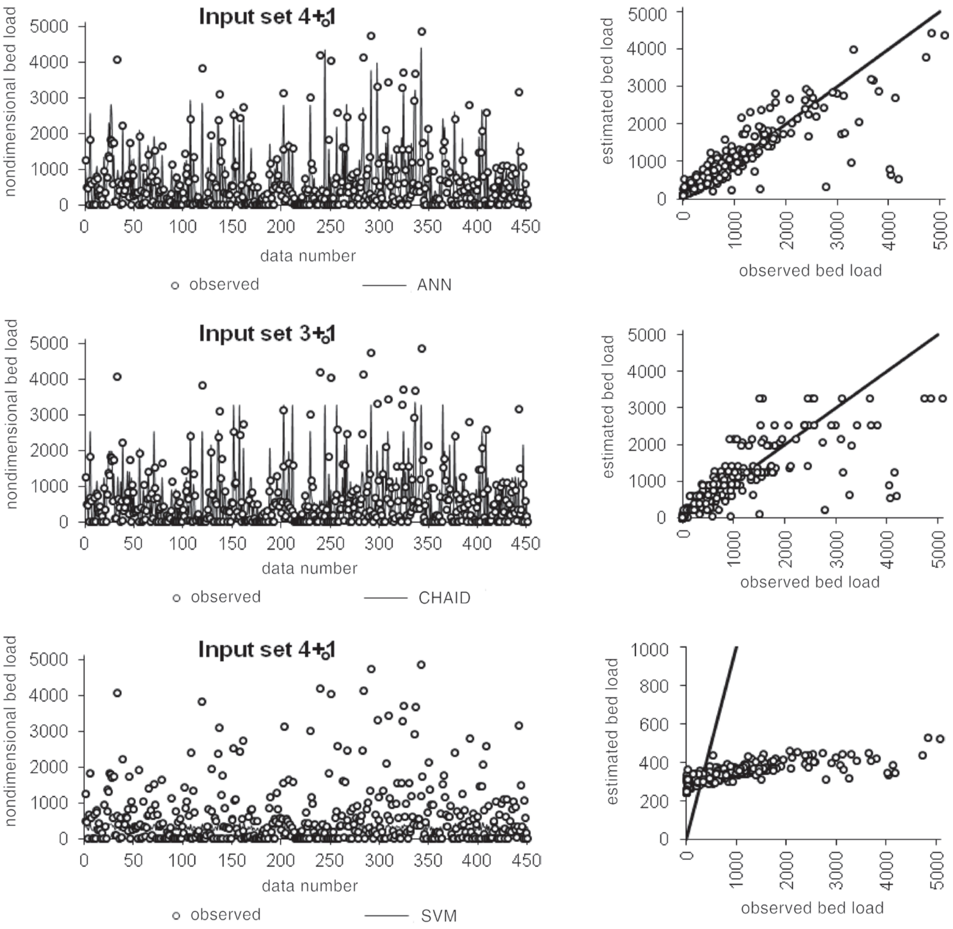

**Figure 5.** The best model of each method *input set* $x+1$.

grams in Fig. 4 and 5, it can be asserted that ANN models give slightly better results than the CHAID methods. Furthermore, both models overestimated the low values and underestimated peak values.

As shown in Figs. 4 and 5, applied models are usually made under predictions since the PDF of observed values have long tails for high values. The models' accuracy is investigated by the sense of over prediction ratios. This ratio simply calculated by dividing the numbers of data points to the total data number where the model has over estimated. So the very close and very far estimations of $y=x$ line have both scored as 1 if these are overestimations. Table 5 shows the over prediction ratios of models. ANN and CHAID models are overestimating, but the overestimation ratio in SVM models is close to underestimation ratios.

*Table 5. Over-prediction ratios of models for all input sets.*

| Over-prediction ratios (%) | *Input set x* | | | *Input set x*+1 | | |
|---|---|---|---|---|---|---|
| | ANN model | CHAID model | SVM model | ANN model | CHAID model | SVM model |
| Input set 1 | 67.63 | 67.85 | 48.34 | 71.18 | 69.40 | 48.56 |
| Input set 2 | 68.74 | 67.18 | 50.33 | 78.05 | 68.29 | 50.11 |
| Input set 3 | 72.06 | 64.97 | 48.56 | 71.62 | 66.96 | 48.56 |
| Input set 4 | 80.93 | 71.40 | 48.34 | 78.05 | 68.96 | 48.34 |

### *5.2. Evaluation of the role of suspended sediment load in models*

As shown in Fig. 3, addition of suspended load into models has increased the model performances for all input sets. For all input sets the accuracy of the *input set x* + 1 is higher than the accuracy of the *input set x*. Especially in ANN models this increment is significant. The most accurate model out of 12 models is the ANN model for the *input set* 4 + 1. The effect of the contribution of dimensionless suspended sediment load in this model could be seen by comparing the variable importance coefficients at Figs. 6a–b. In Fig. 6a the most significant parameter,

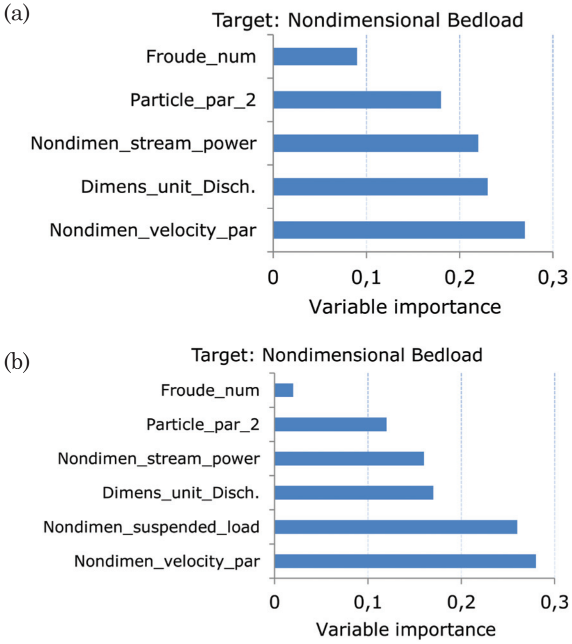

**Figure 6.** The relative variable importance coefficients for *Input set* 4 + 1 ANN model.

non-dimensional *velocity parameter,* has a variable importance coefficient about 0.27. When the *dimensionless suspended load* variable was added to the model this variable became the most significant variable of the model with an importance coefficient of 0.33 (Fig. 6b). Variable importance coefficients were calculated for all applied models and presented after normalization to obtain a total value of 1. After the training process of applied models, these coefficients were calculated on the testing input set by calculating the variance of model response via immobilizing all variables except the investigated variable. The records (data points) of investigated variable were changed by scaled increments and the response of the stated model was stored to calculate variable importance coefficients (SPSS neural network, 2010). The calculated variable importance coefficients of best predictive models for *input sets* $x+1$ is presented in Tab. 6. As shown in the table, the suspended load has highest variable importance rank for the CHAID model of *input set* $1+1$, and has a second degree importance for ANN models of *input set* $3+1$*, input set* $4+1$.

For *input set* $1+1$*,* the CHAID rule induction system starts with branching the root into 8 nodes by using the parameter dimensionless suspended load. This tree generation process is indicating that this parameter is the most significant parameter for CHAID model.

*Table 6. Variable importance coefficients of the best models for input sets* $x+1$*.*

| Input set 1+1 (CHAID Model) | Importance | Input set 3+1 (ANN Model) | Importance |
|---|---|---|---|
| Width | 0.0355 | nondimen_flow_depth | 0.0447 |
| Slope | 0.0371 | nondimen_particle_diameter | 0.0754 |
| Depth | 0.0671 | Depth_scale_ratio | 0.0868 |
| d50 | 0.0980 | Nondimen_Suspended_load | 0.3726 |
| Flow_velocity | 0.1170 | Froude_number_particle | 0.4205 |
| Nondimen_Suspended_load | 0.6453 | Input set 4+1 (ANN Model) | Importance |
| Input set 2+1 (ANN Model) | Importance | Froude_number_Flow | 0.0066 |
| Grain_size_Reynolds | 0.0000 | Particle_parameter_2 | 0.1189 |
| Nondimen_Suspended_load | 0.0049 | Nondimen_unit_stream_power | 0.1690 |
| Rous_number | 0.0052 | dimensionles_unit_discharge | 0.1722 |
| nondimen_particle_diameter | 0.0935 | Nondimen_Suspended_load | 0.2599 |
| Froude_number_particle | 0.8964 | Nondimen_velocity_parameter | 0.2733 |

### *5.3. Evaluation of input sets*

As shown in Fig. 7, for a method that all the specifications (parameter tunings) are immobilized, the performance ($R$ value) has a smooth variability depending upon input sets. The SVM model performances increasing linearly through from the *input set* 1 to *input set* 4, and from the *input set* 1 + 1 to *input set* 4 + 1. The same situation is observed in ANN models (Fig. 7) for *input sets* $x+1$, the most predictive ANN model that is the best predictive model of all applied models, is achieved by using the *input set* 4+1. For the *input sets x* the most accurate ANN model is obtained by using *input set* 3, the $R$ value between model output and observed values are 0.83 meanwhile the performance of *input set* 4 is $R = 0.82$ which is close to 0.83. For CHAID models all the input sets give similar performances, but the most accurate models are obtained by using the *input set* 1 and *input set* 1 + 1. It can be concluded that the *input set* 4 + 1 yields best results for applied ANN and SVM models, while the *input set* 1 + 1 that consist of dimensional observable quantities gives higher results (but close to others) for the applied CHAID model.

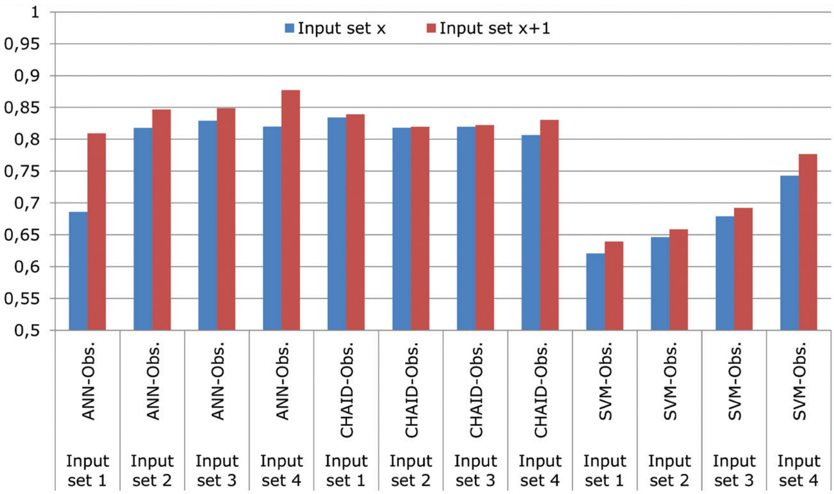

**Figure 7.** Model performances varying with input sets.

## 6. Conclusion

In this study, artificial neural networks, support vector machines and CHAID decision tree models were used to predict the bed load sediment load. To push the envelope of model performances a high range and skewed dataset was used.

It was observed that although the two models return similar results, the ANN model slightly outperforms the CHAID models. On the other hand the performance of SVM models found comparatively insufficient. Both ANN and CHAID models overestimated the low values and underestimated peak values, but SVM model is usually created underestimated results. The *input set* $4+1$ gives more predictive accuracy for ANN and SVM models, on the other hand, in CHAID models the *input set* $1+1$ gives the most accurate results. The addition of suspended load is boosted model accuracies for all input sets. So, if it is possible, it is recommended to use the Suspended load to predict bed load. In further studies the influences of model tuning parameters (model specifications) would be investigated on a single input set. The results of the study are highly encouraging and suggest that an ANN or Decision Tree CHAID approach is reasonable for modeling sediment load prediction.

## References

Abdel-Fattah, S., Amin, A. and Van Rijn, L. C. (2004): Sand transport in Nile River, Egypt, *J. Hydraul. Eng.-ASCE*, **130**, 488–500, DOI: 10.1061/(ASCE)0733-9429(2004)130:6(488).

ASCE Task Committee (2000): Artificial neural networks in hydrology. I: Preliminary concepts, *J. Hydrol. Eng.-ASCE*, **5**, 115–123, DOI: 10.1061/(ASCE)1084-0699(2000)5:2(115).

Bhattacharya, B., Price, R. K. and Solomatine, D. P. (2007): Machine learning approach to modeling sediment transport, *J. Hydraul. Eng.-ASCE*, **133**, 440–450, DOI: 10.1061/(ASCE)0733-9429(2007)133:4(440).

Ciğizoğlu, H. K. (2002a): Suspended sediment estimation and forecasting using artificial neural networks, *Turkish J. Eng. Env., TÜBITAK*, **26**, 15–25.

Ciğizoğlu, H. K. (2002b): Suspended sediment estimation for rivers using artificial neural networks and sediment rating curves, *Turkish J. Eng. Env., TÜBITAK*, **26**, 27–36.

Emmett, W. W., Burrows, R. L. and Parks, B. (1978): *Sediment transport in the Tanana River in the vicinity of Fairbanks, Alaska*. U.S. Department of the Interior Geological Survey, Anchorage, Alaska, Open-File Report, 78-290, 28 pp.

Gao, P. (2011): An equation for bed-load transport capacities in gravel-bed rivers, *J. Hydrol.*, **402**, 297–305, DOI: 10.1016/j.jhydrol.2011.03.025.

Gray, J. R., Laronne, J. B. and Marr, J. D. G. (2007): Measuring bed load discharge in rivers – Bedload-Surrogate Monitoring Workshop, Minneapolis, Minnesota, 11–14 April 2007, *Eos Trans. AGU*, **88(45)**, 471–471, DOI: 10.1029/2007EO450008.

IBM SPSS Decision Trees 20 (2011): Copyright SPSS Inc. 1989, 2011, 108 pp, available at http://www.csun.edu/sites/default/files/decision-trees20-64bit.pdfhttp://www.csun.edu/sites/default/files/decision-trees20-64bit.pdf

IBM SPSS Modeler 14.2 Users Guide (2011): Copyright IBM Corporation 1994, 2011, 261 pp, available at http://faculty.smu.edu/tfomby/eco5385/data/SPSS/SPSS%20Modeler_14_2_UsersGuide.pdfhttp://faculty.smu.edu/tfomby/eco5385/data/SPSS/SPSS Modeler_14_2_UsersGuide.pdf

IBM SPSS Neural Networks 19. (2010): Copyright SPSS Inc. 1989, 2010, 100 pp, available at http://www.csun.edu/sites/default/files/neural-network19.pdfhttp://www.csun.edu/sites/default/files/neural-network19.pdf

Jain, S. K. (2001): Development of integrated sediment rating curves using ANNs, *J. Hydraul. Eng-ASCE,* **127**, 30–37, DOI: 10.1061/(ASCE)0733-9429(2001)127:1(30).

Kircher, J. E. (1983): Interpretation of sediment data for the South Platte River in Colorado and Nebraska, and the North Platte and Platte Rivers in Nebraska, in: *Hydrologic and geomorphic studies of the Platte River basin*, U.S. Department of the Interior Geological Survey, Washington, D.C., Professional Paper 1277-D, 37 pp.

Kisi, Ö., Dailr, A. H., Çimen, M. and Shiri, J. (2012): Suspended sediment modeling using genetic programming and soft computing techniques, *J. Hydrol.*, **450–451**, 48–58, DOI: 10.1016/j.jhydrol.2012.05.031.

Long, Y. and Liang, G. (1994): Data base of sediment transport in the Yellow River. *Technical Report No.* 94001, Institute of Hydraulic Research, Yellow River Conservation Commission, Zhengzhou, P. R. China, 15 pp (in Chinese).

McBean, E. A. and Al-Nassri, S. (1988): Uncertainty in suspended sediment transport curves, *J. Hydraul. Eng-ASCE*, **114**, 63–74, DOI: 10.1061/(ASCE)0733-9429(1988)114:1(63).

Misra, D., Oommen, T., Agarwal, A., Mishra, S. K. and Thompson, A. M. (2009): Application and analysis of support vector machine based simulation for runoff and sediment yield, *Biosyst. Eng.*, **103**, 527–535, DOI: 10.1016/j.biosystemseng.2009.04.017.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L.; Harmel, R. D. and Veith, T. L. (2007): Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *T. ASABE*, **50**, 885–900, available at http://swat.tamu.edu/media/1312/moriasimodeleval.pdfhttp://swat.tamu.edu/media/1312/moriasimodeleval.pdf

Nagy, H. M., Watanabe, K. and Hirano, M. (2002): Prediction of sediment load concentration in rivers using artificial neural network model, *J. Hydraul. Eng-ASCE*, **128**, 588–595, DOI: 10.1061/(ASCE)0733-9429(2002)128:6(588).

Nakato, T. (1990): Tests of selected sediment-transport formulas, *J. Hydraul. Eng-ASCE*, **116**, 362–379, DOI: 10.1061/(ASCE)0733-9429(1990)116:3(362).

Oehler, F., Coco, G., Green, M. O. and Bryan, K. R. (2012): A data-driven approach to predict suspended-sediment reference concentration under non-breaking waves, *Cont. Shelf Res.*, **46**, 96–106, DOI: 10.1016/j.csr.2011.01.015.

Öztürk, F., Apaydın, H. and Walling, D. E. (2001): Suspended sediment loads through flood 0events for streams of Sakarya River Basin, *Turk. J. Engin. Env. Sci.*, *TÜBITAK* **25**, 643–650, available at http://dergipark.ulakbim.gov.tr/tbtkengineering/article/view/5000025162/5000025399http://dergipark.ulakbim.gov.tr/tbtkengineering/article/view/5000025162/5000025399

Pektaş, A. O. (2015): Determining the essential parameters of bedload and suspended sediment load, *Int. J. Global Warm.,* in press.

Raghuwanshi, N. S., Singh, R. and Reddy, L. S. (2006): Runoff and sediment yield modeling using artificial neural networks: Upper Siwane River, India, *J. Hydrol. Eng.-ASCE*, **11**, 71–79, DOI: 10.1061/(ASCE)1084-0699(2006)11:1(71).

Sasal, M., Kashyap, S., Rennie, C. D. and Nistor, I. (2009): Artificial neural network for bedload estimation in alluvial rivers, *J. Hydraul. Res.*, **47**, 223–232, DOI: 10.3826/jhr.2009.3183.

Şen, Z. and Altunkaynak, A. (2003): Fuzzy awaking in rainfall-runoff modelling, *Nord. Hydrol.*, **35**, 31–43.

Sinnakaudan, S. K., Ab Ghani, A., Ahmad, M. S. and Zakaria, N. A. (2006): Multiple linear regression model for total bed material load prediction, *J. Hydraul. Eng.-ASCE*, **132**, 521–528, DOI: 10.1061/(ASCE)0733-9429(2006)132:5(521).

Tayfur, G. and Guldal, V. (2005): Artificial neural networks for estimating daily total suspended sediment in natural streams, *Nord. Hydrol.*, **37,** 69–79.

Tsai, C. T., Tsai, C. H., Weng, C. H., Bair, J. J. and Chen, C. N. (2010): Calculation of bed load based on the measured data of suspended load, *Paddy Water Environ.*, **8**, 371–384, DOI: 10.1007/s10333-010-0216-4.

Vanoni, V. A. (1971): Sediment discharges formulas, *J. Hydraul. Div.-ASCE*, **97**, 523–567.

Yang, C. T. (1972): Unit stream power and sediment transport, *J. Hydraul. Div.-ASCE*, **98**, Proceeding Paper 9295, 1805–1826.

Yang, C. T. (1996): *Sediment transport: Theory and practice*. McGraw-Hill , USA, 396 pp.

Yang, C. T, and Wan, S. (1991): Comparisons of selected bed-material load formulas, *J. Hydraul. Eng.-ASCE*, **117**, 973–989, DOI: 10.1061/(ASCE)0733-9429(1991)117:8(973).

Yu, G., Wang, Z., Zhang, K., Chang, T. C. and Liu, H. (2009): Effect of incoming sediment on the transport rate of bed load in mountain streams, *Int. J. Sed. Res.*, **24**, 260–273, DOI: 10.1016/S1001-6279(10)60002-9.

## SAŽETAK

### Prognoza nanosa putem suspendiranog nanosa pomoću metoda mekog računarstva

*Ali Osman Pektaş i Emrah Doğan*

Za kvantitativne i kvalitativne studije vodnih resursa od ključne je važnosti prikladna i prihvatljiva prognoza nanosa prenošenog vodotocima. Premda je najkonzistentnija metoda za određivanje nanosa *in situ* mjerenje stope nanosa, takva su mjerenja veoma skupa te se ne mogu provoditi na velikom broju vodotoka poput mjerenja nanosa suspendiranog sedimenta. Stoga je u ovoj studiji ispitana uloga suspendiranog nanosa u prognozi nanosa, pri čemu je primijenjena analiza osjetljivosti. Kako se konvencionalnim krivuljama stope sedimentacije i konvencionalnim jednadžbama ne mogu točno prognozirati sedimentni nanosi, u posljednje vrijeme jako porasla upotreba algoritama strojnog učenja. U skladu s tim, u ovoj studiji su primjenjene metode mekog računarstva. Poimence, primijenjeni su ovi modeli: umjetne neuronske mreže (ANN), metoda potpornih vektora (SVM) i modeli stabla odlučivanja (CHAID), koji su po prvi put upotrijebljeni u istraživanju sedimenata. Pojedini parametri često se koriste u metodama mekog računarstva pri kreiranju ulaznih skupova podataka. Ovdje su upotrijebljena tri uobičajena ulazna skupa te novi generirani skup, koji su najprije poslužili kao ulazni podaci za prognozu nanosa, a zatim je tim ulaznim skupovima dodana varijabla suspendiranog nanosa. Međusobno su uspoređene performanse modela s obzirom na ulazne skupove. Kako bi se generirali rezultati i smanjila ograničenja modela, s različitih lokacija prikupljeni su vrlo pristrani i heterogeni podaci. Rezultati pokazuju da su performanse ANN i CHAID modela stabla odlučivanja dobre u usporedbi sa SVM modelima. Upotreba suspendiranog nanosa kao dodatne ulazne varijable poboljšava performanse svih modela i značajno im doprinosi.

*Ključne riječi*: prognoza sedimenta, nanos, suspendirani nanos, umjetne neuronske mreže, metoda potpornih vektora, CHAID modeli stabla odlučivanja

*Corresponding author's address*: Ali Osman Pektaş, Bahcesehir University, Department of Civil Engineering, 34353 Beşiktaş, Istanbul, Turkey, e-mail: aliosmanpektas@gmail.com